UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF

Zentrum für Experimentelle Medizin, Institut für Angewandte Medizininformatik

Direktor der Einrichtung
Prof. Dr. med. Frank Ückert

Improving Interoperability and Generalizability through Technology in Parkinson's Disease

Dissertation

zur Erlangung des Doktorgrades PhD
an der Medizinischen Fakultät der Universität Hamburg.

vorgelegt von:

Christopher Lukas Gundler
aus Braunschweig

Hamburg 2025

(wird von der Medizinischen Fakultät ausgefüllt)

Angenommen von der
Medizinischen Fakultät der Universität Hamburg am: 26.01.2026

Veröffentlicht mit Genehmigung der
Medizinischen Fakultät der Universität Hamburg.

Prüfungsausschuss, der/die Vorsitzende: Prof. Dr. med Frank Ückert

Prüfungsausschuss, zweite/r Gutachter/in: Prof. Dr. med. Monika Pötter-Nerger

Prüfungsausschuss, dritte/r Gutachter/in: Prof. Dr. med. Martin Scherer

Contents

1 Introduction and theoretical background

The early twenty-first century has been marked by an rapid increase in the sheer volume, heterogeneity, and velocity of data generation across all spheres of human activity. This exponential growth in data, frequently described as a hallmark of contemporary society, derives primarily from the digitalization of previously analog processes (Hilbert and López 2011). Advances in information technology, networking infrastructure, computation, and data storage have created an environment in which data grew about 20–30% annually in the past decades (Hilbert 2016) but does so with increasing variety and veracity (Bellazzi 2014). In the specific context of health and medicine, the effects of these developments are obvious. Modern imaging modalities such as multi-slice computed tomography scans and high-field magnetic resonance imaging scanners produce multiple gigabytes of data in a single clinical session. Next generation sequencing could easily generate terabytes (Bellazzi 2014).

Besides the high-quality data obtained in controlled clinical environments and observational studies, the rise of secondary data sources adds another layer of complexity and opportunity (Näher et al. 2023). This broad class of data includes sources such as insurance claims, electronic health records, and patient registries. These have become increasingly important in regulatory and research settings and offer potential insights into the effectiveness, safety, and value of interventions outside the strict parameters of controlled clinical trials (Sherman et al. 2016). Accordingly, legislative initiatives like the *European Health Data Space* (Marcus et al. 2022) specifically discuss the potential of these newer and noisier data sources.

Another reason for the rise in data quantity is the democratization of data collection. Whereas, historically, data generation was the exclusive domain of academic centers, government agencies, or corporate laboratories, there has been a marked shift toward

the involvement of individuals. This "quantified self" movement illustrates the movement from passive subject to active data contributor (Swan 2013). Individuals now routinely collect datasets encompassing daily activities, physiological parameters, environmental exposures, nutritional intake, sleep patterns, and even report mood (Huckvale et al. 2019; Giannakopoulou et al. 2022). Consumer-grade wearable devices and health-focused smartphone applications have achieved global penetration, with most adults in developed economies possessing smartphones equipped with capabilities to measure heart rate, activity, oxygen saturation, and other health metrics (Piwek et al. 2016; Topol 2019). Beyond personal devices, participatory data initiatives have emerged that further decentralize and democratize knowledge creation. Patient-driven health registries, community science projects, and disease-specific platforms such as PatientsLikeMe (Wicks et al. 2010) and Open Humans (Greshake Tzovaras et al. 2019) enable users to voluntarily share granulated, long-term health information, disease experiences, and outcomes.

The result of these factors is the emergence of a remarkable ecosystem where data originates from multiple sources: institutionally governed clinical studies, large-scale administrative databases, real-world behavioral monitoring, and socially networked patient-reported outcomes. The traditional notion of the domain expert as the principal data creator is replaced by more distributed and inclusive models. The broader scientific community is increasingly characterized by the shift from discrete, well-defined datasets of the past to ongoing, high-frequency, multifaceted data streams (Roski et al. 2014). These developments emphasize the urgent need to reconsider multimodal data management, integration, and analytic frameworks, especially as science stands at the crossroads between data abundance and actionable knowledge.

Focus of the thesis

Despite advances in data collection and analytic capabilities, the translation of raw health data into improved clinical decision-making, reduced uncertainty, or novel scientific insights remains limited by both practical and theoretical barriers. Simply increasing the quantity or speed of data acquisition does not ensure greater scientific understanding or clinical utility (Sedlakova et al. 2023). In data-rich environments,

the essential difficulty is to distinguish meaningful signals from ever-increasing background noise (Goldstein et al. 2017).

Within the framework of information theory, information is defined by its potential to reduce entropy or uncertainty within a system (Cover and Thomas 2005). However, entropy is not automatically diminished by accumulating more data; rather, effective entropy reduction requires thoughtful curation, contextual understanding, and systematic interpretation. In medical informatics, these challenges are particularly pressing. The proportion of clinically relevant information in large-scale datasets, whether derived from routine care, electronic health records, or patient-generated data such as that from wearable sensors, is often low (Sedlakova et al. 2023). The presence of redundant, insufficiently annotated, or noisy data can, paradoxically, impair the performance of analytic models, leading to misleading or erroneous conclusions (Sperrin et al. 2019).

This problem is intensified by the increasing adoption of modern machine learning approaches. The so-called "scaling hypothesis" holds that training larger models on ever greater volumes of data will always result in superior performance. However, empirical findings consistently demonstrate that performance plateaus, saturation effects, and even declines can emerge as model complexity and dataset size increase (Diaz and Madaio 2024). Models trained on uncurated data risk overfitting, where the system learns spurious patterns, noise, or artifacts rather than robust, generalizable relations (Rajkomar et al. 2019). This not only undermines generalizability to new settings, but also introduces additional risks such as the unintentional incorporation of harmful data or a loss of model performance over time as data-generating processes change (Lenert et al. 2019).

Consequently, the central question arises: how can the increasing availability and diversity of multimodal health data be effectively leveraged while avoiding the risks of information overload, irreproducible results, or invalid model outputs? In particular, what methodological solutions exist both for standardized clinical assessments and for high-density time series data from consumer devices? This thesis proposes some solutions, using Parkinson's disease as an exemplary domain to explore opportunities and limitations in medical informatics and machine learning.

Outline of the thesis

This thesis contains a synopsis of my Parkinson-related research from the last few years. Within the background section, fundamental concepts linking data quality, modeling, and generalizability are defined, and their close relationship is presented. The main part connects these rather abstract principles to the practical realities of data related to Parkinson's disease. To this end, I will briefly summarize my related research and connect the corresponding publications. Finally, the discussion draws general implications and proposes best practices for scientific progress in data-rich, model-intensive settings such as research on neurodegenerative disorders.

Theoretical background

The following section will introduce the underlying concepts regarding generalizability from the perspective of medical informatics. While descriptions like data quality and generalizability are often used in different context, the following section should link these terms, show their close relationship, and place them in the scope of transfer learning as a technical solution for more sustainable results.

Data, information, and knowledge

Discussions about the promise and perils of data abundance and issues of generalizability often invoke central epistemological concepts like data, information, knowledge, and occasionally higher-level concepts (Rowley 2007). These categories, foundational in philosophy and information science, underpin regulatory standards, computational workflows, and ethical considerations within medical informatics and machine learning. Their boundaries remain objects of debate (Frické 2009), extending from classical dichotomies (empiricism vs. rationalism, as already in Aristotle's works) through to modern epistemology and cognitive science (Floridi 2011). Along with the spread of computer science around the world, modern definitions are commonly discussed in the literature. In a comprehensive survey by Zins (2007), over 130 distinct definitions advanced by 45 scholars were documented.

Understanding the operational meaning of these terms and, more importantly, the transitions between them is essential for establishing effective and efficient use of clinical data (Lehne et al. 2019), for example for machine learning.

Data: Raw, Context-Free Observations

At its core, data may be conceptualized as atomic facts like numbers, signals, or characters that are uninterpreted and context-free. Formally, one may denote a dataset as $X = \{x_1, ..., x_i\}$, where each x_i is a single measurement or recorded instance. In healthcare, such atomic data points might correspond to individual laboratory values, a single element of an electrocardiographic signal, readings from accelerometers, or a single pixel of a radiologic slice. On their own, these data points hold no intrinsic value; it is the process of contextualization and aggregation that imbues them with meaning (Lehne et al. 2019). Raw data can be voluminous, high-frequency, and seemingly precise, but until shaped by subsequent processes, it offers no guarantee of interpretability or relevance (Bellazzi 2014).

Information: Data with context and structure

Information emerges when data are enriched with context, structure, or interpretation (Rowley 2007). In the healthcare setting, assigning metadata to observations, such as values from blood samples, or grouping selected elements from electronic health record entries into patient timelines, are illustrative examples. This transformation requires not merely technical processing, but domain-specific expertise in what constitutes meaningful concepts (Lehne et al. 2019).

This enrichment of context is crucial, however, requires careful attention to the assumptions made. As an example, if one defines a data point as value of a blood sample, one cannot just define it as a prototypical blood sample, but must keep in mind it is a specific blood sample from a specific, study-dependent timepoint under potential study-dependent conditions (Benchimol et al. 2015). More formally, one may define the data points X from above to be elements of a feature space $\mathbb{X}$, $x_i \in \mathbb{X}$. The context is embedded by implicitly assuming a probability distribution $P(X)$ generating these samples. At this point, one often does not have any idea how this distribution is defined mathematically, but it encodes all beliefs regarding the context of the data. The

specification of both the feature space and the generative distribution defines then the domain $D = \{\mathbb{X}, P(X)\}$ of the information (Zhuang et al. 2021).

In this context, the concept of data quality gets important. The term does not cover only the basic correctness and accuracy of the data point itself (Lewis et al. 2023). While the definitions of the different dimensions of data quality are not consistent in literature (Bian et al. 2020), characteristics such as currency (timeliness of recording), completeness (presence of data), or plausibility (degree to which values reflect real-world processes) commonly appear in publications (Lewis et al. 2023). To a large extent, these dimensions are influenced by the implicit context given to data to make them information.

The absence or poor assurance of data quality, along with insufficient specification of the domain through the underlying generating distribution $P(X)$, is not merely a technical inconvenience in subsequent analyses. Rather, these factors have the potential to introduce systematic biases that may distort analytical outcomes (Lehne et al. 2019). The process of contextualizing data is inherently dependent on the conditions under which it is recorded, and there is no universally appropriate method for enriching data with context (Degtiar and Rose 2023). Consequently, the selection of contextual information and associated assumptions may impose specific limitations on the interpretability of the resulting information. Sticking to the previous example, the aggregation and contextualization of blood values collected from study participants aged 18 to 65 years is entirely valid when the objective is to analyze a particular disease within this specific age cohort. In this scenario, the implicit generating distribution describes the population of interest accordingly. However, this specific context and underlying distribution are not representative of the entire population, and extending conclusions from this dataset to populations outside of the recorded age range would be scientifically unsound (Alliende et al. 2023). Therefore, it is essential to ensure transparency regarding the context, assumptions, and limitations under which information is generated, particularly when it comes to the generalization and reuse of data in further analyses (Degtiar and Rose 2023). These forms of bias, if unaddressed, could then propagate through subsequent analytic stages.

Whether or not interoperability and accessibility are "just" another dimension of data quality or a distinct concept is not entirely clear. Some authors like Prasser et al. (2018) have listed them as distinct concepts while other authors have considered "Usability/Ease-of-Use", as observed by Bian et al. (2020). Interoperability would be

a necessary requirement for this "ease of use", similar to the term "portability" used in the ISO 25012. The FAIR (Findable, Accessible, Interoperable, and Reusable) data principles for scientific data management highlight this specific dimension explicitly, too (Wilkinson et al. 2016). In the context of data, the concept of interoperability refers to the capacity to be meaningfully compared, pooled, and analyzed across disparate sources, timeframes, or institutional boundaries (Cheng et al. 2024). Without sufficient interoperability and accessibility, information becomes trapped in data silos, leading to ineffective integration, duplication of effort, and barriers to scientific or clinical progress (Szarfman et al. 2022). Interoperability exists at multiple levels, including syntactic (consistent data structure) and semantic interoperability (shared understanding of meaning) (Lehne et al. 2019).

Knowledge: Synthesis, modeling, and validation

Only in very few situations, information is collected for its own sake. Normally, it is collected to gain knowledge, when information is synthesized, modeled, and validated to reveal actionable patterns, generalizations, or theoretical frameworks. The methods of obtaining the knowledge and the reasons for doing it depend on the situation under study. However, the concept of "learning" is applied for this purpose and conducted in day-to-day life, as part of a scientific study, or by data scientists through machine learning models (Jordan and Mitchell 2015; Lake et al. 2017).

From a probabilistic perspective, the acquisition of knowledge can be viewed as the *approximation* of the generative process $P(X)$ underlying the observed information in the domain D (Murphy 2022). When translating data into information as presented in the previous chapter, one just *implicitly defined* this process by the context in which the data was obtained. Learning is now the process of defining an explicit model $P(X|\Theta)$ characterized by parameters Θ instead (Bishop 2006). These explicit models might be established in a large number of ways like rules, mathematical equations, executable logic, or computational architectures. It could be a mental model encoded within the brain, physical models where only some parameters are updated, or complex machine learning models using their flexibility to adapt to the information (Lake et al. 2017). In each case, the model is an abstraction, a necessary simplification of reality through the mentioned assumptions about the underlying process and methodological considerations. The final process of learning the model (in that case defined as

linking data with models) may proceed by maximum likelihood estimation, Bayesian updating, or other inferential approaches (Bishop 2006; Murphy 2022).

Within the thesis, the acquisition of knowledge relies primarily on machine learning. Within the broader field of artificial intelligence, machine learning might be defined as a set of techniques by which knowledge is extracted directly from data (Jordan and Mitchell 2015). As an alternative, "classical" approaches to artificial intelligence emphasized explicit knowledge representation often by crafting logic rules, ontologies, and inference engines by hand (Russell and Norvig 2021). Those techniques could still be state-of-the-art for problems where explainability remains important (Samek et al. 2021). However, particularly since the 1980s machine learning approaches are getting more and more "information-driven" with the availability of large datasets and increases in computational capacity (LeCun et al. 2015). These models commonly shift away from explicit knowledge encoding to learning statistical associations, often without causal understanding (Pearl and Mackenzie 2018).

While the field of machine learning is highly diverse, in the following I focus on supervised machine learning. In this case, the final aim is to predict some quantity given potential covariates. More formally, besides the existing X, one may assume additional output data $Y = \{y_1, ..., y_n\}$ from a label space $\mathbb{Y}$. Then, given a training set $S = \{\mathbb{X}, \mathbb{Y}\}$ containing both the information and the associated labels, one may try to find a predictive function $f : \mathbb{X} \rightarrow \mathbb{Y}$, mapping the input data to the output label (Shalev-Shwartz and Ben-David 2014). In the probabilistic formulation from above, the model corresponds now to $p(Y|f, X, \Theta)$. The task the machine learning model should solve is then defined as $T = \{\mathbb{Y}, f\}$.

Machine learning and classical statistics share the reliance on these model-data connections but differ in their goals. While statistics seeks to draw robust population-level conclusions from observed data, machine learning aims principally at working with unseen data (Bzdok et al. 2018). With an appropriate model, the obtained performance should not only hold on existing samples but on new, dynamically changing inputs. This desired behavior corresponds to generalization (Belkin et al. 2019) in the sense commonly understood by humans.

Generalization and transfer learning

While the extraction and formalization of knowledge from data constitutes a critical step in science, the ultimate utility of such knowledge is determined by its ability to generalize (Yarkoni 2022; Shalev-Shwartz and Ben-David 2014). Generalization reflects the principle that established findings, when transported to new or broader contexts, retain their validity and predictive power. This capacity is foundational for sustainable scientific progress, enabling the expansion of theory-building based on reproducible and transferable results rather than isolated observations (Dwork et al. 2015). In medical science, for instance, the aim is not merely to describe phenomena within a single cohort or institution, but to achieve insights applicable across diverse patient populations, settings, and time frames (Beam and Kohane 2018; Varoquaux and Cheplygina 2022).

For humans, generalizing knowledge to novel circumstances is a fundamental cognitive ability, intuitively shaped by experience and reasoning (Lake et al. 2017). In contrast, achieving generalization in (computational) models remains a substantial challenge (Varoquaux and Cheplygina 2022). That is not necessarily a flaw of the underlying methods or algorithms. All models are inherently dependent on assumptions, assumed data distributions $P(X)$ and relationships embedded during the process, for example when contextualizing the raw data into information or when defining the task (Futoma et al. 2020). When these assumptions encoded in the domain D or predictive function f no longer hold, models are prone to fail. Two such changes with specific names in the literature are covariate shift where the distribution of X varies between settings and concept drift where the actual relationship f between features and labels changes over time or context (Moreno-Torres et al. 2012).

As a consequence, the implications of insufficient generalization are not limited to technical domains but contribute substantially to the broader reproducibility crisis in science (Beam and Kohane 2018). Numerous studies, particularly in preclinical biomedical research, have reported reproducibility failures in more than half of published findings (Begley and Ioannidis 2015; Yarkoni 2022; Beaulieu-Jones et al. 2024). These failures reflect the inability to replicate results when methods are applied to new cohorts or slightly altered protocols, highlighting the challenges in translating insights from controlled environments to real-world contexts (Camerer et al. 2018).

This lack of generalizability is especially problematic for machine learning models, for example in the medical domain. While modern approaches yield highly accurate predictive models on development data, their performance frequently deteriorates when deployed in clinical environments distinct from their original training context (Varoquaux and Cheplygina 2022). One underlying reason is the bias-variance dilemma (Belkin et al. 2019). Current advances in computational power and algorithmic sophistication enable models of substantial complexity. Many such models, including deep neural networks, are universal function approximators and might capture virtually any underlying signal present in the data. However, this flexibility entails a trade-off: highly complex models, if not appropriately constrained, may overfit to the noise of the training set (Shalev-Shwartz and Ben-David 2014). These nuances are commonly the differences between the implicitly assumed $P(X)$ and the "real" $P(X)$ when contextualizing the data. Then, the captured spurious associations and context-specific samples are artifacts that do not translate outside the original environment in which the data was collected (Futoma et al. 2020). The additional complexity of many computational models raises further issues of interpretability, as such "black box" models are often not comprehensible with human understanding (Samek et al. 2021). As a consequence, one must rely even more upon the quality of the input information. Acknowledging these risks, regulatory authorities including the Food and Drug Administration, European Medicines Agency, and the German Federal Institute for Drugs and Medical Devices are increasingly tightening requirements for evidence of generalizable and robust model performance, extending to post-market surveillance and ongoing validation (Aboy et al. 2024).

Besides data quality issues, additional and important sources of compromised generalization are methodological errors such as data leakage. Data leakage arises when information from outside the training data, which would be unavailable in a real-world application, inadvertently enters the modeling process (Kaufman et al. 2012). This mistake frequently leads to overestimated model performance, undermining the reliability of results when moving into new settings. Once such errors are excluded, different strategies could be utilized to support generalizability within data-driven medical research:

1. Efforts to ensure generalizability should already start on the level of the data. One may therefore try to improve the accessibility and interoperability of data, ensuring that samples used for training are representative and diverse.

2. One may limit the complexity of the model by incorporating domain knowledge while designing the model p to make it less dependent on the specific data X.
3. When more flexible models are necessary, transfer learning approaches should be considered. These models are able to "re-use" existing knowledge, for example derived from larger datasets than those available in the medical area. More formally, given a source domain D_s, a learning task T_s, a target domain D_t and another learning task T_t, transfer learning aims to improve the learning of the target predictive distribution f_t in D_t using the knowledge in D_s and T_s, where $D_s \neq D_t$ or $T_s \neq T_t$ (Zhuang et al. 2021). In *inductive* transfer learning, knowledge learned from one source task is used to improve performance on a different but related target task. In *transductive* transfer learning, the source and target tasks are the same while the domains are different.

Within this thesis, the benefits and limitation of these approaches for different data modalities in Parkinson's disease are demonstrated.

2 Narrative summary of methods and results

In general, the preceding background on the effective usage of data to gain generalizable knowledge is not specific to any disease. Within this thesis, the clinical application is centered on the field of neurology, focusing specifically on Parkinson's disease. First systematically described by James Parkinson in 1817, it is now established as the second most prevalent neurodegenerative disease globally, surpassed only by Alzheimer's disease (Giannakopoulou et al. 2022). More precisely, Parkinson's disease is a progressive neurodegenerative syndrome that primarily impairs motor function, but also presents a wide range of non-motor symptoms impacting nearly all aspects of daily life (Bloem et al. 2021). The increasing incidence of Parkinson's disease in recent decades is largely attributed to demographic shifts, particularly aging populations, as well as heightened clinical awareness and improved diagnostic capabilities (Ben-Shlomo et al. 2024).

The central symptoms of Parkinson's disease comprise classic motor features such as resting tremor, bradykinesia, muscular rigidity, and postural instability. These symptoms primarily arise from the degeneration of dopaminergic neurons in the substantia nigra, a midbrain region essential for motor control. While dopaminergic therapies such as Levodopa provide effective symptomatic relief, the progressive nature of the neurodegeneration contributes to the eventual development of motor fluctuations and complications, including dyskinesias, that complicate long-term management (Bloem et al. 2021). Importantly, Parkinson's disease also encompasses a diverse spectrum of non-motor symptoms, including cognitive impairment, mood and behavioral changes, autonomic dysfunction, sleep disturbances, pain, and various sensory deficits. Increasing understanding of these non-motor manifestations has underscored the clinical heterogeneity of Parkinson's disease and reaffirmed the importance of comprehensive patient evaluation (Vizcarra et al. 2019).

Recent progress in Parkinson's disease research and care has been closely linked to the increasing availability and diversity of data sources (Sigcha et al. 2023). These extend beyond traditional clinical observations to encompass biomolecular profiles, neuroimaging, digital sensor data, patient-reported outcomes, and administrative health records. As the volume and variety of observational studies and published datasets have grown, concerns have been raised regarding the alignment of such data with real-world clinical scenarios. Authors such as Beaulieu-Jones et al. (2024) have therefore advocated for additional investigations into the generalizability of research findings derived from these datasets. While each data type may provide evidence regarding diagnosis, monitoring, and the advancement of research, it is also presents new methodological and technical challenges.

Standartized neurological examinations remain foundational for diagnosis and disease monitoring. In general, they are conducted through the application of structured rating instruments such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) (Goetz et al. 2008) and the modified Hoehn and Yahr scale (Goetz et al. 2004). These instruments ensure consistent assessment of motor symptoms, are systematically documented in electronic health records, and promote longitudinal research as well as multi-center collaborations by providing semantic interoperability. Digital sensor and wearable data have emerged as valuable assets for continuous, objective monitoring of motor symptoms in naturalistic environments (Giannakopoulou et al. 2022). Wearable devices, incorporating accelerometry, gyroscopes, and smartphone-based sensors, facilitate detailed assessment of movement patterns, symptom fluctuations, and treatment effects outside the clinical setting (Del Din et al. 2021). These technologies simplify the development of digital endpoints for clinical trials and have opened new possibilities for remote patient monitoring and personalized care (Adams et al. 2021). Patient-reported outcome data capture aspects of disease experience that may not be apparent in clinical or sensor-based assessments (Vizcarra et al. 2019). Through structured questionnaires, diaries, and digital surveys, these data sources provide insights into non-motor symptoms, quality of life, emotional well-being, fatigue, and treatment satisfaction (Lee et al. 2022). Accordingly, patient-reported outcomes support more holistic, patient-centered models of care and constitute essential endpoints in many clinical investigations.

Beyond these data modalities, which constitute the primary focus of this thesis, other types of data are also actively researched but are not explored further here. As an

example, advances in molecular profiling have enhanced the collection and analysis of biomolecular and genetic data from blood, cerebrospinal fluid, and other biospecimens (Bloem et al. 2021). Possible measurements now include genetic variants, transcriptomic profiles, and protein biomarkers such as alpha-synuclein and neurofilament light chain. These biomarkers offer promise for improving diagnostics, characterizing disease subtypes, and identifying targets for new treatments. Large-scale initiatives and biobanks, such as the Parkinson's Progression Markers Initiative (Marek et al. 2011), are setting standards for the systematic collection and sharing of these biospecimens to support translational research. Similarly, structural and functional neuroimaging, including magnetic resonance imaging and dopamine transporter single photon emission computed tomography, provide detailed representations of brain anatomy and function.

The seven publications, highlighted in the following, focus on the three data types patient-reported outcomes, clinical data, and wearable-derived data described above. Specifically, these works examine how improvements in interoperability and data quality can enable the development of models with greater robustness in novel clinical settings, as well as more effective methods for reusing existing knowledge.

Sustainable research through improved interoperability

Building on the previously discussed background, it is evident that robust, generalizable conclusions in Parkinson's disease research depend on access to appropriately sized, interoperable datasets. Interoperability then enables the integration of data from diverse sources, mitigating local biases and broadening the scope of research insights akin to those generated by multicenter studies. Achieving this level of data harmonization in Parkinson's disease research necessitates systematic work at several interoperability layers which are described in four publications.

Establishing syntactic interoperability

Despite continual expansion in clinical data collection, the practical reusability of these datasets is impeded by shortcomings in syntactic interoperability. As introduced earlier, syntactic interoperability refers to the capacity of distinct information

systems to exchange data in standardized, structured formats suitable for automated processing. Achieving this requires strict adoption of standardized data formats, data schemas, and communication protocols. The mere digitization of clinical documents, for example by scanning paper forms or employing optical character recognition, does not inherently enforce data structure or adherence to recognized standards. Genuine syntactic interoperability is realized only when consistent, standardized conventions are applied, allowing systems to accurately interpret and exchange information without ambiguity.

Conveniently, in this specific context of Parkinson's disease, semantic interoperability may be inherently supported upon achieving syntactic interoperability for many clinical assessments. This is largely due to the widespread usage of internationally validated, standardized instruments for disease assessment. The previously mentioned Hoehn and Yahr scale, the MDS-UPDRS, and the MoCA are used as de facto standards in both research and clinical practice. These instruments systematically capture a broad spectrum of disease attributes, spanning motor, non-motor, and quality-of-life components.

This dissertation investigates two complementary strategies for enhancing interoperability of these established instruments. The first focuses on the prospective digital acquisition of patient-reported outcomes through systems that directly encode data into interoperable formats. The second explores retrospective structuring of existing unstandardized documents using vision language models as a potentially straightforward way for clinicians to obtain structured data without expert knowledge in information technology.

Prospective improvements (Publication 1)

The initial study within this thesis examines the design and deployment of a digital assessment system for the standardized collection of patient-reported outcomes, built to ensure seamless mapping to recognized standards for digital medicine. Although the adoption of digital questionnaires administered on tablets and smartphones is growing, notable barriers remain. Key assessment instruments, such as the MDS-UPDRS, are often subject to licensing that prohibits modification of their structure or

appearance. This restriction constrains interface design and excludes interface adaptation through common digital elements like dropdown menus and checkboxes. Furthermore, the heterogeneity of digital literacy in the Parkinson's disease population, in combination with frequent motor symptoms, introduces additional usability considerations.

Addressing these factors requires highly accessible digital interfaces that maintain strict adherence to validated instrument formats, while ensuring practical usability for individuals with variable motor and digital abilities. The designed and implemented approach within the publication leveraged consumer-grade tablets with high-contrast screens similar to eBook readers. The device, with a similar size to DIN A4 paper, preserved the familiar layout and interaction patterns of paper forms, while enabling digital storage and interoperability via standard health data interfaces, through a custom backend service for the device. A machine learning model running on the device mapped the handwritten digits to their digital counterpart and provided an endpoint compatible with the Fast Healthcare Interoperability Resources standard from Health Level 7 (HL7-FHIR) to facilitate both immediate usability and downstream interoperability.

To assess the feasibility and acceptance of this system, a usability study was conducted among individuals with Parkinson's disease at the University Medical Center Hamburg-Eppendorf. Participation was voluntary, anonymized, and conducted strictly in accordance with institutional guidelines and ethical requirements, including informed consent procedures. The study was reported as a scientific case according to the regulations of the Ärztekammer Hamburg and was waived from consultation with the Ethics Committee (2024-300491-WF). Participants were introduced to a tablet device and invited to complete selected fields of the assessment instrument, simulating typical real-world use. They were asked to navigate the tool, interact with both the device and stylus, and were encouraged to pose questions regarding any aspect of the technology or instrument. Subsequently, usability was assessed using the established *System Usability Scale* (SUS) in a validated German translation.

Results from nine participants, average age 69 years, indicated high usability, with an average SUS score of 83.01 (SD = 9.11). Participants generally found the tool approachable and effective, appreciating the similarity to paper-based documentation. However, important ergonomics-specific findings were identified: the tablet's flat form factor and stylus attachment required further adaptation for ease of use

among patients with impaired fine motor skills. These practical insights underscore the necessity for iterative refinement of consumer hardware and digital workflows to optimize accessibility in this population.

Despite these ergonomic challenges, participants generally expressed strong motivation to engage with the digital system and confidence in its utility. Mimicry of paper-based questionnaires was particularly valued, as it helped bridge familiarity and ensured user confidence. Accordingly, this pilot implementation demonstrates that high-quality, syntactically interoperable data collection can be successfully achieved in routine care with minimal disruption to established procedures.

Retrospective improvements (Publication 2)

For clinicians, the ability to efficiently extract structured data from existing clinical documentation could streamline workflows and support interoperable secondary data use if that is not associated with significantly more effort. Simplicity and minimal manual effort are essential for routine integration. Recent advances in multimodal neural networks with zero-shot learning capabilities appear promising for this purpose, as they can interpret both visual and textual information within documents without retraining or changes in paper-based processes.

The second publication examined the performance of these neural models for retrospectively structuring data from routine neurological assessments. The dataset included three types of documents: a summary form, detailed scores from the MDS-UPDRS, and the Montreal Cognitive Assessment (MoCA), in German, Arabic, and Russian, which varied in language due to patient needs. A subset of paper documents was processed, with sensitive data redacted, to create a controlled evaluation. Within it, key variables such as specific test scores and assessment items were manually annotated, digitized, and used as reference standards for evaluating model accuracy.

Model selection prioritized practical deployment on premise, requiring local operation on a single GPU, open-source code, and pre-trained weights for transparency and reproducibility. The models ranged from older architectures specialized for document understanding like Donut and BLIP to recent state-of-the-art systems such as MiniCPM-Llama3-2.5 and InternVL. Prompts were designed to directly ask for numerical values, such as “What is the value for ‘MOCA’?” or “What is the value of

the first test from the top?", with pattern-matching techniques employed when the models failed to output single, clear numbers.

The dataset comprised 18 reports from 17 patients, representing 24 annotated variables. Document heterogeneity, including free-text corrections, non-standard layouts, and multilingual content, posed substantial challenges. Despite the focus on seemingly simple extraction tasks, model performance was impaired by this document variability and complexity. The most accurate models achieved just over 77% accuracy overall; performance on complex, multi-language forms such as the MoCA was markedly lower. Accuracy was higher on simple summary documents; however, performance was still insufficient for widespread clinical application.

These results indicate that, although current neural models show potential, several obstacles remain for automatic, robust structuring of clinical documents. Addressing these will require the development of diverse, representative datasets for further model optimization and benchmarking. Additionally, evolving prompt designs and interaction workflows tailored to clinical documentation may yield further performance gains

Establishing semantic interoperability (Publication 3)

While the previously discussed technical solutions for the clinical assessments result in syntactic and semantic interoperability for information *X*, research involving sensor data for Parkinson's disease faces additional challenges. Wearable devices such as smartwatches and activity trackers generate large volumes of high-frequency motion data. However, significant heterogeneity in device types, measurement protocols, and data structures does not allow naive combination into a single research dataset.

The third study in this thesis set out to address this gap by developing an architecture for the aggregation and standardized management of wearable sensor data in Parkinson's disease research, guided by the FAIR principles of data management. The objective was to implement and validate a platform capable of harmonizing accelerometer data from different datasets and allowing real-time collection, enabling reproducible multi-site collaboration and advanced analytic approaches while adhering to established standards.

An initial literature review highlighted widespread reliance on proprietary and individualized data formats, limiting data sharing and scientific reproducibility; the few available public datasets lacked standardization. Therefore, an adaptable, interoperable solution for the processing and harmonization of the sensor data appears to be required. In response, a modular architecture was designed and implemented to support the extraction, transformation, and loading of existing sensor datasets and metadata. The resulting data model prioritized three core entities: patient, time-stamped acceleration measurements, and linked clinical assessments. Consistent with the principle of data minimization, only essential participant identifiers and allocation details were retained initially, with capacity for additional metadata. Sensor data were encoded to capture sampling frequency, body location, device specifics, and measurement units.

Clinical assessments were represented using resource-centric schemas flexible enough to support both standardized and tailored protocols, capturing the necessary context for aligning symptom labels with motion data streams. The entire software was designed for interoperability by providing an HL7-FHIR-compliant, open-source endpoint capable of efficient deployment and real-time data management. Validation demonstrated reliable handling of large-scale, high-resolution motion data, with accurate synchronization of sensor data and clinical records. In comparison to existing HL7-FHIR servers, the presented implementation was able to process significantly more samples. The platform has since been successfully applied in research later described in this thesis for the management and analysis of wearable data.

Accessible data for validating hypotheses (Publication 4)

While interoperability is essential for effective use of datasets, accessibility remains a critical challenge. The fourth publication addressed in this thesis details the development, deployment, and evaluation of a clinical research platform designed for secondary use of clinical data, considering not only technical and regulatory requirements but also the practical needs of end-users. This platform should bridge routine clinical documentation and the needs of clinicians for efficient, privacy-preserving hypothesis testing.

The platform *Datenhotel* was designed through close collaboration with clinical stakeholders, prioritizing use cases including rapid hypothesis testing and in-house validation of external research findings. Design and implementation explicitly adhered to local and federal data protection regulations, including Hamburg's legal framework for pseudonymized retrospective analysis (§ 12 HmbKHG). This was ensured through robust trustee and data security mechanisms.

Technically, the system was built upon a modular architecture incorporating the existing data integration center established for the German Medizininformatik-Initiative for interfacing with clinical systems and a metadata repository structured according to international standards like ISO/IEC 11179. Through a separate pseudonymization service and an optimized user interface, the platform ensures both compliance with data security standards and usability for clinicians not specialized in data science or medical informatics.

Demonstration of the platform's capabilities after implementation was conducted by replicating a research study focused on tremor subtypes in Parkinson's disease. Using comprehensive clinical documentation available within the local hospital's information systems, secondary data of 777 patients was extracted and analyzed. The analysis allowed replication of previous findings by confirming a significantly higher prevalence of action tremor among patients with rest tremor, compared to those without rest tremor. This closely mirrored the findings reported in larger, externally curated datasets.

In summary, the publication demonstrates that with appropriate technical, regulatory, and usability considerations, it is feasible to realize sustainable, interoperable, and accessible research data infrastructures. The developed platform provides an operational model for hypothesis testing with real-world data, validated through successful replication of clinically relevant findings and steady user adoption since the release of the plattform.

Ensuring generalizability on interoperable data

Given the previous considerations and improvements on interoperability, the harmonized data could serve as the basis for establishing and evaluating methods which

perform not only on the datasets they are trained on but could generalize well in clinical setups, too. If simply making more information X accessible through interoperability is not an option, one may improve the generalizability by focusing on the model $P(X)$ instead. At least two options are viable: One may limit the flexibility of the model to match the clinical intuition, which may lead to more robust learning by avoiding overfitting. Alternatively, one may utilize transfer learning setups to base $P(X)$ on more information than available through X alone. A summary of the separate publications for these two approaches can be found in the following.

Incorporating human knowledge in movement data (Publication 5)

The objective of the fifth publication was to investigate how different movement data representations, one grounded in clinical, task-specific knowledge and another rooted in automatic, large-scale time-series feature extraction, impacted the performance and generalizability of machine learning models designed for dyskinesia detection.

In clinical practice, the management of Parkinson's disease relied substantially on regular adjustment of dopaminergic therapy, during which levodopa-induced dyskinesia frequently emerged as a significant side effect requiring close surveillance. Conventional assessments depend on intermittent clinical evaluations using rating scales which capture only isolated moments in a patient's daily life. In contrast, commercially available wrist-worn accelerometers facilitated continuous monitoring and have the potential to enhance clinical ratings with objective and temporally rich movement data. While substantial research had explored wearable-based detection of tremor and bradykinesia, automated dyskinesia assessment remained less investigated despite its importance for therapeutic decision-making and early identification of adverse drug reactions.

This study addressed the requirement for efficient and interpretable data representations in the development of machine learning methods for dyskinesia detection. Two contrasting approaches were compared: a semantically informed technique that employed a principal component analysis of the raw sensor data and a subsequent biomechanical feature extraction, and an automatic large-scale time-series feature extraction method utilizing advanced computational tools. Models were initially devel-

oped and evaluated on comprehensive, multi-center public datasets made available by the Michael J. Fox Foundation, and subsequently tested on a new, prospectively collected clinical dataset for the study. This clinical dataset consisted of wrist-worn accelerometer measurements, each labeled by a physician during the inpatient stay of patients for therapy optimization.

The results indicated that both representation strategies achieved promising F_1 scores near 0.7 when evaluated on public datasets. However, their generalizability varied considerably. Models based on automatic feature extraction achieved strong results on the public datasets, which were relatively homogeneous and structured, but showed diminished performance on the more heterogeneous dataset from routine stays, suggesting that they were still prone to overfitting. In contrast, models using semantically derived representations, as well as hybrid models that combined semantic dimensionality reduction with automatic feature selection, demonstrated higher and more consistent performance in the new clinical context, too. The features extracted through the semantic pathway directly corresponded to observable movement patterns, allowing for transparent model validation and user trust. These findings supported the hypothesis that semantically informed models were less susceptible to overfitting and delivered features that were understandable by clinicians, an essential property for any tool influencing clinical decision-making.

In conclusion, the study delivered scientific evidence by confirming that models built upon human-understandable, task-specific representations were more resilient to clinical variability and dataset shifts than models relying solely on automatic feature extraction. The incorporation of clinical knowledge at the data preprocessing and feature engineering stage was shown to be crucial, not only for interpretability but also for achieving enhanced reliability and generalizability in clinical machine learning applications.

Using transfer learning to re-use existing knowledge

Unlike the previous case where a human-designed model led to better generalizability, using simpler models might not always be an option due to the complexity of the problem or missing human intuition. In this case, transfer learning represents a promising paradigm for addressing data quality challenges in clinical and research

informatics. This section details two complementary transfer learning strategies explored in the context of clinical data quality enhancement and activity recognition in Parkinson's disease cohorts.

Inductive transfer learning (Publication 6)

The sixth manuscript of this cumulative dissertation investigated solutions for missing data in clinical assessment data, specifically examining the potential of inductive transfer learning to support data quality in clinical routine. The study provided empirical evidence that models trained on well-annotated research cohorts can be adapted and effectively applied to the more heterogeneous and fragmented clinical routine datasets, which frequently suffer from high rates of missing values due to diverse documentation standards and practical constraints. Traditional approaches to imputation often treat the problem as a data cleaning step conducted after data collection. In contrast, this work integrated advanced imputation methods directly into the data acquisition pipeline, utilizing knowledge gained from comprehensive research studies. This integration aimed to improve not only the completeness but also the practical utility of real-world clinical datasets from the outset.

The primary training resource was a large-scale, systematically collected observational cohort. This dataset represented the target population for both model training and evaluation, taking advantage of the high data quality in controlled studies. Validation was subsequently conducted using routine clinical data from the University Medical Center Hamburg-Eppendorf, which was accessed through the "Datenhotel" established in the fourth publication. To further evaluate the performance, the project included additional study data as an additional benchmark, providing insight into how the models performed on even smaller sample sizes where classical imputation algorithms might not work sufficiently well.

The study evaluated inductive transfer learning by utilizing knowledge learned from one source task for a different but related target task while the source and target domains were kept the same. Methodologically, the selected model for this purpose was a self-supervised deep learning model tailored for representation learning. Once trained, it was adapted and evaluated on the evaluation datasets to assess the ability of transferred feature representations to support imputation. Alongside this primary

approach, the study also benchmarked several established imputation strategies, including deep generative models, tree-based ensemble methods, classical regression-based techniques, and simple univariate approaches, to provide a comprehensive comparative analysis.

The results demonstrated the viability of inductive transfer learning as a strategy for clinical data imputation. The self-supervised deep learning approach, when trained exclusively on research data and applied to both routine and external clinical datasets, achieved superior imputation performance. Specifically, it obtained a mean macro F1-score of 0.35, outperforming all comparator methods, including established techniques such as Multivariate Imputation by Chained Equations (MICE) and other machine learning-based approaches. These findings indicate that patterns and representations learned from well-curated research data can meaningfully enhance missing value handling in more noisy clinical settings. Unlike better interpretable but less flexible classical models, which remain valuable for training efficiency and are optimized for a single task only, the deep learning approach produced latent feature representations that could be easily employed for recycling knowledge.

Transductive transfer learning (Publication 7)

The seventh and final manuscript systematically investigated transductive transfer learning for the recognition of physical activities related to motor examinations in Parkinson's disease. Similar to the representations learned in the previous publication, the research leveraged foundation models pre-trained on large accelerometer datasets of healthy individuals, aiming to improve the recognition of motor examination activities in patients with Parkinson's disease, whose movement patterns are distinctly affected by their symptoms. Accordingly, the evaluation paradigm contained difference in populations (healthy versus Parkinson's disease), while maintaining the same activity recognition source and target task.

The study utilized three distinct datasets for assessing the generalizability between different recording setups. Besides two published study datasets with motor examinations, the collected movement data from the sixth publication was used, using the interoperable data structure developed in the third publication. Activity labels

were harmonized into clinically and practically relevant categories, including walking, sitting, rotating hands (an MDS-UPDRS examination item), and a general "other" activities class.

The foundational model used for transductive transfer learning was a publicly available deep learning model pre-trained on thousands of days of accelerometer data from healthy individuals. This model extracted generic movement features, assumed to be transferable to related tasks. Within the study, the model was fine-tuned using Parkinson's cohort data. Three adaptation strategies were analyzed: complete training from scratch with no external information; partial fine-tuning involving retraining only the last classification layers while freezing the remaining weights; and full fine-tuning allowing gradient descent to adjust all parameters of the model. Then, the performance on comparable samples through a cross-validation, and generalizability on the more clinical dataset were assessed.

The main findings indicated that models initialized with knowledge from the healthy source domain outperformed those trained solely on Parkinson's data. Models using either partial or full fine-tuning achieved test F1 scores near 0.7 on observational data, versus 0.55 achieved from scratch. This strongly supported the hypothesis that representations learned from healthy movements for the same task were beneficial, even when motor symptoms significantly altered movement patterns in the patient population. Importantly, the differences between fine-tuning strategies revealed subtle trade-offs: partial fine-tuning conferred greater stability, while full fine-tuning enabled higher peak performance but was more sensitive to hyperparameters such as the learning rate.

When assessing the generalizability, application to the more clinical dataset with differing sensor setups and annotation processes led to a drop in performance. Nevertheless, models benefiting from foundation model knowledge maintained higher F_1 scores (0.48) than models trained exclusively on Parkinson's disease data (0.33), supporting the efficacy of transductive transfer learning for managing domain discrepancies in clinical sensor data. This methodological approach thus represents a pragmatic and scalable pathway to translate sensor-based monitoring innovations from the research domain into clinical practice, increasing the reliability and accessibility of digital disease monitoring tools for real-world use.

3 Overall discussion

The digital transformation of society, marked by the escalating use of information technologies, has resulted in a dramatic increase in the volume, variety, and speed of health data generation. In medicine, these advances have generated both significant possibilities and new challenges. On one hand, the growing pool of high-frequency, heterogeneous health data offers the opportunity to deliver more accurate and individualized patient care, as well as to accelerate discovery within data-driven research. On the other hand, the process of converting these data into meaningful, robust, and clinically actionable knowledge remains challenging. The abundance of data does not guarantee actionable insight or knowledge; rather, it often introduces new issues, such as information overload, the propagation of noise, challenges in data harmonization, and risks to reproducibility and generalizability.

This cumulative dissertation examines, using Parkinson's disease as a primary example, how methodological, infrastructural, and analytical innovations can drive sustainable progress despite these emerging challenges. The findings from the seven manuscripts collectively address the core question of how the diversity and accessibility of health data in Parkinson's research can be leveraged to produce reliable, transferable, and clinically useful models. The discussion that follows reflects on these dimensions, focusing on interoperability and accessibility, robust and generalizable modeling, the role of transfer learning, and broader implications for medical informatics.

Interoperability and accessibility

The transformation of health data into actionable knowledge requires not only adequate data volume and intrinsic quality, but also effective integration across systems

and contexts. As highlighted early in this thesis, both data quality and interoperability are essential for supporting meaningful data reuse. Despite the rapid expansion of digital health data collection, challenges in the aggregation and analysis of information from disparate sources still limit the scalability and efficacy of both research efforts and clinical care. This work focused on two data modalities: structured questionnaires and movement data obtained through wearable devices. Each modality demonstrates distinct complexities in standardization, interoperability, and usability.

For structured data based on established assessment instruments, achieving semantic interoperability appears relatively feasible. For instance, the utilized MDS-UPDRS benefits from detailed item descriptions, interpretive guidelines, and partly existing concepts in standards like Logical Observation Identifiers Names and Codes (LOINC) and HL7 FHIR. Nevertheless, the three manuscripts covering this topic within the thesis consistently demonstrate that digital health data attains its greatest value when standardization and user-centered system design are emphasized early in the design of data collection.

The second manuscript specifically examined the challenge of retrospective interoperability. Since extensive clinical information still resides in paper-based or heterogeneously formatted digital archives, extracting structured data from these sources is a promising strategy. Contemporary vision-language models were evaluated for their potential to transform unstructured clinical documentation into interoperable, structured data. However, the study identified substantial limitations in the performance of current artificial intelligence-based methods, particularly in handling the diversity of layout, content, and language that characterizes real-world clinical documents. Despite their theoretical promise, practical implementation is hindered by inconsistencies in document structure and linguistic heterogeneity, limiting the reliability of large-scale aggregation from routine care records. Thus, simply introducing an automated digitalization step to existing analogue workflows, even with state-of-the-art systems, may be insufficient for producing high-quality, standardized datasets.

Acknowledging the need for digital data, many processes in a modern hospital might already involve the digital collection of data, often by involving humans manually inputting the information. However, technical, regulatory, and organizational barriers can render even those structured datasets inaccessible to practitioners and researchers. The fourth manuscript addressed these challenges by implementing

and evaluating a comprehensive operational platform for clinical research named the "Datenhotel". While the platform primarily enhances data accessibility, it does not enforce interoperability, as the responsibility for form design and use of standards remains with the clinicians at the university medical center. Nonetheless, when clinicians utilize established instruments and standards regarding Parkinson's disease, retrospective access to routine care data enabled research with cohort sizes rivaling the largest internationally available datasets. Consistent with international publications, this dataset allowed the successful replication of established epidemiological findings such as tremor subtype prevalence among local Parkinson's patients and demonstrated the platform's validity for use in research. During the evaluation, usability, compliance, and sensitivity to institutional culture emerged as critical factors influencing clinician acceptance. Ongoing use indicates that many clinicians in non-neurological specialties become aware of data quality concerns only at the stage of data analysis, which is often too late for effective intervention. Early enforcement of semantic interoperability requirements could alleviate these issues.

An example of such a de-novo design of interoperable data collection is highlighted in the first publication. Unlike the clinical assessments conducted in the previous parts, the pipeline focused on patient-reported outcomes in routine care. Rather than introducing entirely new digital interfaces, the study implemented a digital questionnaire that mirrored the familiar paper-based instrument. The hand-written digits were analyzed on the device and transformed into HL7 FHIR resources. During its evaluation, this approach facilitated high usability among Parkinson's patients while ensuring strict adherence to internationally recognized standards for interoperability. This compliance enabled immediate interoperability by removing the need for extensive post-collection harmonization. Therefore, interoperability can be achieved at minimal additional cost when the physical and cognitive needs of both patients and clinicians are adequately considered in the data acquisition system's design.

Collectively, these findings highlight that the acquisition of high-quality, sustainable information for research in Parkinson's disease and likely beyond depends on careful prospective design of digital systems. When data are prospectively collected in a standards-adherent digital form, it might be more robust for further analysis than if it is harmonized after-the-fact from legacy records. This even holds for well-adopted instruments like the MDS-UPDRS which ensure some level of semantic interoperability. It is important to recognize that simply digitizing structured forms does not guaran-

tee high data quality or compliance. Usability considerations, tailored to the needs of both clinicians and patients as end-users, remain vital.

Interoperability takes on even greater complexity in the context of wearable sensor data and high-frequency time series, as detailed in the third manuscript. Unlike standardized clinical assessment scales with clear definitions and community consensus, wearable datasets currently lack universally accepted ontologies and harmonized acquisition protocols. The lack of both syntactic and semantic standards complicates aggregation, synchronization, and analysis, as reflected in the literature. The platform developed and assessed in this work demonstrates that robust architectures are capable of handling large-scale wearable data streams and can comply with HL7 FHIR. By making major published datasets compatible, the system allows for joint analyses. However, current international health data standards do not yet fully accommodate the detailed metadata and contextual specificity needed for advanced analysis of motion data.

In summary, the studies within this thesis demonstrate that achieving high-quality, actionable health data in Parkinson's disease research is an interdisciplinary challenge. Success requires not only technical solutions, but also consideration of interoperability, accessibility, privacy, usability, and long-term sustainability. The experience gained through developing and applying semantic standards, emphasizing user-centered design, and building robust digital infrastructures underscore the potential to greatly enhance data quality, utility, and ultimately clinical impact. At the same time, persistent challenges, ranging from legacy information technology systems and regulatory limitations to human factors and emerging data types, underline the urgent need for ongoing, collaborative efforts among patients, clinicians, researchers, IT experts, and policymakers.

Learning from interoperable data

To highlight the benefits of using this interoperable structured and unstructured data, the second part the thesis focuses on using those standartized datasets for tackling the initially described challenge of developing models that not only achieve high performance in narrow settings but also generalize to real-world, heterogeneous clinical

environments. The proliferation of machine learning in medicine has led to expectations that larger datasets and improved algorithmic sophistication will translate directly into better diagnostic and prognostic tools. As already discussed in the introduction, these expectations often fail when exposed to the realities of clinical practice; their utility drastically declines in the face of real-world complexity, evolving data distributions, and population heterogeneity. As the capabilities and complexities of the models appear to grow faster than the amount of interoperable data, overfitting the model to the sparse available samples is likely the primary reason for these performance issues.

One countermeasure might be the systematic inclusion of clinical and domain knowledge throughout the modeling pipeline in order to reduce the complexity of the utilized model. Beyond analytic robustness, models informed by clinical knowledge may provide much-needed trust, transparency, and interpretability. Such attributes are prerequisites for regulatory acceptance and adoption in clinical workflows. Results in the fifth manuscript provide evidence for this claim when interoperable movement data from observational studies to automatically assess dyskinesia was used. The machine learning model based upon features derived from clinical experience showed a significantly better generalizability when evaluated on the novel test dataset. In contrast, models based solely on data-driven, large-scale feature sets were less robust to deviations and distributional shifts present in the patient population. Accordingly, models aligned with domain understanding, causal relationships, and real-world constraints exhibit less overfitting and greater resilience to variability.

While integration of expert knowledge can mitigate some issues regarding generalizability even for unstructured sensor data, limiting the complexity of the model may be insufficient when the problem of interest is particularly complex and human intuition is missing. Then, transfer learning on the interoperable data arises as a promising paradigm.

The inductive transfer learning employed in the sixth publication demonstrated benefits in addressing the problem of missing data in clinical assessments. When the learned representations on the interoperable research datasets were used for imputing missing values, that setup demonstrated superior accuracy in comparison to state-of-the-art imputation methods. By learning from the rich structure of research data, these models appear to capture complex dependencies among variables that would

otherwise not be observable in smaller or noisier clinical datasets. This is particularly interesting as the performance of the imputation is independent of the size of the dataset which is required to be imputed and can be conducted online.

Finally, transductive transfer learning on interoperable data allowed pre-trained models to adapt to new populations. The seventh publication showed that fine-tuning foundation models previously trained on accelerometer data from healthy participants significantly improved the classification of movements associated with assessments, despite the inherent differences in movement patterns between both populations. These benefits of using this pre-training even hold when the models fine-tuned on study data were applied to a clinical test set.

In summary, the results demonstrate that transfer learning methods built upon interoperable datasets consistently outperform models trained exclusively on smaller or isolated datasets. By using interoperable datasets, which become exchangeable, it becomes possible to reuse knowledge efficiently even in clinical environments, where data might be messier and more fragmented than in controlled observational studies. However, transfer learning also imposes new requirements for the maintenance and monitoring of models over time. As clinical practices, patient demographics, and technologies evolve, adaptation strategies must be continually refined. Additionally, the complexity and opacity of models produced by deep learning can pose challenges for interpretability, clinical auditability, and regulatory compliance. Accordingly, the case-by-case comparison with simpler models based on the intuition of clinical experts similar to the fifth publication remains necessary.

Limitations and future work

While this thesis demonstrates the potential of technology to advance interoperability, accessibility, and analytical use of health data for Parkinson's disease, several limitations require further consideration.

First, the studies predominantly relied on integrating large observational datasets with smaller datasets generated in prospective research settings. Although these approaches facilitate proof-of-concept and methodological exploration, resulting sample sizes may still limit the generalizability of findings across the broader patient

population. Furthermore, the thesis focused on structured questionnaire data and sensor-based movement data as principal sources of information, but did not comprehensively address other potentially informative data modalities. In clinical practice, the management and understanding of Parkinson's disease often necessitate the integration of diverse data types, including medical imaging, genetic or genomic profiles, laboratory test results, and narrative clinical documentation. Future work should therefore strive to integrate these multimodal data sources to enable more comprehensive modeling of disease progression, patient subtypes, and therapeutic response. The ongoing development of machine learning methodology and advancements in data standards offer opportunities to combine cross-modal health data while maintaining standards of interoperability and data privacy. Extension of the analysis to encompass the full range of routinely collected clinical data could significantly enhance the scope and relevance of digital health approaches in Parkinson's disease.

Second, sustainable integration of the proposed technological solutions into hospital settings require attention not only to technical challenges but also to regulatory, legal, and organizational aspects. Factors such as data protection regulations, institutional policies, and established clinical workflows can substantially constrain the adoption and scalability of digital systems. Opportunities may arise within national and international legal frameworks, including recent provisions supporting in-house digital health developments in Germany and Europe. However, these provisions may not easily translate to cross-institutional usage. The considerable variability in data accessibility, quality, and interoperability across different hospitals further emphasizes the need for validation in multi-center and multi-national studies. Such efforts are critical to ensure the generalizability, effectiveness, and scalability of the approaches proposed in this thesis.

A critical technical challenge that remains is the lack of widely adopted standards for wearable sensor data and high-frequency time series data. The absence of robust, community-supported ontologies and metadata schemas complicates the reliable aggregation, reuse, and comparison of wearable data across studies and clinical sites. Future work should focus on contributing to international interoperability initiatives, such as LOINC for standardized questionnaire coding, and on the continued development of analytical and data integration methods for time-dependent clinical data. Advancing these areas could substantially facilitate multi-site studies and accelerate knowledge transfer.

Finally, although this thesis placed emphasis on technical outcomes such as data quality, interoperability, and predictive performance, the true clinical value of the proposed digital health systems will be determined by their impact on healthcare delivery and patient outcomes in real-world settings. Evaluation within routine clinical environments is essential to assess the robustness and adaptability of algorithms in the presence of domain shifts, non-stationary data, and evolving clinical documentation standards.

In summary, the advances described in this thesis represent foundational steps toward improving digital health infrastructures for Parkinson's disease. Future research should address the stated limitations by incorporating larger and more representative data collections, extending to additional data modalities, enhancing interoperability across diverse clinical environments, and rigorously assessing real-world clinical impacts.

4 Conclusion

The body of work presented in this cumulative dissertation offers both practical solutions and theoretical insights for maximizing the value of large, heterogeneous datasets in medical informatics. Starting from foundational improvements in interoperability for Parkinson's disease research, the studies demonstrate that systematic structuring and harmonization of clinical and sensor data are both achievable and impactful. Interoperability enables not only improved research efficiency but also the replication and extension of scientific findings across diverse clinical contexts.

Through a focus on Parkinson's disease as a model disorder, the thesis demonstrates in practical terms how investments in interoperability and accessibility, realized through standards-based digital tools, translational research platforms, and stakeholder engagement, enable analytic approaches, including transfer learning, that are resilient in the face of heterogeneity. Human expertise remains central; models informed by domain understanding, and augmented but not supplanted by scalable machine learning techniques, offer a resilient path toward reliable and generalizing clinical deployment. Semantically-informed feature engineering and robust transfer learning approaches each address the limitations associated with raw data volume and complexity, moving the field closer to models that remain accurate, transparent, and trustworthy when they are deployed beyond their initial development context.

Despite the substantial progress documented here, continued evolution of standards, data handling frameworks, and analytic methodologies is necessary. Areas such as advanced multimodal data integration, ongoing benchmarking of language and vision models in clinical settings, and systematic assessment of real-world impact will require sustained attention. Organizational and regulatory frameworks must also evolve to support responsible data reuse, sufficient transparency, and reliable monitoring of models after deployment.

5 Bibliography

Aboy, Mateo, Timo Minssen, and Effy Vayena. 2024. "Navigating the EU AI Act: Implications for Regulated Digital Medical Products." *Npj Digital Medicine* 7 (1): 237. https://doi.org/10.1038/s41746-024-01232-3.

Adams, Jamie L., Karlo J. Lizarraga, Emma M. Waddell, et al. 2021. "Digital Technology in Movement Disorders: Updates, Applications, and Challenges." *Current Neurology and Neuroscience Reports* 21 (4): 16. https://doi.org/10.1007/s11910-021-01101-6.

Alliende, Luz Maria, Teresa Vargas, and Vijay Anand Mittal. 2023. "Representation Challenges in Large Clinical Datasets." *Schizophrenia Bulletin* 49 (6): 1414–17. https://doi.org/10.1093/schbul/sbad109.

Beam, Andrew L., and Isaac S. Kohane. 2018. "Big Data and Machine Learning in Health Care." *JAMA* 319 (13): 1317. https://doi.org/10.1001/jama.2017.18391.

Beaulieu-Jones, Brett K., Francesca Frau, Sylvie Bozzi, et al. 2024. "Disease Progression Strikingly Differs in Research and Real-World Parkinson's Populations." *Npj Parkinson's Disease* 10 (1): 1–11. https://doi.org/10.1038/s41531-024-00667-5.

Begley, C. Glenn, and John P.A. Ioannidis. 2015. "Reproducibility in Science: Improving the Standard for Basic and Preclinical Research." *Circulation Research* 116 (1): 116–26. https://doi.org/10.1161/CIRCRESAHA.114.303819.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. "Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off." *Proceedings of the National Academy of Sciences* 116 (32): 15849–54. https://doi.org/10.1073/pnas.1903070116.

Bellazzi, Riccardo. 2014. "Big Data and Biomedical Informatics: A Challenging Opportunity." *Yearbook of Medical Informatics* 23 (01): 08–13. https://doi.org/10.15265/IY-2014-0024.

Benchimol, Eric I., Liam Smeeth, Astrid Guttmann, et al. 2015. "The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement." *PLOS Medicine* 12 (10): e1001885. https://doi.org/10.1371/journal.pmed.1001885.

Ben-Shlomo, Yoav, Sirwan Darweesh, Jorge Llibre-Guerra, Connie Marras, Marta San Luciano, and Caroline Tanner. 2024. "The Epidemiology of Parkinson's Disease." *The Lancet* 403 (10423): 283–92. https://doi.org/10.1016/S0140-6736(23)01419-8.

Bian, Jiang, Tianchen Lyu, Alexander Loiacono, et al. 2020. "Assessing the Practice of Data Quality Evaluation in a National Clinical Data Research Network Through a Systematic Scoping Review in the Era of Real-World Data." *Journal of the American Medical Informatics Association* 27 (12): 1999–2010. https://doi.org/10.1093/jamia/ocaa245.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Information science and statistics. Springer.

Bloem, Bastiaan R, Michael S Okun, and Christine Klein. 2021. "Parkinson's Disease." *The Lancet* 397 (10291): 2284–303. https://doi.org/10.1016/S0140-6736(21)00218-X.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. 2018. "Statistics Versus Machine Learning." *Nature Methods* 15 (4): 233–34. https://doi.org/10.1038/nmeth.4642.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–44. https://doi.org/10.1038/s41562-018-0399-z.

Cheng, Cindy, Luca Messerschmidt, Isaac Bravo, et al. 2024. "A General Primer for Data Harmonization." *Scientific Data* 11 (1): 152. https://doi.org/10.1038/s41597-024-02956-3.

Cover, Thomas M., and Joy A. Thomas. 2005. *Elements of Information Theory*. Wiley. https://doi.org/10.1002/047174882X.

Degtiar, Irina, and Sherri Rose. 2023. "A Review of Generalizability and Transportability." *Annual Review of Statistics and Its Application* 10 (1): 501–24. https://doi.org/10.1146/annurev-statistics-042522-103837.

Del Din, Silvia, Cameron Kirk, Alison J. Yarnall, Lynn Rochester, and Jeffrey M. Hausdorff. 2021. "Body-Worn Sensors for Remote Monitoring of Parkinson's Disease

Motor Symptoms: Vision, State of the Art, and Challenges Ahead." *Journal of Parkinson's Disease* 11 (Suppl 1): S35–47. https://doi.org/10.3233/JPD-202471.

Diaz, Fernando, and Michael Madaio. 2024. "Scaling Laws Do Not Scale." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (October): 341–57. https://doi.org/10.1609/aies.v7i1.31641.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. 2015. "Generalization in Adaptive Data Analysis and Holdout Reuse." In *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/bad5f33780c42f2588878a9d07405083-Paper.pdf.

Floridi, Luciano. 2011. *The philosophy of information*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199232383.001.0001.

Frické, Martin. 2009. "The Knowledge Pyramid: A Critique of the DIKW Hierarchy." *Journal of Information Science* 35 (2): 131–42. https://doi.org/10.1177/0165551508094050.

Futoma, Joseph, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. "The Myth of Generalisability in Clinical Research and Machine Learning in Health Care." *The Lancet Digital Health* 2 (9): e489–92. https://doi.org/10.1016/S2589-7500(20)30186-2.

Giannakopoulou, Konstantina-Maria, Ioanna Roussaki, and Konstantinos Demestichas. 2022. "Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review." *Sensors* 22 (5): 1799. https://doi.org/10.3390/s22051799.

Goetz, Christopher G., Werner Poewe, Olivier Rascol, et al. 2004. "Movement Disorder Society Task Force Report on the Hoehn and Yahr Staging Scale: Status and Recommendations." *Movement Disorders* 19 (9): 1020–28. https://doi.org/10.1002/mds.20213.

Goetz, Christopher G., Barbara C. Tilley, Stephanie R. Shaftman, et al. 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): 2129–70. https://doi.org/10.1002/mds.22340.

Goldstein, Benjamin A, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. 2017. "Opportunities and Challenges in Developing Risk Prediction Models with

Electronic Health Records Data: A Systematic Review." *Journal of the American Medical Informatics Association* 24 (1): 198–208. https://doi.org/10.1093/jamia/ocw042.

Greshake Tzovaras, Bastian, Misha Angrist, Kevin Arvai, et al. 2019. "Open Humans: A Platform for Participant-Centered Research and Personal Data Exploration." *GigaScience* 8 (6): giz076. https://doi.org/10.1093/gigascience/giz076.

Hilbert, Martin. 2016. "Big Data for Development: A Review of Promises and Challenges." *Development Policy Review* 34 (1): 135–74. https://doi.org/10.1111/dpr.12142.

Hilbert, Martin, and Priscila López. 2011. "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332 (6025): 60–65. https://doi.org/10.1126/science.1200970.

Huckvale, Kit, Svetha Venkatesh, and Helen Christensen. 2019. "Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety." *Npj Digital Medicine* 2 (1): 88. https://doi.org/10.1038/s41746-019-0166-1.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60. https://doi.org/10.1126/science.aaa8415.

Kaufman, Shachar, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data* 6 (4): 1–21. https://doi.org/10.1145/2382577.2382579.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences* 40: e253. https://doi.org/10.1017/S0140525X16001837.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Lee, JuHee, Insun Yeom, Misook L. Chung, Yielin Kim, Subin Yoo, and Eunyoung Kim. 2022. "Use of Mobile Apps for Self-Care in People With Parkinson Disease: Systematic Review." *JMIR mHealth and uHealth* 10 (1): e33944. https://doi.org/10.2196/33944.

Lehne, Moritz, Julian Sass, Andrea Essenwanger, Josef Schepers, and Sylvia Thun. 2019. "Why Digital Medicine Depends on Interoperability." *Npj Digital Medicine* 2 (1): 79. https://doi.org/10.1038/s41746-019-0158-1.

Lenert, Matthew C, Michael E Matheny, and Colin G Walsh. 2019. "Prognostic Models Will Be Victims of Their Own Success, Unless. . . ." *Journal of the American Medical Informatics Association* 26 (12): 1645–50. https://doi.org/10.1093/jamia/ocz145.

Lewis, Abigail E, Nicole Weiskopf, Zachary B Abrams, et al. 2023. "Electronic Health Record Data Quality Assessment and Tools: A Systematic Review." *Journal of the American Medical Informatics Association* 30 (10): 1730–40. https://doi.org/10.1093/jamia/ocad120.

Marcus, J. Scott, Bertin Martens, Christophe Carugati, Anne Bucher, and Ilsa Godlovitch. 2022. "The European Health Data Space." *SSRN Electronic Journal*, ahead of print. https://doi.org/10.2139/ssrn.4300393.

Marek, Kenneth, Danna Jennings, Shirley Lasch, et al. 2011. "The Parkinson Progression Marker Initiative (PPMI)." *Progress in Neurobiology* 95 (4): 629–35. https://doi.org/10.1016/j.pneurobio.2011.09.005.

Moreno-Torres, Jose G., Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. "A Unifying View on Dataset Shift in Classification." *Pattern Recognition* 45 (1): 521–30. https://doi.org/10.1016/j.patcog.2011.06.019.

Murphy, Kevin P. 2022. *Probabilistic machine learning: an introduction*. Adaptive computation and machine learning. The MIT Press.

Näher, Anatol-Fiete, Carina N Vorisek, Sophie A I Klopfenstein, et al. 2023. "Secondary Data for Global Health Digitalisation." *The Lancet Digital Health* 5 (2): e93–101. https://doi.org/10.1016/S2589-7500(22)00195-9.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. First edition. Basic Books.

Piwek, Lukasz, David A. Ellis, Sally Andrews, and Adam Joinson. 2016. "The Rise of Consumer Health Wearables: Promises and Barriers." *PLOS Medicine* 13 (2): e1001953. https://doi.org/10.1371/journal.pmed.1001953.

Prasser, Fabian, Oliver Kohlbacher, Ulrich Mansmann, Bernhard Bauer, and Klaus Kuhn. 2018. "Data Integration for Future Medicine (DIFUTURE): An Architectural and Methodological Overview." *Methods of Information in Medicine* 57 (S 01): e57–65. https://doi.org/10.3414/ME17-02-0022.

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. 2019. "Machine Learning in Medicine." *New England Journal of Medicine* 380 (14): 1347–58. https://doi.org/10.1056/NEJMra1814259.

Roski, Joachim, George W. Bo-Linn, and Timothy A. Andrews. 2014. "Creating Value In Health Care Through Big Data: Opportunities And Policy Implications." *Health Affairs* 33 (7): 1115–22. https://doi.org/10.1377/hlthaff.2014.0147.

Rowley, Jennifer. 2007. "The Wisdom Hierarchy: Representations of the DIKW Hierarchy." *Journal of Information Science* 33 (2): 163–80. https://doi.org/10.1177/0165551506070706.

Russell, Stuart J., and Peter Norvig. 2021. *Artificial intelligence: a modern approach*. Fourth Edition. With Ming-wei Chang, Jacob Devlin, Anca Dragan, et al. Pearson Series in Artificial Intelligence. Pearson.

Samek, Wojciech, Gregoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Muller. 2021. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications." *Proceedings of the IEEE* 109 (3): 247–78. https://doi.org/10.1109/JPROC.2021.3060483.

Sedlakova, Jana, Paola Daniore, Andrea Horn Wintsch, et al. 2023. "Challenges and Best Practices for Digital Unstructured Data Enrichment in Health Research: A Systematic Narrative Review." *PLOS Digital Health* 2 (10): e0000347. https://doi.org/10.1371/journal.pdig.0000347.

Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding machine learning: from theory to algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019.

Sherman, Rachel E., Steven A. Anderson, Gerald J. Dal Pan, et al. 2016. "Real-World Evidence — What Is It and What Can It Tell Us?" *New England Journal of Medicine* 375 (23): 2293–97. https://doi.org/10.1056/NEJMsb1609216.

Sigcha, Luis, Luigi Borzì, Federica Amato, et al. 2023. "Deep Learning and Wearable Sensors for the Diagnosis and Monitoring of Parkinson's Disease: A Systematic Review." *Expert Systems with Applications* 229 (November): 120541. https://doi.org/10.1016/j.eswa.2023.120541.

Sperrin, Matthew, David Jenkins, Glen P Martin, and Niels Peek. 2019. "Explicit Causal Reasoning Is Needed to Prevent Prognostic Models Being Victims of Their Own Success." *Journal of the American Medical Informatics Association* 26 (12): 1675–76. https://doi.org/10.1093/jamia/ocz197.

Swan, Melanie. 2013. "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery." *Big Data* 1 (2): 85–99. https://doi.org/10.1089/big.2012.0002.

Szarfman, Ana, Jonathan G. Levine, Joseph M. Tonning, et al. 2022. "Recommendations for Achieving Interoperable and Shareable Medical Data in the USA." *Communications Medicine* 2 (1): 86. https://doi.org/10.1038/s43856-022-00148-x.

Topol, Eric J. 2019. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence." *Nature Medicine* 25 (1): 44–56. https://doi.org/10.1038/s41591-018-0300-7.

Varoquaux, Gaël, and Veronika Cheplygina. 2022. "Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future." *Npj Digital Medicine* 5 (1): 48. https://doi.org/10.1038/s41746-022-00592-y.

Vizcarra, Joaquin A., Álvaro Sánchez-Ferro, Walter Maetzler, et al. 2019. "The Parkinson's Disease e-Diary: Developing a Clinical and Research Tool for the Digital Age." *Movement Disorders* 34 (5): 676–81. https://doi.org/10.1002/mds.27673.

Wicks, Paul, Michael Massagli, Jeana Frost, et al. 2010. "Sharing Health Data for Better Outcomes on PatientsLikeMe." *Journal of Medical Internet Research* 12 (2): e19. https://doi.org/10.2196/jmir.1549.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Yarkoni, Tal. 2022. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: e1. https://doi.org/10.1017/S0140525X20001685.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, et al. 2021. "A Comprehensive Survey on Transfer Learning." *Proceedings of the IEEE* 109 (1): 43–76. https://doi.org/10.1109/JPROC.2020.3004555.

Zins, Chaim. 2007. "Conceptual Approaches for Defining Data, Information, and Knowledge." *Journal of the American Society for Information Science and Technology* 58 (4): 479–93. https://doi.org/10.1002/asi.20508.

6 Publications

German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI240870

Digitalizing Handwritten Digits of Patients with Parkinson's Disease Utilizing Consumer Hardware and Open-Source Software

Christopher GUNDLER[a,1], Alexander Johannes WIEDERHOLD[a], and Monika PÖTTER-NERGER[b]
[a] *Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[b] *Department of Neurology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
ORCiD ID: Christopher Gundler https://orcid.org/0000-0001-9301-8872

Abstract. Introduction Parkinson's disease represents a burdensome condition with complex manifestations. A licensed, standardized paper-based questionnaire is completed by both patients and physicians to monitor the progression and state of the disease. However, integrating the obtained scores into digital systems still poses a challenge. **Methods** Paper-based handwriting is intuitive and an efficient mode of human-computer interaction. Accordingly, we transformed a consumer-grade tablet into a device where an exact digital copy of the disease-specific questionnaire can be filled with the supplied pen. Utilizing a small convolutional neural network directly on the device and trained on MNIST data, we translated the handwritten digits to appropriate LOINC codes and made them accessible through a FHIR-compatible HTTP interface. **Results** When evaluating the usability from a patient-centric point of view, the System Usability Score revealed an excellent rating (SUS = 83.01) from the participants. However, we identified some challenges associated with the magnetic pen and the flat design of the device. **Conclusion** In setups where certified medical devices are not required, consumer hardware can be used to map handwritten digits of patients to appropriate medical standards without manual intervention through healthcare professionals.

Keywords patient-reported outcome; usability testing; Unified Parkinson's Disease Rating Scale; Parkinson's disease; human-computer interaction

1. Introduction

Parkinson's disease (PD) represents a central challenge in modern healthcare, posing a significant burden on both individuals and society getting older at large [1]. Its complex manifestations, ranging from motor impairments to cognitive and psychological symptoms, render it a multifaceted condition to manage [2]. Advancements in both

[1] Corresponding Author: Christopher Gundler, c.gundler@uke.de, Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany

pharmacological and device-assisted therapy could nowadays significantly reduce the suffering of those affected [3]. However, this medical progress requires constant and labor-intensive adjustments often over decades.

Central to the systematic evaluation of Parkinson's disease progression and therapeutic efficacy are standardized assessments, which serve as indispensable tools in clinical practice and research endeavors. Among these, the Hoehn & Yahr scale, the Unified Parkinson's Disease Rating Scale (UPDRS), and its revised version by the Movement Disorder Society (MDS-UPDRS) [4] stand out as gold standards due to their widespread international adoption and rigorous validation. These instruments offer a comprehensive framework for assessing the multifaceted manifestations, encompassing motor and non-motor experiences of daily living as well as direct evaluation of motor complications. Importantly, they facilitate a holistic evaluation by integrating inputs not only from clinicians but also from patients, recognizing the subjective experiences and perspectives that are integral to understanding the impact of the disease in daily life.

Digital technologies and intelligent systems for corresponding disease management range from hand-derived algorithms over traditional machine learning approaches to deep learning architectures, offering novel options for enhanced accuracy, objectivity, and efficiency [5–7]. Despite these magnitudes of possible modalities and technological advancements, challenges remain in integrating these systems into clinical practice [8]. Issues such as data privacy, algorithm robustness, and regulatory compliance necessitate careful consideration and do not yet allow usage in a medical context [9]. Accordingly, the utilization of such assessments remains predominantly paper-based, limiting scalability, efficiency, and accessibility.

While there has been a trend towards deploying questionnaires on digital devices such as tablets or smartphones in both clinical and patient-facing contexts [10], several challenges persist, hindering seamless integration and adoption of digitalized MDS-UPDRS questionnaires. Firstly, in contrast to personalized surveys, the questionnaire is an original work that has been approved, validated and licensed for standardized severity assessments during clinical routine. It is against the license agreement to alternate the mode of administration or to interfere with its general structure. Thus, alternative appearances, such as dropdown menus or radio buttons, are impossible to use, making the questionnaire's design less flexible for developers [10]. Secondly, the inherent demographic diversity of Parkinson's patients [1] translates into varying levels of digital literacy and proficiency in using electronic devices. Some patients may struggle to navigate complex interfaces or may be apprehensive about using unfamiliar technology, posing barriers to effective questionnaire completion and data collection. Finally, the lack of standardized data exchange formats, exemplified by the absence of Fast Healthcare Interoperability Resources (FHIR) integration, further complicates the digitalization efforts and integration into existing systems [11].

Addressing these challenges necessitates the development of user-friendly, partly unsupervised recording setups that harness existing knowledge, particularly in written assessments, and capitalize on readily available consumer devices. By bridging the gap between traditional paper-based assessments and digital methodologies, such solutions promise to enhance the efficiency, accuracy, and accessibility of Parkinson's disease evaluations, ultimately improving patient outcomes and streamlining healthcare delivery. This paper explores the development of such a system and its integration with a focus on the patient's perspective, aiming to give hints about the digitalization of MDS-UPDRS scores without substantial changes in their given and validated shape. This objective is achieved through the convergence of consumer hardware, open-source software, and the

FHIR standard. Finally, to prove the presented concept this paper presents a usability evaluation of the proposed architecture from diagnosed patients, using the well-established System Usability Scale (SUS) [12].

2. Methods

2.1. Rational of the study

In a preceding project within the Department of Neurology of the University Medical Center Hamburg-Eppendorf, we investigated the potential to expedite the digitalization process for patients frequently enrolled in clinical data collection. Specifically, our focus was on patients with Parkinson's disease, who require regular reassessment and thus would benefit from a more autonomous method of health status recording. We considered digitalizing the MDS-UPDRS questionnaire by leveraging existing technology, such as MoPat [13], to update a FHIR database. However, the Movement Disorder Society (MDS) strictly prohibits any modifications to the form's structure, including the use of radio buttons [10]. Given these constraints, we designed an architecture that circumvents formatting restrictions by utilizing technology capable of reading handwritten digits from locally saved PDF documents.

2.2. Selection of the questionnaire

Within our study, the MDS-UPDRS serves as a validated clinical questionnaire for assessing PD symptoms. Available in German and validated for reliability and validity, MDS-UPDRS forms have been optimized for usage by both patients and doctors. Traditionally paper-based, these forms consist of different sections where patients and doctors assign scores ranging from 0 (normal) to 4 (severe) to predefined fields within, capturing various motor and non-motor symptoms associated with PD. While the UPDRS is fully mapped to LOINC (Logical Observation Identifiers Names and Codes) and available in the corresponding panel 77717-7, the elements of the newer MDS-UPDRS are currently not. Accordingly, we specified them as surrogates in the same style until the inclusion process into the standard is completed.

2.3. Selection of hardware for optimal human-computer interaction

Given the diverse population affected by PD, we established criteria for selecting a device to facilitate effective usage by both patients and clinicians. Handwriting presents an intuitive and efficient mode of human-computer interaction, minimizing potential bias induced by new modalities. Traditionally, patients receive questionnaires on paper, which they can handle well. Using handwriting in this context closely mirrors standard practices, which enhances both its acceptability and practicality in routine clinical settings. Accordingly, we sought consumer-grade devices offering features such as pen support, good contrast resembling paper, low energy consumption, and dimensions akin to an A4 size sheet. E-ink displays emerged as attractive options. Additionally, we prioritized devices with WiFi capability that do not rely on vendor servers, ensuring data privacy, accessibility, and integration into existing healthcare infrastructure.

For developing purposes, we searched for systems allowing us to integrate custom software. As we were unable to identify a popular device with open interfaces and all previous requirements, we fell back to those systems utilizing strict open licenses with a copy-left clause. While the manufacturers are no longer responsible nor must provide support, the corresponding components must legally be changeable and adoptable by an end user. Consequently, we selected the consumer-grade device reMarkable 2 (reMarkable, Oslo, Norway), which is supplied with a pen (Figure 1).

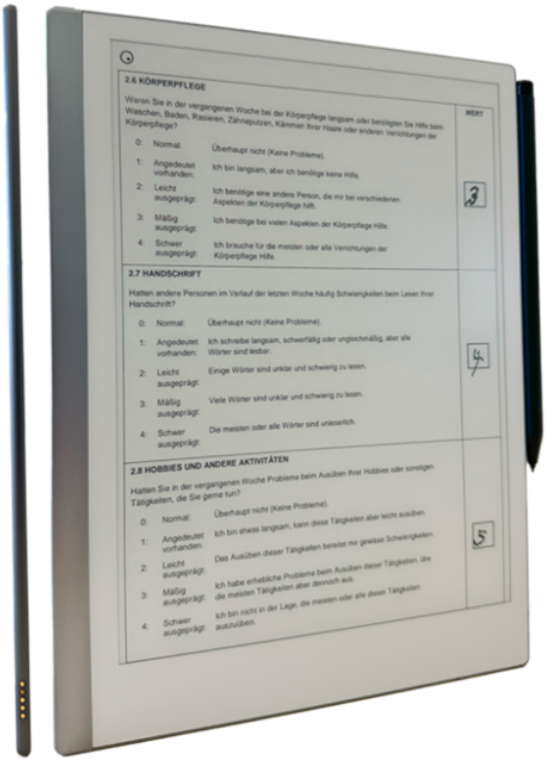


Figure 1. An exemplary part of the MDS-UPDRS presented on the selected device with the associated pen. Independent of the environmental light, the contrast of the display resembles the contrast of a sheet of paper.

2.4. Implementation

For the frontend utilized by patients and physicians, we utilize the original graphical user interface of the device. The individual questionnaires are locally stored as standard PDF documents on the device. To ensure proper privacy protection, we did not utilize any MDS-UPDRS fields including person-identifying markers. Instead, a freely selectable file name serves as an easy surrogate for proper pseudonymization. The original frontend provided navigation within the document and the support for the pen to write the scores in the corresponding fields. By being as near to the optimized graphical user interface as possible, we expect to avoid those challenges already solved by the original manufacturer and profit from its optimizations in that regard.

The manufacturer ensured its compliance regarding the GPL license by providing access to the backend of the device through SSH to the underlying Linux operation system. We used this access to deploy the developed software described in the following.

The central component of our proposed system represents a custom server component continuously running in the background. Writing in the programming

language Rust, the software is cross-compiled to the specific processor architecture and optimized for speed in the resource-limited environment of the device. Further, Rust offers memory safety, has an inbuilt concurrency support, emphasizes robust error handling and is supported by a vibrant community with an expanding ecosystem of libraries (crates). Considering these points, Rust is a strong candidate for building efficient and reliable medical software for resource-constrained devices. Then, through its FHIR-compatible HTTP interface, the software presents the locally stored questionnaires as individual patients with the filename corresponding to the chosen pseudonym. For a fully separated connection of the device to the potential sensible clinical network, a local WiFi network without internet access and appropriate WPA2 encryption could be utilized. The external system could then access our server component through the chosen port and extract the individual observations for the patient of interest (Figure 2).

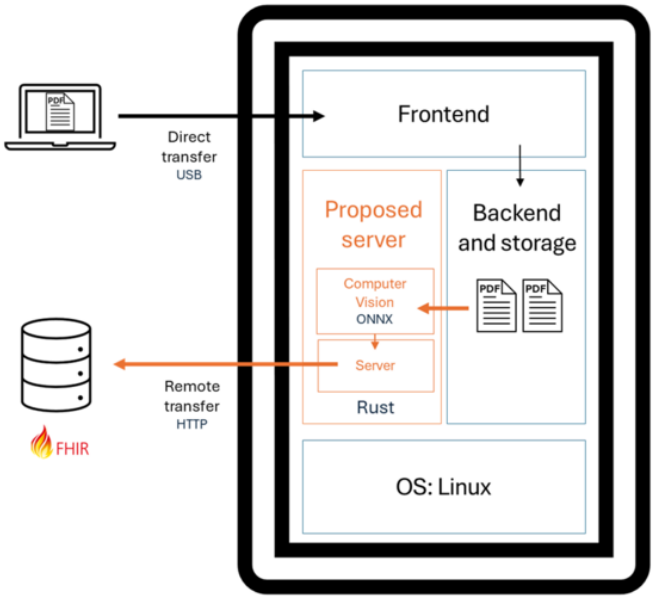


Figure 2. Complete pipeline of our proposed architecture including the tablet device and the FHIR database.

The translation from the handwritten results to the LOINC codes is conducted once queried through the FHIR interface. We choose the pull-based strategy for utilization of the limited computational components on the device where the values are only generated upon request. Given the known position of the boxes to be filled in the automatically generated thumbnails generated by the device, corresponding patches are automatically extracted. Subsequently, the actual task of digit recognition resembles the popular MNIST (Modified National Institute of Standards and Technology) dataset [14], a dataset classically used for evaluating newly developed methods of statistics and machine learning. As this task is considered solved by some authors, we used an existing neural network architecture showing good results with limited computational resources required [15]. For executing the model in the most model-independent manner, we utilized an optimized ONNX runtime. The most likely element from the range between 0 – 4 is then used as a value encoded within a corresponding LOINC element. Crucially, the complete pipeline did not interfere with the document itself nor modify it so that a required human validation is easily possible (Figure 2).

2.5. Usability analysis

To understand the usability of the proposed system in a clinical setting, we conducted a usability study with a patient-centric focus. Patients with PD currently in therapy at the University Medical Center Hamburg-Eppendorf were asked if they would be able to give feedback that would not influence their application. We randomly selected patients with a confirmed diagnosis of PD who were admitted to the Department of Neurology at the time of our survey in March 2024. The whole survey was carried out in accordance with relevant regulations and after written informed consent for the questionnaire was obtained by a physician, participants received the reMarkable 2 device. The device was briefly explained and introduced to the participants, who then should try out filling some fields of the MDS-UPDRS with random values between 0 and 4. During the survey, participants were instructed to open and close the questionnaire, navigate to the next or previous page, and attach the pen. Additionally, we addressed any questions pertaining to the device, the pen, or the MDS-UPDRS questionnaire. Their experience was anonymously reported through the paper-based and well-established System Usability Score (SUS) in a German translation [12] and interpreted according to the guideline of Bangor et al. [16].

3. Results

A total of nine patients diagnosed with Parkinson's disease participated in this study, consisting of 6 males and 3 females. These participants completed the SUS questionnaire to evaluate the usability of the hardware and software systems. The mean age of the participants was 69 years (SD = 13) and the SUS yielded an average score of 83.01 (SD = 9.11). According to the selected interpretation guideline, this score corresponds to an excellent rating, indicating a high level of usability and user satisfaction.

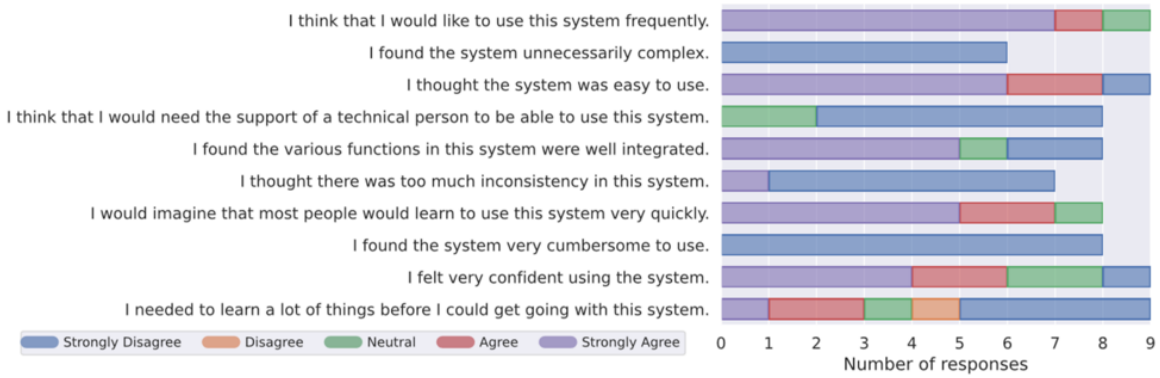


Figure 3. The obtained number of ratings for the different items of the System Usability Scale. The bars correspond to the five possible choices on the associated Likert scale and are always sorted from strongly agree to strongly disagree. For all items with even numbers, more agreement corresponds to better usability while participants should disagree with odd numbers for the same result. If a value was never chosen, its bar does not appear inside the figure. The plot shows the good usability of the system from the perspective of the patients.

The details regarding the participants' responses are illustrated in Figure 3. Notably, all participants expressed their willingness to use the system frequently. The majority of participants reported that they found the system easy to use and did not anticipate requiring support from a technical person, as seen in Figure 3. Furthermore, participants

expressed confidence in their ability to use the system effectively and suggested that most individuals would be able to utilize it in that way.

When observing the patients during their interaction with the device, we identified two caveats. Firstly, the patients utilized the magnetic pen quite naturally. However, once they attached it slightly off the intended position on the frame, the pen fell to the ground, rendering it unfeasible for them to retrieve the pen unaided owing to their motoric symptoms. While that would be the same case for normal pens, the convenient and intuitive option of attaching the pen to the device magnetically worked like a trap leading to behavior with a significant impact on usability. Secondly, the flat design of the device, and potentially all tablet-like devices, poses a significant challenge when lying flat on the table. As most patients were unable to grab below the device, they often tried to shift the device to the border of the table to be able to get a better grip.

4. Discussion

Despite significant efforts to transition to digital systems, paper-based processes remain prevalent in hospital settings. In particular, commercial licenses often impose restrictions on the modification of questionnaires, thereby limiting the possibilities for alternative user interactions. For instance, the MDS explicitly prohibits the incorporation of radio buttons or the addition of comment boxes. This implies not only to the MDS-UPDRS questionnaire but all forms owned by the MDS. In contrast to existing software, such as the MoPat application for patient-centered surveys, which employs various tools such as radio buttons and text boxes [13], we needed to develop an architecture that preserves the integrity of the original questionnaire. Therefore, bridging this gap necessitates smart approaches that optimize existing patterns and processes while minimizing disruption for physicians and patients. A soft digitalization, such as using pens on paper-like tablets for data input while automatically analyzing and mapping results to appropriate medical standards, might offer a feasible solution. Due to its resemblance to printed questionnaires, the majority of participants felt confident using the device, as this approach maintains clarity comparable to traditional paper methods while also facilitating seamless standardization for data exchange. Nevertheless, some patients experienced difficulties with handling the device. Specifically, lifting the flat, tablet-like device from a table and reattaching the magnetic pen to its slim side were not always as intuitive as intended. While designing a device to be as flat as possible to resemble paper may seem logical, utilizing a special case with a designated slot for the pen and options to grab below the device may significantly simplify the usage for patients with PD.

When we evaluate the usability of the technology from the perspective of the patients, the potential of the technology is apparent. In general, the patients despite their different levels of expertise, were highly motivated to utilize the novel technology. They praised the simplicity of the technology and generalized from their short experience with the device to others. From this perspective, the implementation of the device appears to be relatively straightforward.

While our approach highlights the opportunities presented by open-source licenses to implement self-developed software alongside standard consumer hardware in clinical studies, several challenges need to be addressed in future research. As we are adapting hardware for our specific needs, it is unlikely that the toolkit itself will receive approval as a medical device suitable for routine clinical use. Given the lack of official support for the hardware within this particular configuration, issues such as guaranteed liability

and backward compatibility must be taken into account. However, for in-lab conditions during research projects, where the use of uncertified medical devices is permissible, the technology offers an easy, time-saving, and cost-effective alternative for administering questionnaires. Additionally, our study evaluated the usability only from a patient-centric perspective. Although the unsupervised setup may reveal the most beneficial effects, future research should include a quantitative evaluation from the clinician's perspective to provide a more comprehensive assessment of the system's effectiveness. Finally, future work may further improve the utilized algorithms and enrich them, for example, with an outlier detection. Despite those limitations, the proposed system demonstrates the chances of bringing MDS forms into clinics in the least laborious way.

5. Conclusion

In conclusion, our study demonstrates that a novel digital system for administering paper-based questionnaires to patients with Parkinson's disease exhibits high usability and user satisfaction. It further demonstrates the potential of custom software to facilitate innovative applications of consumer hardware in healthcare settings. Through the implementation of smart digitalization strategies, we have shown that it is feasible to leverage existing technologies for the intelligent extraction and mapping of scores related to the progression of Parkinson's disease from paper-simulating tablets. This approach preserves the integrity of written evidence and retains archival options. Participants appreciated the ease of use and the minimal need for technical support, which suggests a promising implementation in clinical settings. Observations revealed practical challenges, such as issues with the magnetic pen and the flat design of the device, indicating areas for improvement. Despite these challenges, the digital system's resemblance to traditional paper methods contributed to patient confidence and ease of adoption. Overall, our study underscores the viability of digitalization to enhance healthcare delivery without significant disruption in hospital settings.

Declarations

Conflict of Interest:

The authors declare no conflict of interest.

Author contributions

CG planned the study, developed the application, and wrote the first version of the manuscript. AJW conducted the evaluation regarding the usability and edited the manuscript. MPN edited the manuscript significantly.

Ethics

The entire survey was conducted in accordance with relevant guidelines and regulations. Participants were included in the survey only after obtaining informed consent by a physician. The study was reported as a scientific case according to the regulations of the

Ärztekammer Hamburg and waived from consultation with the Ethics Committee (2024-300491-WF).

References

[1] Pringsheim T, Jette N, Frolkis A, Steeves TDL. The prevalence of Parkinson's disease: a systematic review and meta-analysis. Mov Disord. 2014;29(13):1583–90. doi: 10.1002/mds.25945.

[2] Richter D, Bartig D, Muhlack S, Hartelt E, Scherbaum R, Katsanos AH, et al. Dynamics of Parkinson's Disease Multimodal Complex Treatment in Germany from 2010–2016: Patient Characteristics, Access to Treatment, and Formation of Regional Centers. Cells. 2019;8(2):151. doi: 10.3390/cells8020151.

[3] Deuschl G, Antonini A, Costa J, Śmiłowska K, Berg D, Corvol JC, et al. European Academy of Neurology/Movement Disorder Society-European Section Guideline on the Treatment of Parkinson's Disease: I. Invasive Therapies. Mov Disord. 2022;37(7):1360–74. doi: 10.1002/mds.29066.

[4] Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment. Mov Disord. 2008;23(15):2129–70. doi: 10.1002/mds.22340.

[5] Giannakopoulou KM, Roussaki I, Demestichas K. Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review. Sensors (Basel). 2022;22(5):1799. doi: 10.3390/s22051799.

[6] Ancona S, Faraci FD, Khatab E, Fiorillo L, Gnarra O, Nef T, et al. Wearables in the home-based assessment of abnormal movements in Parkinson's disease: a systematic review of the literature. J Neurol. 2022;269(1):100–10. doi: 10.1007/s00415-020-10350-3.

[7] Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-Worn Sensors for Remote Monitoring of Parkinson's Disease Motor Symptoms: Vision, State of the Art, and Challenges Ahead. Mirelman A, Dorsey ER, Brundin P, Bloem BR, editors. JPD. 2021;11(s1):S35–47. doi: 10.3233/JPD-202471.

[8] Bloem BR, Post E, Hall DA. An Apple a Day to Keep the Parkinson's Disease Doctor Away? Ann Neurol. 2023;93(4):681–5. doi: 10.1002/ana.26612.

[9] Espay AJ, Bonato P, Nahab F, Maetzler W, Dean JM, Klucken J, et al. Technology in Parkinson disease: Challenges and Opportunities. Mov Disord. 2016;31(9):1272–82. doi: 10.1002/mds.26642.

[10] Monje MHG, Fuller RLM, Cubo E, Mestre TA, Tan AH, Stout JC, et al. Toward e-Scales: Digital Administration of the International Parkinson and Movement Disorder Society Rating Scales. Mov Disord Clin Pract. 2021;8(2):208–14. doi: 10.1002/mdc3.13135.

[11] Gundler C, Zhu QR, Trübe L, Dadkhah A, Gutowski T, Rosch M, et al. A Unified Data Architecture for Assessing Motor Symptoms in Parkinson's Disease. Stud Health Technol Inform. 2023;307:22–30. doi: 10.3233/SHTI230689.

[12] Gao M, Kortum P, Oswald FL. Multi-Language Toolkit for the System Usability Scale. Int J Hum-Comput Interact. 2020;36(20):1883–901. doi: 10.1080/10447318.2020.1801173.

[13] Blitz R, Storck M, Baune BT, Dugas M, Opel N. Design and Implementation of an Informatics Infrastructure for Standardized Data Acquisition, Transfer, Storage, and Export in Psychiatric Clinical Routine: Feasibility Study. JMIR Ment Health. 2021;8(6):e26681. doi: 10.2196/26681.

[14] Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research. IEEE Signal Process Mag. 2012;29(6):141–2. doi: 10.1109/MSP.2012.2211477.

[15] Open Neural Network Exchange. ONNX: Models for MNIST [Internet]. [cited 2024 Apr 3]. Available from: https://github.com/onnx/models/tree/main/validated/vision/classification/mnist

[16] Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Studies. 2009;4(3):114–23.

MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI250840

Digitalizing Medical Forms Through Visual Question Answering: Are We There Yet?

Christopher GUNDLER[a,1], Alexander Johannes WIEDERHOLD[a], and Monika PÖTTER-NERGER[b]
[a] *Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf*
[b] *Department of Neurology, University Medical Center Hamburg-Eppendorf*
ORCiD ID: Christopher Gundler https://orcid.org/0000-0001-9301-8872, Alexander Johannes Wiederhold https://orcid.org/0009-0005-1987-9427, Monika Pötter-Nerger https://orcid.org/0000-0001-7680-2147

Abstract. This study investigates the potential of multimodal neural networks to convert data from unstructured paper-based medical documents into a structured format. Utilizing advancements in Visual Question Answering, we curated a dataset from neurological documents at the University Medical Center Hamburg-Eppendorf. Different models were assessed for their effectiveness. While models from 2024 showed improved performance, accuracy remained below clinical standards, revealing significant challenges in adapting such technology to complex and heterogeneous medical records. The findings emphasize the need for larger, diverse datasets and ongoing refinement to bridge the gap between current model capabilities and human-level performance, underscoring the complexity of automating data extraction in clinical settings.

Keywords. Multimodal neural network, Visual Question Answering, data extraction, clinical documentation, Parkinson's disease, zero-shot learning

1. Introduction

Despite the growing significance of medical informatics in clinical practice, a substantial portion of patient data continues to exist in unstructured form within paper-based documentation [1]. These records, sometimes accumulated over decades, represent a vast resource for data-driven approaches in the medical field. Transforming these documents into structured data usable for analyses still commonly requires extensive manual work. This study investigates the potential of using modern, multimodal neural networks to extract data from scanned paper-based medical records instead. Their capabilities regarding generalizability, observable in the ability of so-called zero-shot learning, are central to their current popularity [2,3]. If zero-shot learning is applicable to medical documentation, it would no longer be necessary to fine-tune the models for specific document types. Instead, extracting knowledge would merely require providing a digitized document and posing a question or prompt in natural language. By enabling natural language as input, these systems could become accessible to virtually any

[1] Corresponding Author: Christopher Gundler, Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, c.gundler@uke.de.

medical scientist [4]. The foundation of this work is based on the significant progress in the field of Visual Question Answering (VQA) and Document Question Answering [5], both of which have been greatly influenced by recent advancements in large language models. VQA is a technique where a question in a prompt about an image is answered in natural language. It involves the combination of visual and textual embeddings to create a unified representation that can understand and respond to questions about visual content. Examples of VQA applications in the medical domain are increasingly seen, ranging from radiology reports to pathology images [6].

2. Methods

2.1. Data

We randomly selected a set of real-world samples from clinical routines conducted at the Department of Neurology at the University Medical Center Hamburg-Eppendorf (Germany) derived from an extensive collection documenting the complex diagnostic assessment of Parkinson's disease. From this collection, we selected three distinct types of medical documentation for our study. The first type was a summary form encompassing the entire examination and an overview of aggregated values. The second type was a detailed documentation of scores obtained in the MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [7], which featured a complex table and detailed information recorded by clinicians. The third type was the Montreal Cognitive Assessment (MoCA) [8], a test used to evaluate the patient's cognitive state and includes a variety of tasks. Unlike the other documents, which were in German, the language of the MoCA test varied depending on the patient's primary language. To establish a valid ground truth for evaluation, we selected four elements of the summary form, the first part of the MDS-UPDRS, and the individual scores of the MoCA test as variables of interest. We then digitized a small subset of the paper-based forms available and blacked out any sensitive information. We labeled the selected key variables as baseline, excluding any missing values from further analysis.

2.2. Selection of multimodal models and prompts

Model selection was guided by the need to ensure practical applicability and privacy compliance within hospital settings. Essential criteria included the capability for local deployment to prevent the transmission of sensitive data to external entities and the reproducibility of results through self-maintenance. All models were required to operate on a single GPU to ensure accessibility without the necessity of a high-performance computing cluster. Furthermore, we required the availability of open-source code and pre-trained weights to ensure reproducibility and transparency. Models were chosen based on their size, established impact, and performance on a task-specific leaderboard. Consequently, we included well-regarded but older models such as Donut [9] and BLIP [10], NanoLLaVA-1.5 as an example of a particularly small model [11], and state-of-the-art models like MiniCPM-Llama3-V 2.5 [12] and InternVL2-1B [13]. Depending on the form, we used prompts like "What is the value for 'MOCA'? Answer only with a single number between 0 and 30" or, when no explicit label was available, "What is the value of the first test from the top? Answer only with a single number between 0 and 5." The prompts contained both a description of which information we wanted to extract and

the numerical value as an additional hint. Whenever the model did not return a single number, we utilized a regular expression to isolate the numerical values.

3. Results

3.1. Data

The final dataset comprised 18 reports from 17 unique patients, with 18 summaries, 14 MDS-UPDRS and MoCA forms, and a total of 24 annotated numerical variables across those forms. During the annotation process, we identified several challenges that must be addressed by the algorithms when applied in clinical settings. Over time, we observed that some of the documents, while containing the same information, differed in layout. While most forms were in German, MoCA forms were written in German, Arabic, and Russian. Additionally, spelling mistakes appear to occur often in clinical practice and are subsequently corrected through crossed-out scores, requiring the model to identify and interpret the correct value adjacent to or even outside the corresponding table cells. Similarly, essential information was sometimes noted beside the tables. Despite the seemingly straightforward task of extracting numerical values, the general noise level inherent in paper-based documentation appeared relatively large.

3.2. Performance of the models

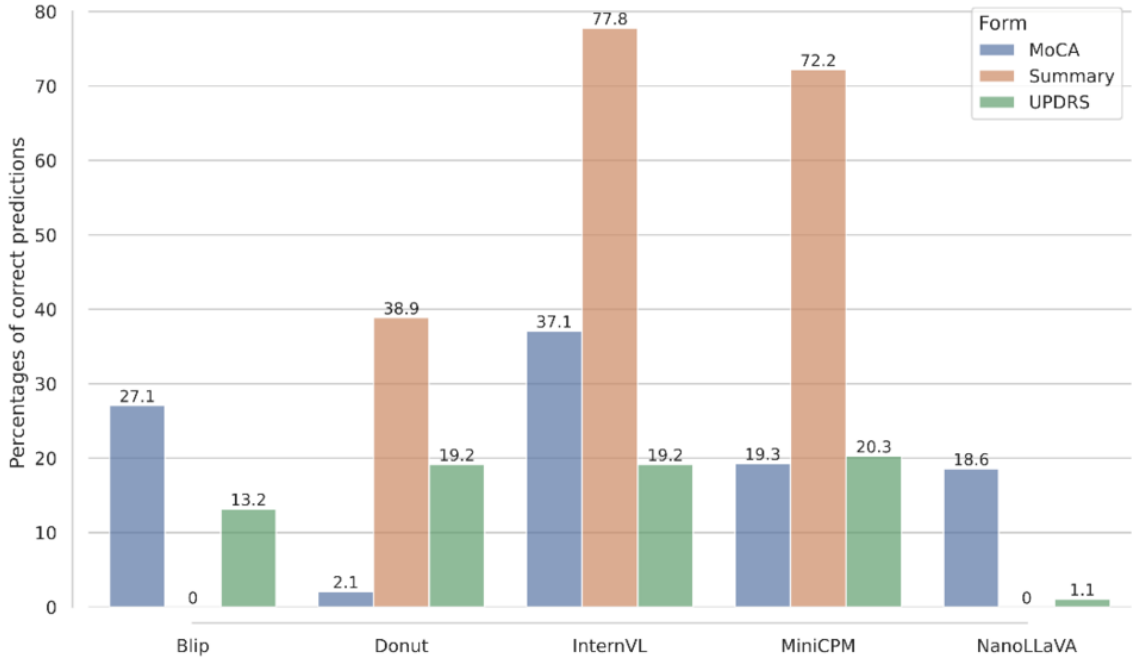


Figure 1. Comparison of the correct predictions across various forms utilized reveals substantial differences in performance outcomes. Higher values indicate superior performance.

When running the models locally on the data and comparing the obtained values with those annotated by humans, the classifiers exhibited widely varying performance levels. The percentage of correct predictions for the different models is illustrated in Figure 1. In terms of absolute performance, the overall results were suboptimal. The peak

performance attained was merely marginally above 77%, which is insufficient for meeting the standards required for clinical applications. Nonetheless, the most recent models significantly outperformed the older models. The progression over the years is noteworthy: earlier models such as Blip and Donut from the early 2020s performed considerably worse than more recent models like InternVL and MiniCPM.

The performance also varied significantly across different forms. No single model excelled across all types of documentation. Generally, the best results were obtained with the summary form, which features a relatively simple tabular structure. Performance declined markedly with the more complex MDS-UPDRS form, which includes multiple adjacent columns. The poorest performance was observed with the MoCA form. Despite having additional information through the number of maximum points, the best performance on this form was only 37%.

4. Discussion

Given the current enthusiasm surrounding the use of multimodal models, the prospect of effortlessly extracting structured data from unstructured documents seems promising. However, although there is evidence suggesting that we are on the right path towards developing capable models, our findings indicate that no current model fully matches the performance of a human. Several factors likely contribute to the mediocre performance observed in these models. Firstly, the generalizability that is often assumed does not appear to fully materialize. Significant differences exist between the training images and questions and the complex realities of clinical documentation. These complexities include a mixture of handwritten and printed text, sophisticated forms used by experts, as well as corrections and annotations frequently made in clinical practice. This indicates the need for a large and diverse dataset of medical documents and associated questions. Once challenges regarding privacy, language diversity and structural integrity are accounted for, such a dataset would be essential both for fine-tuning models and for objectively measuring the performance of newer models once they are published.

Our study has some limitations that should be considered when interpreting the results. Firstly, we used only a selected subset of the extensive amount of paper-based information available. Given their neurological context, the results may not generalize to all medical domains, particularly those with simpler forms designed for patients. Furthermore, there is potential for optimization in the prompts used. While the standardized prompts employed in this study facilitate comparisons, they may not represent the most effective approach for achieving optimal outputs. Lastly, it is important to recognize that each evaluation captures only a snapshot in a rapidly evolving field. The observed improvements in model performance over time suggest that ongoing research is needed. Therefore, standardized and regularized benchmarks that can be replicated and updated with the release of new models are essential.

5. Conclusion

Using recent advances in multimodal models to extract structured data from large volumes of unstructured forms appears to be a promising method for improving processes in healthcare. However, our results indicate that the practical application of this approach in research contexts, particularly in neurology, remains a challenging goal.

Although modern models demonstrate improved capabilities, the variations observed highlight the significant challenges in applying zero-shot learning to the complex and heterogeneous datasets typical for medical records. The rapid advancements in this field necessitate ongoing critical evaluations to ensure generalizability and transferability are accurately assessed before assuming applicability to similar challenges.

6. Declarations

Conflict of Interest: The authors declare no conflict of interest.
Author contributions: CG conducted the study and wrote the first version of the manuscript. AJW annotated the data and edited the manuscript. MPN edited the manuscript significantly.
Ethics: The entire survey was conducted in accordance with relevant guidelines and regulations. The study was reported as a scientific case according to the regulations of the Chamber of Physicians Hamburg (2024-300504-WF).

References

[1] Fröhlich D, Bittersohl C, Schroeder K, Schöttle D, Kowalinski E, Borgwardt S, et al. Reliability of Paper-Based Routine Documentation in Psychiatric Inpatient Care and Recommendations for Further Improvement. Front Psychiatry. 2020;10:964. doi: 10.3389/fpsyt.2019.00954.

[2] Bian C, Yuan C, Ma K, Yu S, Wei D, Zheng Y. Domain Adaptation Meets Zero-Shot Learning: An Annotation-Efficient Approach to Multi-Modality Medical Image Segmentation. IEEE Trans Med Imaging. 2022;41(5):1043–56. doi: 10.1109/TMI.2021.3131245.

[3] Ge Y, Guo Y, Das S, Al-Garadi MA, Sarker A. Few-shot learning for medical text: A review of advances, trends, and opportunities. J Biomed Inform. 2023;144:104458. doi: 10.1016/j.jbi.2023.104458.

[4] AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. J Med Internet Res. 2024;26:e59505. doi: 10.2196/59505.

[5] Manmadhan S, Kovoor BC. Visual question answering: a state-of-the-art review. Artif Intell Rev. 2020;53(8):5705–45. doi: 10.1007/s10462-020-09832-7.

[6] Lin Z, Zhang D, Tao Q, Shi D, Haffari G, Wu Q, et al. Medical visual question answering: A survey. Artif Intell Med. 2023;143:102611. doi: 10.1016/j.artmed.2023.102611.

[7] International Parkinson and Movement Disorder Society. MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [Internet]. [cited 2025 Jan 8]. Available from: https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS.htm

[8] Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. J American Geriatrics Society. 2005;53(4):695–9. doi: 10.1111/j.1532-5415.2005.53221.x.

[9] Kim G, Hong T, Yim M, Nam J, Park J, Yim J, et al. OCR-free Document Understanding Transformer [Internet]. arXiv.org. 2021 [cited 2024 Jul 19]. Available from: https://arxiv.org/abs/2111.15664v5

[10] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation [Internet]. arXiv; 2022 [cited 2024 Jul 19]. Available from: http://arxiv.org/abs/2201.12086

[11] Quan N. nanoLLaVA-1.5 [Internet]. 2024 [cited 2024 Jul 19]. Available from: https://huggingface.co/qnguyen3/nanoLLaVA-1.5

[12] Hu J, Yao Y, Wang C, Wang S, Pan Y, Chen Q, et al. Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages [Internet]. arXiv; 2024 [cited 2024 Jul 19]. doi: 10.48550/arXiv.2308.12038. Available from: http://arxiv.org/abs/2308.12038

[13] Chen Z, Wang W, Tian H, Ye S, Gao Z, Cui E, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites [Internet]. arXiv; 2024 [cited 2024 Jul 19]. doi: 10.48550/arXiv.2404.16821. Available from: http://arxiv.org/abs/2404.16821

German Medical Data Sciences 2023 — Science. Close to People.
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI230689

A Unified Data Architecture for Assessing Motor Symptoms in Parkinson's Disease

Christopher GUNDLER[a,1], Qi Rui ZHU[a], Leona TRÜBE[a], Adrin DADKHAH[b], Tobias GUTOWSKI[b], Moritz ROSCH[b], Claudia LANGEBRAKE[b], Sylvia NÜRNBERG[a], Michael BAEHR[b], and Frank ÜCKERT[a]
[a] *Applied Medical Informatics, University Medical Center Eppendorf, Germany*
[b] *Pharmacy, University Medical Center Eppendorf, Germany*

Abstract. Introduction The diagnosis and treatment of Parkinson's disease depend on the assessment of motor symptoms. Wearables and machine learning algorithms have emerged to collect large amounts of data and potentially support clinicians in clinical and ambulant settings. **State of the art** However, a systematical and reusable data architecture for storage, processing, and analysis of inertial sensor data is not available. Consequently, datasets vary significantly between studies and prevent comparability. **Concept** To simplify research on the neurodegenerative disorder, we propose an efficient and real-time-optimized architecture compatible with HL7 FHIR backed by a relational database schema. **Lessons learned** We can verify the adequate performance of the system on an experimental benchmark and in a clinical experiment. However, existing standards need to be further optimized to be fully sufficient for data with high temporal resolution.

Keywords. Parkinson's disease, motor symptoms, inertial sensors, acceleration data, machine learning algorithms

1. Introduction

1.1. Background

Parkinson's disease (PD) is one of the neurodegenerative diseases with the greatest implications for contemporary societies [1]. While PD also manifests in non-motor symptoms with significant burdens on the patient, research mainly focuses on the characteristic motor symptoms [2]. The causes of the disease are relatively well understood, but effective treatment remains a challenge. Consequently, computer-aided support for managing medical assessments and therapies is an active area of research [3].

Various kinds of sensors are currently under investigation for their clinical feasibility [1]. Inertial sensors of so-called wearables have emerged to effortlessly collect large amounts of data. These portable computers are worn as smartwatches, fitness bracelets, or similar devices on different body parts. Developed for the consumer market, the collected data is usually used to measure activities and sleep quality. However, the miniaturization of the necessary sensor technology also enables its usage in clinical and

[1] Corresponding Author, Christopher Gundler, Applied Medical Informatics, University Medical Center Eppendorf, Martinistraße 52, 20246 Hamburg, Germany; E-mail: c.gundler@uke.de.

ambulant settings. However, the vast amount of generated data in the process poses a challenge.

Machine learning algorithms have emerged as valuable tools for analyses of motion data acquired in monitoring motor symptoms. In research, classification of severeness, differential diagnosis, medication management, and similar tasks are nowadays evaluated accordingly [1]. These techniques can analyze large amounts of data with high accuracy. Therefore, these tools may reduce subjective differences in perception between clinicians and simplify daily clinical routines. Given the potential of machine learning algorithms, development and evaluation of such technologies increased in recent years.

However, multiple authors also stated limitations and pitfalls requiring consideration within that use case [4,5]. The accuracy and effectiveness of machine learning algorithms largely depend on the quantity and quality of the data. Many studies investigating the use of wearables for motion monitoring in PD often have small sample sizes and short study durations limiting the quantity. Symptoms and progress of PD can differ significantly between affected individuals. Accordingly, procedures primarily based on population statistics may result in a bias to the detriment of an individual. Other limitations result from the available variety of wearable devices, machine learning algorithms, and evaluation criteria preventing comparability. Apart from this, the use of artificial intelligence in healthcare raises several ethical and legal concerns, such as privacy, bias, and accountability. Most of these challenges could be managed by a secure, sufficiently large, and diverse collection of standardized data from suitable sources for ensuring proper generalization.

1.2. Objective and Requirements

This work aims to develop a unified data architecture for research on accelerometer data acquired in PD studies. A central goal is the interoperability and combination of different datasets according to the findable, accessible, interoperable, and reusable (FAIR) principles. Two research questions can be formulated:

1. Do existing studies keep their data in a format that allows easy interoperability and combination?
2. Which kind of structure is needed to optimally link different datasets and make them accessible for future research?

The data architecture aims to serve as a basis for future research with wearables in PD motion monitoring.

2. State of the art

Giannakopoulou *et al*. identified 53 publications using inertial sensors for various analyses in PD research [1]. To assess data availability and formats, all referenced publications were systematically searched for related statements. Most of the included studies provided no access to the raw data. Naturally, sensitive samples are commonly protected by data protection acts. At the same time, further research might be limited due to missing access. A minority of the included publications allow access to the underlying measurements through a transparent authorization process. The most popular examples are the mPower study [6, e.g. used by 7], the MJFF Levodopa Response Study [8], and suitable parts of the UK Biobank dataset [9, e.g. used by 10]. Another notable dataset without any access regulation is Daphnet [11, e.g. used by 12].

While the underlying platforms ensure authenticated and authorized access to the data, we were unable to identify a platform for receiving and storing datasets complying with the FAIR principles. Instead, the data are mainly stored in study-dependent formats requiring customized preprocessing for individual research. Such semantic incompatibilities are amplified by differing research questions and various paradigms in the data acquisition of inertial sensors. Since the platforms are primarily built to publish static data and do not offer real-time interfaces, they are not suitable for storing data with high temporal resolution encountered during experiments. Consequently, providing an alternative appears as an efficient way to simplify research.
Generally, only "very few studies or research projects have investigated [appropriate data standardization for wearable data] or proposed standardization procedures" [13]. The publications focus primarily on harmonizing general movement data on local systems powered by MATLAB [13,14]. While those approaches are powerful in their range of supported data modalities, they are neither tailored to the specific requirements of Parkinson´s disease nor optimized to be used as platforms.

In summary, sustainable use of the data published remains conditionally possible. The formats currently in use are neither standardized nor capable of real-time analyses. A unified data architecture would enable the integration of data from different sources, such as databases, individual files containing tabular data, and external research data platforms, into a single system. Consequently, developed data processing pipelines can be used independently of the actual data origin. In the following, we present a possible concept and implementation of a unified data architecture for heterogeneous datasets containing accelerometer data.

3. Concept: Designing a uniform data architecture

The conception and implementation of the uniform data architecture for the monitoring of motor symptoms in PD can be divided into separate steps. After the identification of feasible data sources in the literature, a classic Extract, Transform, Load (ETL) process was applied. This was followed by the definition, development, and testing of a suitable interface standardized for data storage, modification, and retrieval.

The datasets of Daneault *et al*. and Bot *et al*. were included to exemplary create a unified data architecture [6,8]. After receipt of the datasets, their content and metadata were analyzed. In the second step, we defined the transformation rules. In contrast to classical ETL processes in existing data warehouses, a unified data architecture was identified based on the analysis of the included data, the information from the previously mentioned literature review, and the discussions with clinicians. The three central structures identified in this process are (1) the subjects, (2) the inertial data, and (3) the medical assessments.

3.1. Integration of the subjects

Similar to existing standards, the unified data model is subject-centered. For reasons of data minimalization, their description is by default rudimentary. In addition to an internal identification number, classification to a certain cohort is primarily relevant. The original study can only be identified at this point within the architecture. The possibility of assigning further descriptions can also be used to ensure further connections.

3.2. Integration of acceleration data with a high temporal resolution

The central element of motion recording is the data with high temporal resolution collected by inertial sensors of wearables. The movement data are complex and can be collected using sensors from different manufacturers. Accordingly, they may differ in temporal resolution, measurement ranges, and associated measurement errors [13]. The impact of these variances may potentially influence further analyses. Therefore, the sensor type and group structure must be reflected by the data architecture. Depending on the manufacturer, the inertial sensors can use only accelerometers or combinations of accelerometers, gyroscopes, and other sensors. Due to its popularity in research regarding PD [1], we only consider the first type of sensor within this study. The unit of those measurements can vary; both the SI unit m/s^2 and the mean acceleration due to gravity are common. Since conversion between both units is possible, the proposed architecture needs to ensure its integration. Of greater relevance is the measurement site on the body. The placement of sensors differs between studies and might reflect differences in manifestations of motor symptoms on distinct body sites in PD patients [3].

Based on the resources described, the storage of the actual acceleration data is relatively simple: Both, the identification numbers of the sensor used on a specific body part and the identification number of the patient in combination with a timestamp function as the key element for the measured sample. This sample is serialized in the standardized SI unit m/s^2 in all three dimensions with timestamps transferred from the time zone of their recording to the Coordinated Universal Time (UTC). Consequently, the data become easily comparable across studies. This enables the efficient mass storage of the data with high temporal resolution and provides the basis for subsequent analyses.

3.3. Integration of medical assessments

The unified data model specified so far is suitable for making acceleration data recorded over a course of time available and easily retrievable. This enables, for example, continuous monitoring of motor symptoms in an ambulant setting. In terms of machine learning models, unsupervised methods can thus already be applied. In literature, however, supervised methods are more frequently used. The integration of medical assessments appears to be necessary and represents the third central structure of the unified data architecture.

The basis of any assessment is the use of a test procedure as standardized as possible. In this context, they are commonly represented as physical exercises highlighting specific motor symptoms in PD. However, different options are available often based on the personal experiences of clinicians or individual hospital guidelines. A certain standardization can be achieved by using rating scales, for example, the motor section of the unified Parkinson's disease rating scale (UPDRS). While both reference datasets supported this classification, the schema optionally allows the use of proprietary exercises. With this design decision, we choose flexibility and straightforward applicability for custom studies over inter-study comparability.

A separate resource is created for each exercise. It contains the test procedure used and the patient examined. It also offers the possibility to define a precise start and end point of the exercise. As a result, a patient's exercise documentation can overlap, even if they are in principle performed sequentially to each other. In the context of diagnoses, several motor symptoms including tremor, dyskinesia, and bradykinesia can be assessed

in parallel during each exercise. Their expressions are encoded on different scales. Hence, the joint encoding of the motor symptom and the scale demonstrate comparability and differences.

The final classification can be used as a label for machine learning. It is distinctively defined as a rating of a specific instance of a task and can be particularized to individual body parts. The value then encodes the severity of the motor symptom defined within the applied scale.

3.4. Formalization of a relational database schema

Based on the previously described definitions, a relational database schema was formalized which is shown in Figure 1. The resulting definition of resources represents the lowest common denominator between the different data sources extracted. According to clinical requirements, the relational database scheme possesses the necessary flexibility to be compatible with a wide range of study designs and evaluation methods. Furthermore, the use of an optimized database management system enables a performance that is sufficient for large amounts of data and can handle a range of workloads. The SQL definition of the schema should be compatible with most available systems. The actual data storage can thus take place within stable and well-tested systems.

To finalize the ETL process, the two exemplary datasets of Daneault *et al*. and Bot *et al*. were loaded into the unified architecture [6,8]. To perform the required syntactic and semantic transformations, we developed independent command-line tools.

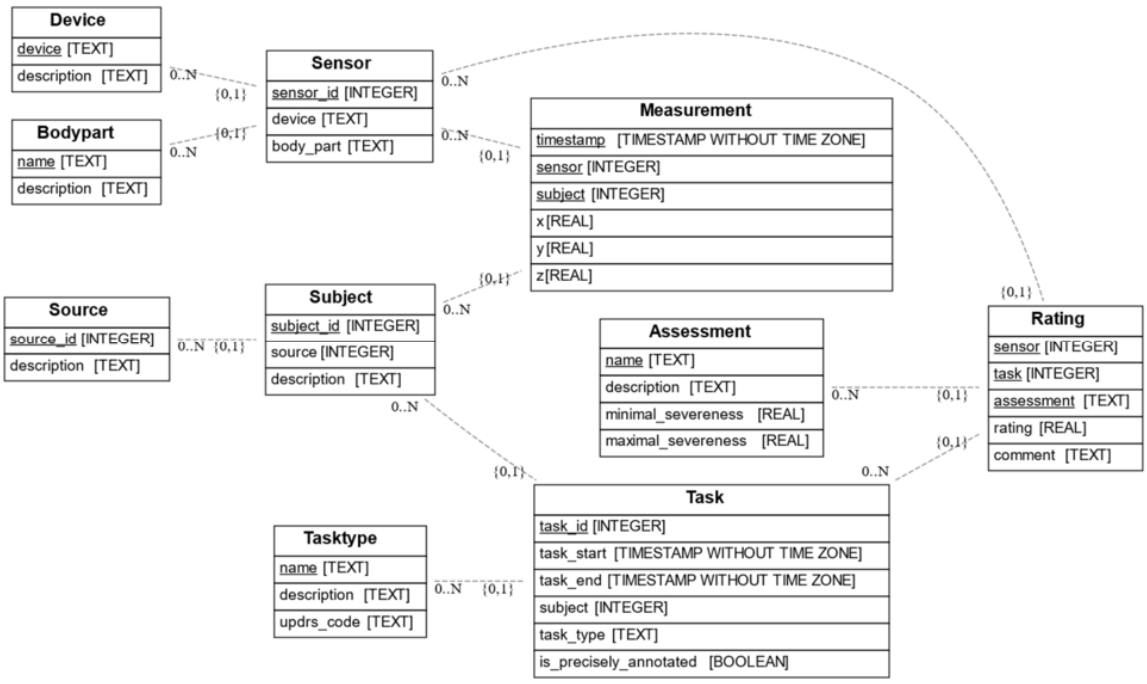


Figure 1. Visualization of the proposed relational database schema

4. Implementation: FHIR as an interface to the uniform data architecture

After applying the appropriate schema and the ETL process, the data are available in a relational database management system. They can be added, edited, and retrieved via SQL. This forms the foundation for efficient storage and retrieval of the acceleration data acquired in studies. However, this type of data management is insufficient to ensure a reusable and expandable data architecture. Therefore, fast healthcare interoperability resources (FHIR), established by Health Level Seven International (HL7), was introduced as an interface to approach the optimized data management formulated at the beginning.

HL7 FHIR is a standard for data exchange between different software solutions in healthcare. By design, it is defined to be compatible with FAIR principles. The associated overhead not strictly required for fulfilling the interoperability aspect of the latter might not be necessary for every use case. In our case, providing an FHIR-compatible interface while referencing the SQL database schema in the background facilitates the integrability into existing systems. In addition, the development of the customized front end enables a more defined integration of machine learning algorithms and models. Accordingly, there are numerous possible integrations within clinical infrastructures.

For creating an appropriate interface for data insertion and querying, we developed a Java-based server. The HAPI FHIR library was used as the technical foundation to ensure proper and well-tested parsing and conformity to the standard. Inside the software, the values are mapped on-the-fly to and from PostgreSQL as the underlying database management system. Easily deployable through Docker, the code is available under a permissive open-source license and available to research groups interested in PD[2].

4.1. Results

At least two quality measures are important for assessing the appropriability of the proposed tool. The general correctness of the underlying transformation processes and the stored data are ensured through a battery of integration tests. However, if the system should be used for data collection in real-time, an appropriate insertion performance is required. Accordingly, we benchmarked our server against a reference server developed by the team behind the HAPI FHIR library. While the FHIR interface and parsing functionality are the same, the reference implementation used a general-purpose database schema instead of our proposed backend.

The corresponding benchmark results of four parallel threads and data payload of 100.000 requests without network overhead are stated in figure 2. The boxplot of insertion timings indicates an asymmetrical data distribution skewed towards lower writing times. The median of the proposed system (0.005 seconds per insertion) is substantially lower than the reference HAPI server (0.01 seconds per insertion). However, our naïve and unoptimized implementation shows a larger spread in the interquartile range. Summarizing, the proposed FHIR interface implementation appears useful for real-time applications.

[2] Corresponding Git repository: https://github.com/UKEIAM/de.uke.iam.parkinson_on_fhir

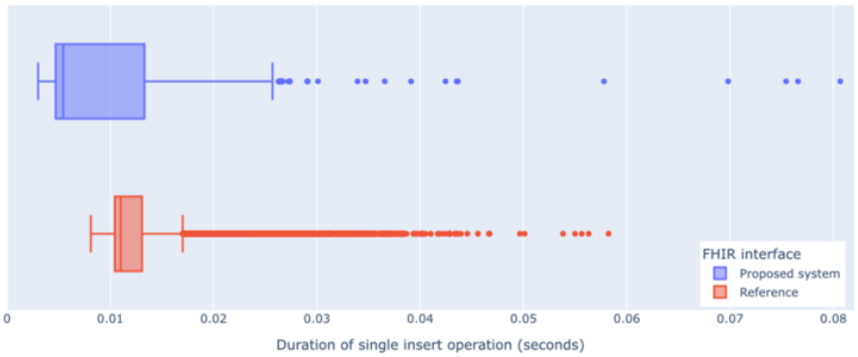


Figure 2. Boxplot of insertion times of the proposed and reference system

4.2. System in use

The proposed relational database schema and FHIR interface are already deployed for storing acceleration data within a PD study currently conducted at the University Medical Center Eppendorf. Utilizing two applications for smartwatches with the popular operating systems iOS and Android OS, the accelerometer data of up to 50 patients are recorded for various movement tasks with a temporal resolution of 50 samples per second. By the time of writing, more than 107.068.493 of these samples have been recorded from up to five subjects simultaneously. Subsequently, the data will be used to research solutions for the optimized treatment of PD.

5. Lessons learned

As demonstrated in both, the lab and the pilot study, storing motion data of patients according to FAIR principles is possible. A hybrid system with a relational backend optimized for performance and a FHIR-compatible frontend can handle requests in real-time. Consequently, integration of such systems is possible within existing clinical environments.

Current limits are set not by the technology itself but by the use of FHIR for this kind of data modality. As an example, encoding each recording sample as an individual observation might appear verbose. FHIR supports sampled data as a type of value stored within this resource. However, such a storage format may hinder the application in the real-time setting as easy insertion is not possible and render longitudinal studies hard. It will be interesting to see if and how future releases may induce further opportunities.

The medical vocabularies used for interoperability are not readily applicable, either. While there exists a LOINC code 80493-0 for an aggregated form of acceleration, individual gravitational components could not be specified. The FHIR specification does neither allows possible alternatives like SNOMED-CT concepts with post-coordination within the Measurement resource. Accordingly, our implementation falls back to not-yet-standardized codes. Getting those codes standardized through the official process remains future work. The code under a permissive open-source license allows adaptation for the required changes in the future. We hope the unified data architecture can further boost research activity regarding PD.

6. Conclusion

The current research landscape in the sensor-based recording of motor symptoms in PD is diverse. Accordingly, existing studies do not maintain data in a format that easily allows interoperability and a combination of different datasets. Within the scope of this project, a uniform data architecture was developed to represent the multitude of possible paradigms for recording movement data in patients with PD. It enables reusable storage of research data, allows interoperable communication between systems, and effective training of machine learning algorithms. The data architecture can be used for primary data acquisition or serve as an endpoint for decentral collected measurement data. Using FHIR as an interface for optimized data management offers the possibility for extensions depending on the requirements of future research projects. For example, in addition to the measurement data, medication data can be stored and included in analyses via the multiple resources provided by the standard.

In summary, the designed, implemented, and tested data architecture can provide a basis for future PD research. By standardization of popular yet inhomogeneous datasets, and by providing sufficient data management with an expandable data architecture, it offers the opportunity to link machine learning methods to daily clinical routines and further boost research activity regarding PD.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.

Contributions of the authors: CG and QZ were involved in the planning and implementation of the project. AD, MR, TG, CL, SN, and MB contributed significantly to the clinical study. CG, QZ, LT, AD, MR, TG, CL, SN, MB, and FÜ were involved in the writing and/or revision of the manuscript.

References

[1] Giannakopoulou KM, Roussaki I, Demestichas K. Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review. Sensors (Basel). 2022 Feb 24;22(5):1799.

[2] Adams JL, Lizarraga KJ, Waddell EM, Myers TL, Jensen-Roberts S, Modica JS, et al. Digital Technology in Movement Disorders: Updates, Applications, and Challenges. Curr Neurol Neurosci Rep. 2021 Mar 3;21(4):16.

[3] Ancona S, Faraci FD, Khatab E, Fiorillo L, Gnarra O, Nef T, et al. Wearables in the home-based assessment of abnormal movements in Parkinson's disease: a systematic review of the literature. J Neurol. 2022 Jan 1;269(1):100–10.

[4] Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-Worn Sensors for Remote Monitoring of Parkinson's Disease Motor Symptoms: Vision, State of the Art, and Challenges Ahead. Mirelman A, Dorsey ER, Brundin P, Bloem BR, editors. JPD. 2021 Jul 16;11(s1):S35–47.

[5] Espay AJ, Bonato P, Nahab F, Maetzler W, Dean JM, Klucken J, et al. Technology in Parkinson disease: Challenges and Opportunities. Mov Disord. 2016 Sep;31(9):1272–82.

[6] Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016 Mar 3;3(1):160011.

[7] Zhang H, Deng K, Li H, Albin RL, Guan Y. Deep Learning Identifies Digital Biomarkers for Self-Reported Parkinson's Disease. Patterns. 2020 Jun 12;1(3):100042.

[8] Daneault JF, Vergara-Diaz G, Parisi F, Admati C, Alfonso C, Bertoli M, et al. Accelerometer data collected with a minimum set of wearable sensors from subjects with Parkinson's disease. Sci Data. 2021 Feb 5;8(1):48.

[9] Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLOS ONE. 2017 Feb 1;12(2):e0169649.

[10] Williamson JR, Telfer B, Mullany R, Friedl KE. Detecting Parkinson's Disease from Wrist-Worn Accelerometry in the U.K. Biobank. Sensors. 2021 Jan;21(6):2047.

[11] Bachlin M, Plotnik M, Roggen D, Maidan I, Hausdorff JM, Giladi N, et al. Wearable Assistant for Parkinson's Disease Patients With the Freezing of Gait Symptom. IEEE Transactions on Information Technology in Biomedicine. 2010 Mar;14(2):436–46.

[12] Li B, Yao Z, Wang J, Wang S, Yang X, Sun Y. Improved Deep Learning Technique to Detect Freezing of Gait in Parkinson's Disease Based on Wearable Sensors. Electronics. 2020 Nov 14;9(11):1919.

[13] Palmerini L, Reggi L, Bonci T, Del Din S, Micó-Amigo ME, Salis F, et al. Mobility recorded by wearable devices and gold standards: the Mobilise-D procedure for data standardization. Sci Data. 2023 Jan 19;10(1):38.

Original Paper

Unlocking the Potential of Secondary Data for Public Health Research: Retrospective Study With a Novel Clinical Platform

Christopher Gundler[1*], MSc; Karl Gottfried[1*]; Alexander Johannes Wiederhold[1], MD; Maximilian Ataian[1]; Marcus Wurlitzer[2], Dr rer nat; Jan Erik Gewehr[2], Dr rer nat; Frank Ückert[1], Dr med

[1]Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[2]Research Data Facility, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[*]these authors contributed equally

Corresponding Author:
Christopher Gundler, MSc
Institute for Applied Medical Informatics
University Medical Center Hamburg-Eppendorf
Martinistrasse 52
Hamburg, 20246
Germany
Phone: 49 40741054979
Email: c.gundler@uke.de

Abstract

Background: Clinical routine data derived from university hospitals hold immense value for health-related research on large cohorts. However, using secondary data for hypothesis testing necessitates adherence to scientific, legal (such as the General Data Protection Regulation, federal and state protection legislations), technical, and administrative requirements. This process is intricate, time-consuming, and susceptible to errors.

Objective: This study aims to develop a platform that enables clinicians to use current real-world data for testing research and evaluate advantages and limitations at a large university medical center (542,944 patients in 2022).

Methods: We identified requirements from clinical practitioners, conceptualized and implemented a platform based on the existing components, and assessed its applicability in clinical reality quantitatively and qualitatively.

Results: The proposed platform was established at the University Medical Center Hamburg-Eppendorf and made 639 forms encompassing 10,629 data elements accessible to all resident scientists and clinicians. Every day, the number of patients rises, and parts of their electronic health records are made accessible through the platform. Qualitatively, we were able to conduct a retrospective analysis of Parkinson disease over 777 patients, where we provide additional evidence for a significantly higher proportion of action tremors in patients with rest tremors (340/777, 43.8%) compared with those without rest tremors (255/777, 32.8%), as determined by a chi-square test (P<.001). Quantitatively, our findings demonstrate increased user engagement within the last 90 days, underscoring clinicians' increasing adoption of the platform in their regular research activities. Notably, the platform facilitated the retrieval of clinical data from 600,000 patients, emphasizing its substantial added value.

Conclusions: This study demonstrates the feasibility of simplifying the use of clinical data to enhance exploration and sustainability in scientific research. The proposed platform emerges as a potential technological and legal framework for other medical centers, providing them with the means to unlock untapped potential within their routine data.

(Interact J Med Res 2024;13:e51563) doi: 10.2196/51563

KEYWORDS

secondary use; hypothesis testing; research platform; clinical data; Parkinson disease; data; health-related research; health data; electronic health record; EHR; tremor

Introduction

In recent years, there has been a growing interest in using clinical routine data, especially electronic medical records, for research [1]. Known as secondary data use, this practice is significantly influenced by legislative actions such as the Health Information Technology for Economic and Clinical Health Act in the United States of America and publicly founded initiatives

like the Medical Informatics Initiative (MII) in Germany [2,3]. University hospitals now serve as central hubs in bridging the gap between research and patient care. Achieving connectivity between previously isolated data silos necessitates adherence to detailed standards, such as the custom FHIR (Fast Healthcare Interoperability Resources) profile "Kerndatensatz" in Germany and effective communication among diverse stakeholders within national health systems [4]. Despite the inherent complexity, these concerted efforts are expected to establish a new foundation for evidence-based science.

Despite the advantages of adopting a state-wide approach to secondary data use, we contend that certain research could be more effectively conducted at the local level within individual hospitals. Specifically, we have identified 2 critical use cases where hypothesis testing on unmapped raw data is essential for advancing evidence-based medicine:

First, to verify hypotheses derived from clinical practice on a larger database, clinicians should be able to validate their experiences easily through data-driven investigations. These studies may involve data elements not comprehensively covered by standardized data sets. In addition, expecting clinicians to create intricate mappings between used data elements and state-wide standards may prove ineffective.

Second, the replication of existing publications to assess their generalizability must always consider the local context. Conducting public health research, for example on large cohorts of patients with Parkinson disease (PD), might miss important external factors [5,6]. Accordingly, testing external validity on a schema applied in practice rather than one developed for collaboration appears more appropriate.

Beyond the clinical perspective, local solutions may better accommodate regional or local legal requirements, as many collaborative standardizations tend to converge on the "smallest common divisor" between partners. Consequently, the implementation of complementary systems for secondary data analysis at both the local and global levels is deemed appropriate.

Developments in recent years have led to the emergence of research platforms that enable the analysis of clinical data in compliance with data protection guidelines. Notable examples include EPOCH and ePRISM (IP-ITT Corporation) [7], KETOS (Friedrich-Alexander University Erlangen-Nürnberg) [8], and Medical-Blocks (University of Bern) [9]. These platforms provide environments that allow clinical scientists to train and deploy statistical models. However, their primary focus is on translating these models back into clinical practice rather than testing hypotheses through the secondary use of data. In addition, works such as EHR4CR (Electronic Health Records for Clinical Research) [10] have implemented infrastructures that enable the use of clinical data across multiple European sites in a secure and privacy-preserving manner without focusing on the subsequent analysis. Given our knowledge, a platform for hypothesis testing on routine data has not been implemented and evaluated in clinical reality.

This paper addresses this gap in research by introducing and evaluating a novel platform explicitly designed for hypothesis testing on clinical routine data. Starting by collecting the requirements of clinicians, we strive to design and implement a modular and, consequently, reusable platform. Similar to other states, the federal law of Hamburg permits pseudonymized retrospective data analysis without patient consent given specific guarantees regarding data protection. The platform ensures those guarantees in accordance with all European and German laws and is directly integrated into the technical infrastructure of the University Medical Center Hamburg-Eppendorf (UKE). The evaluation process encompasses quantitative assessment, exemplified by the replication of a public health finding in the context of PD, and qualitative evaluation through the examination of clinicians' use in real-world clinical scenarios. This dual-pronged evaluation strategy aims to judge both the quantitative efficacy and practical use of the proposed platform in clinical reality.

Methods

Technical Considerations

For developing the platform, we examined the challenges of using routine hospital data for hypothesis testing through extensive communication with the different business divisions of the UKE, like the infrastructure department, division for information technology, research data facility, data protection officers, and internal boards. Informed by the project meetings and discussions with clinicians as later users, we identified and prioritized 4 critical process components necessitating optimization.

Defining Appropriate Hypotheses

Precise hypothesis formulation relies on a thorough understanding of metadata within the clinical information system. For researchers, the accessibility of relevant data fields may not be immediately evident. Challenges arise from both nontechnical limitations and the opacity of data type and structure. Filtering cohorts based on specific criteria may yield statistically inappropriate sizes, and requested data may be inadequately recorded [11]. The feasibility of research ideas is thus not guaranteed, necessitating extensive consultation with data integration experts for hypothesis refinement.

Obtaining Data From the Infrastructure

Efficient storage and retrieval of routine hospital data are crucial for medical treatment and research. Hospitals use diverse IT architectures, often a mix of specialized systems with proprietary data structures and nonstandardized file formats. Access and control vary widely, from centralized systems to more federated approaches led by individual clinics. Clinicians aiming to test hypotheses face challenges in accessing required documentation, understanding these structures, and communicating with the responsible data manager.

Analysis of the Hypotheses

To facilitate hypothesis tests, clinicians expressed a need for a comprehensive and heterogeneous array of tools, encompassing table-based software and standard scripting languages like Python (Python Software Foundation) or R (R Foundation for Statistical Computing). Established research data management

platforms, such as Kaggle (Google) [12], Paperspace Gradient (DigitalOcean) [13], Colab (Google) [14], or CodaLab (Microsoft Research) [15], provide ideal support for efficient data analysis: An integrated and simplified development environment, a separate space for data analysis with access to high-performance computing, and the ability to communicate and collaborate with other users of the research community. Rather than developing a novel solution, leveraging a platform that accommodates diverse analysis methods appears to be a pragmatic approach.

Reuse of Established Components

Based upon the preliminary work of the MII and the existing research landscape, the following tools were explored as relevant in the context of our work.

Data Integration

Data integration centers (DICs) enable the cross-site and cross-institutional use of digital health data from patient care and biomedical research in Germany [2,3,16]. All DICs are located at university medical sites and have access to routinely collected patient data. To this end, they build up interoperable databases with quality-assured and internationally harmonized data (based on HL7 [Health Level Seven International] FHIR) and metadata. These are made available in anonymized form through trustees. DICs make an important contribution to the development of a research-orientated infrastructure for the German health care system. The first use cases using the functionalities are already in operation [17]. These functionalities are reusable and valuable for our work. For further details, we refer to the literature regarding the MII [2,3].

Data Usage Considerations

European, national, and local laws govern the use of sensitive routine data. Those projects necessitate ethical approval and explicit consent, a crucial yet burdensome process for both researchers and ethics committee members. As the legislators have already identified the need for simplification, we were able to use §12 of the "Hamburgisches Landeskrankenhausgesetz" [18]. This statute permits pseudonymized retrospective data analysis without patient consent, allowing us to forego consent-based data usage. With approval from the ethics committee for hypothesis testing, the board and the individual researcher might focus more on the research question rather than time-consuming bureaucratic processes. Without this general approval for hypothesis testing, researchers would normally not be able to query the data without extensive knowledge regarding the infrastructure and the law. Furthermore, the UKE has established an independent trust center, which is largely autonomous in legal terms. This center uses suitable pseudonymization techniques to safeguard patient data identity.

Metadata Processing

The processing of metadata is crucial in the context of data harmonization with multiple data sources, as intended in this project. Metadata repositories (MDRs) enable the structuring of data for the technical extract, transform, and load (ETL) process. They are also applications that make the syntax and semantics of the data understandable for the end user. Both attributes are relevant in our context. Numerous systems have already been tested and evaluated in use [15-17]. In this case, we prefer an MDR that is a further development of the already used Samply.MDR [19], an ISO/IEC 11179-based metadata repository built on a graph-based backend, making the MDR applicable to many hierarchical data structures.

Quantitative and Qualitative Analysis

In the evaluation of the proposed tool, a dual assessment was conducted, encompassing a qualitative analysis of its suitability for replicating a public health-focused study and a quantitative examination of clinicians' usage behavior within the hospital.

Qualitative Analysis: Hypothesis Testing

The capabilities of the proposed platform for testing scientific hypotheses appear to be valuable for replicating studies in other cohorts. Comparable to existing publications [8], we applied the platform to underscore its efficacy in promoting sound scientific practices and for examining the generalizability of findings regarding the circumstances present at a specific hospital.

Due to its notable clinical implications [20,21], we chose PD as a neurodegenerative disorder of interest for which routine data may provide helpful insights. The International Parkinson and Movement Disorder Society (MDS) developed a scoring system to measure the severity of PD motor symptoms. This movement test is called the Unified Parkinson's Disease Rating Scale (UPDRS) and is widely used in clinical routine [22]. While postural, kinetic, and isometric tremor are subcategories of action tremor, the isometric tremor is difficult to measure in routine clinical settings and is not routinely assessed [23,24]. Nonetheless, the exact relationship between these distinct types of tremors remains incompletely understood.

Motivated by the findings of Gupta et al [25], our objective is to validate their proposed correlation between rest tremor and action tremor in patients with PD [26]. Consequently, we aim to replicate their observation of a significantly higher prevalence of action tremor in individuals also experiencing resting tremor.

By leveraging the proposed platform, we gained access to routine data, expanding beyond the use of public data sets used in the original study: The Parkinson Progression Marker Initiative (PPMI) [27], the Fox Investigation for New Discovery of Biomarkers (BioFIND) [28], and the Parkinson's Disease Biomarkers Program (PDBP) [29] data sets are 3 distinct clinical oriented, observational studies collecting relevant disease-specific data from patients with PD. The PPMI study focuses on early-stage patients with PD who have recently been diagnosed and are not yet receiving dopaminergic treatment. In contrast, the BioFIND and PDBP studies encompass patients at varying stages of PD, ranging from moderate to advanced and early to advanced, respectively. Consequently, the PPMI data set exclusively includes patients in the medication-off state, while the latter 2 data sets include patients in both the medication-off and medication-on states.

As the first step of the analysis, we identified those forms within the clinical information system used to store classifications according to the MDS-UPRDS. The platform facilitated the

selection of a well-defined cohort, ensuring precise inclusion criteria for the data query. Accordingly, we included all patients with the designated *ICD-10-GM* (*International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, German Modification*) code for Parkinson disease (G20). Furthermore, we limited our cohort to admissions that occurred between February 24, 2018, and February 24, 2023. Although age limitations can be applied, we did not impose any restrictions for the presented cohort. Leveraging the aforementioned criteria, we executed the query and retrieved all corresponding records stored in the system.

Quantitative Analysis

For the quantitative analysis, we focus on performance indicators critical for assessing the relevance of our platform in clinical reality. The practical use of the platform is measured with the cumulative probability distribution and the absolute number of requests after the initialization of the platform. The waiting times are critical for user experience, which is expected of the researcher's experience when they receive the requested data.

Ethical Considerations

Based on the proposed pipeline for pseudonymization and data security, the ethics committee of the Hamburg chamber of physicians agreed on approval for all hypothesis tests conducted through the platform (2022-100891-BO-ff).

Results

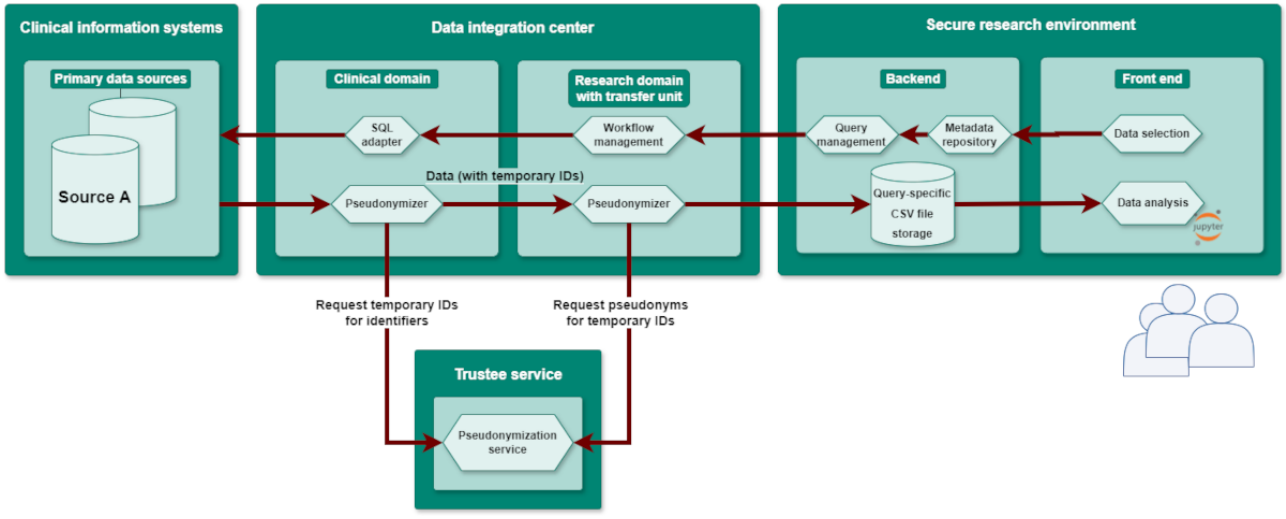
Technical Realization Within Clinical Reality

The central implementation detail of the platform is the usage of the established systems through strategic interfacing within the DIC. Notably, it circumvents the integration phase by directly querying the databases of clinical systems. The platform's backend is incorporated into the DIC's clinical and research domains, using the central trustee service for pseudonymization and the transfer unit for managing workflows and facilitating data delivery to researchers. The underlying architecture is constructed using standardized web technologies, specifically HTTPS and REST (Representational State Transfer). These procedural steps are fully automated, furnishing transparent feedback on the ongoing progress (Figure 1).

The data architecture is organized into 3 primary sections: clinical information systems, data integration center, and secure research environment. A researcher initiates hypothesis testing using the web application in the front end of the secure research environment. Data selection for this process is facilitated through a catalog of data elements supplied by the metadata repository. The data integration center, which is divided into 2 domains, the clinical domain, and the research domain, is responsible for converting the researcher's query into potentially multiple SQL (Structured Query Language) commands to retrieve data from the clinical information system's database. The output is exported as a CSV (comma-separated values) file and subjected to the pseudonymization triangle, where identifiers like medical record numbers and visit numbers are replaced with temporary IDs by the clinical domain before transmitting the data set to the research domain. The trustee service ensures that data from various sources are assigned consistent pseudonyms, maintaining the integrity of the hypothesis testing context. Subsequently, the pseudonymized data is stored in the query-specific CSV file storage within the backend of the secure research environment. Researchers have access to these data sets for further analysis and can use analytical tools such as Jupyter notebooks, which are available on the front end.

Text data or images that contain sensitive information are not exported. Subsequently, the research domain decodes temporary IDs into definitive pseudonyms and stores the data set for subsequent researcher access. Both the clinical and research domains obtain temporary IDs and pseudonyms from the trustee service, configured to issue unique ones for each identifier type (eg, medical record number or visit number). The linkage between original IDs and pseudonyms remains confidential to both domains; solely, the trustee service retains this information throughout the project's duration. After this process, the resulting file is automatically downloaded into a separate network designed for research and stored on a file system accessible only to the requesting clinician.

Figure 1. Components and data processing architecture of the platform. CSV: comma-separated value; SQL: Structured Query Language.



Using a widely recognized solution for both clinicians and data scientists, we incorporated JupyterLab, a prominent open-source web-based development environment, as the principal front end for ensuing data analyses. As a result, the proposed platform empowers users to leverage a diverse array of tools and libraries on the 639 forms encompassing 10,629 data elements that we have made accessible from clinical routine.

Qualitative Results: Example of a Hypothesis Testing

For the qualitative assessment, all patients with the designated *ICD-10-GM* code G20 and admissions that occurred between February 24, 2018, and February 24, 2023, were included. The majority of patients underwent multiple assessments during their hospital visit. We only considered the first assessment to ensure independent samples and discarded subsequent assessments. This was necessary since we were mainly interested in the patient cohort with any of the 3 basic tremor types rather than the overall occurrence of all tremors assessed at any given time point. The decision to choose the first assessment was made since not every subject was assessed more than once, but always at the beginning of their hospitalization, thereby ensuring uniformity in tremor severity assessments shortly after admission. Afterward, we derived the subtypes described in the original work by considering the MDS-UPDRS items 3.17, 3.15, and 3.16 as surrogates for rest tremor, postural tremor of the hands, and kinetic tremor of the hands, respectively. As a result, we were able to include 777 patients in our qualitative assessment.

Table 1 presents the prevalence of the primary tremor types and the association between rest tremor and action tremor. The table includes 4 distinct data sets, with the first 3 data sets obtained from Gupta et al [25] and the fourth data set corresponding to our analysis conducted using the proposed tool (UKE). The provided values for rest tremor, postural tremor, and kinetic tremor represent the count of patients with PD exhibiting each respective tremor type while at rest, while holding their hands stretched out, or during a finger-to-nose maneuver, respectively. The severity rating for each tremor type is equal to or above 1, as outlined in the MDS-UPDRS guideline. The aggregated values presented in the table were derived following the published protocol.

Table 1. Comparison of the tremor subtypes and their occurrences within the cohorts reported by Gupta et al [25]. In addition, the last column shows the obtained results from routine data based on the proposed platform.

	PPMI[a] (N=423), n (%)	BioFIND[b] (N=118), n (%)	PDBP[c] (N=874), n (%)	UKE[d] (N=777), n (%)
Rest tremor	290 (68.6)	75 (63.6)	459 (52)	340 (43.8)
Pure rest tremor	87 (20.6)	15 (12.7)	104 (11.8)	57 (7.3)
Action tremor	156 (36.9)	46 (39)	316 (35.8)	255 (32.8)
Pure action tremor	40 (9.5)	10 (8.5)	87 (9.9)	76 (9.8)
Postural tremor	223 (52.7)	69 (58.5)	412 (46.7)	416 (53.5)
Pure postural tremor	18 (4.3)	8 (6.8)	31 (3.5)	72 (9.3)
Kinetic tremor	217 (51.3)	61 (51.7)	463 (52.5)	301 (38.7)
Pure kinetic tremor	23 (5.4)	6 (5.1)	86 (9.8)	31 (4)
No tremor	52 (12.3)	19 (16.1)	211 (23.9)	258 (33.2)
Any tremor	317 (87.7)	99 (83.9)	663 (75.2)	519 (66.8)
All tremor	116 (27.4)	36 (30.5)	229 (26)	179 (23)

[a]PPMI: Parkinson Progression Marker Initiative.

[b]BioFIND: Fox Investigation for New Discovery of Biomarkers.

[c]PDBP: Parkinson's Disease Biomarkers Program.

[d]UKE: University Medical Center Hamburg-Eppendorf.

Our results represent a cohort of patients with PD irrespective of any dopaminergic treatment since many patients lack information regarding medication status due to the subsequent addition of this data field into the clinical information system. Through our data analysis, we observed a prevalence of 43.8% (340/777) for rest tremors and 7.3% (57/777) for pure rest tremors within the cohort. In contrast, the prevalence of total action tremor was 32.8% (255/777), with a corresponding occurrence of 9.8% (76/777) for pure action tremor. The incidences of postural tremor and pure postural tremor were found to be 53.5% (416/777) and 9.3% (72/777), respectively. We identified a prevalence of 38.7% (301/777) for kinetic tremor and 4.0% (31/777) for pure kinetic tremor. Finally, we calculated the occurrence of patients exhibiting all 3 tremor types simultaneously, the absence of any tremor, and the presence of at least 1 tremor type, resulting in proportions of 23.0% (179/777), 33.2% (258/777), and 66.8% (519/777), respectively. These relative figures closely resemble the reported values from the original authors. Importantly, we also observed a significantly higher proportion of action tremors in patients with rest tremors (43.8%) compared with those without rest tremors (32.8%), as determined by a chi-square test (P<.001).

Our analysis of routine data has yielded additional evidence that aligns with the published findings, suggesting that action tremor may be part of a broader tremor syndrome observed in PD. This discovery emphasizes the need for a more dynamic

approach to tremor classification, considering the progressive worsening of rest tremor severity over time [30] and its potential association with the occurrence of action tremor. Specifically, our data set corroborates the previous findings by Gupta et al [25], which propose a relationship between rest tremor and the emergence of action tremor. The data we have obtained further suggests that action tremor may represent a manifestation of the underlying basal ganglia disease, highlighting the potential requirement for additional neuroimaging studies to elucidate this connection.

Quantitative Results

In the realm of quantitative results, we focus on performance indicators critical for assessing the relevance of our platform within the clinical reality. To that end, we compiled a comprehensive list of successful queries executed using our proposed tool before October 30, 2023. Subsequently, we exclude queries carried out by members of the development teams, as they were primarily intended for debugging purposes.

Figure 2 illustrates the cumulative probability distribution of all incorporated queries across the temporal dimension. A conspicuous observation is the initial absence of queries in the early phase, signifying a notable delay in the adoption of the tool by clinical researchers, spanning nearly 6 months. The discernible rightward shift indicates an increasing interest among researchers following an initial habituation period. Nevertheless, the following data points reveal a marked acceleration in query use, with over 50% of the total queries executed within the most recent 3-month period.

Our platform offers the unique advantage of accessing multiple systems integral to clinical care. However, it's important to note that these platforms are not optimized for the specific nature of the queries in question. Substantial delays in data retrieval could significantly impede the quality of research conducted using our tool. Consequently, we analyzed to evaluate the waiting times experienced by clinical researchers before they received the requested data.

Figure 3 displays the distribution of the time it took for the queried data to become available to the researcher. The plot reveals a notable range of waiting times. While more than 50% of all requests were processed within a time frame of 50 hours, the longest queries extended to nearly a week. The pronounced initial ascent up to the median highlights the prompt reception of a substantial proportion of data despite the existence of instances where requests experience prolonged processing durations.

Figure 2. Cumulative probability distribution of researcher-initiated queries over time, starting from the public announcement of the platform, as extracted from the platform logs.

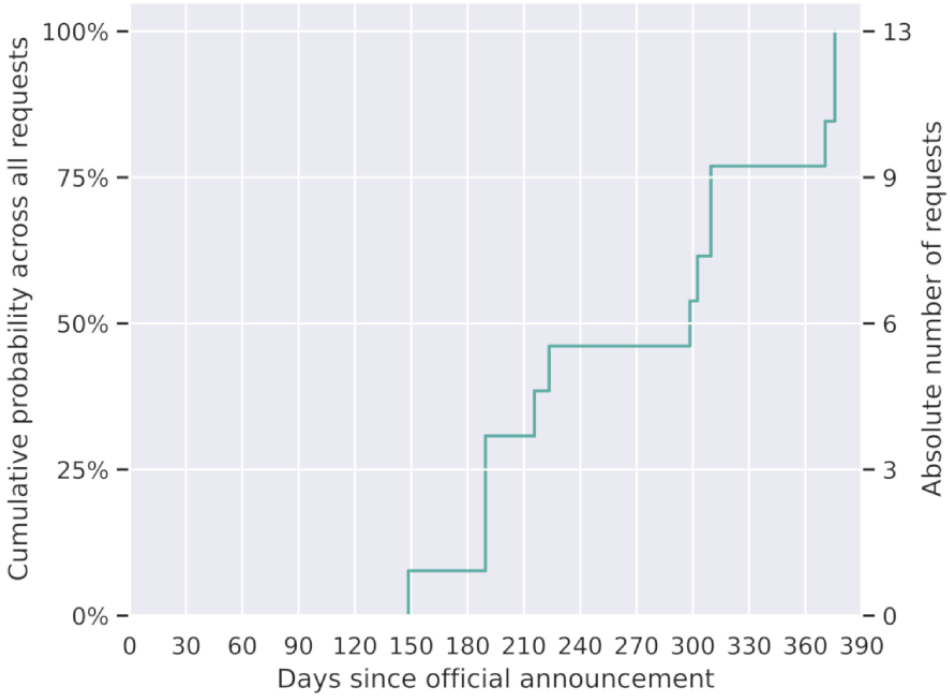
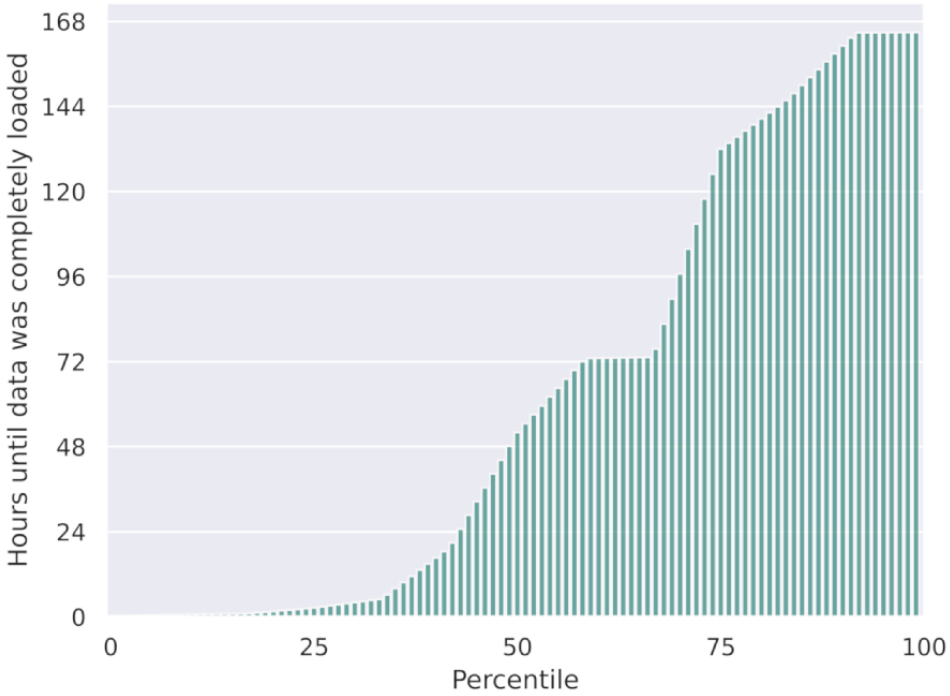


Figure 3. The percentiles illustrate all queries conducted by researchers until the availability of data for analysis extracted from the platform logs.



Discussion

Principal Findings

The development of a novel data platform at the University Medical Center Hamburg-Eppendorf for hypothesis testing on current clinical routine data according to all European and German data protection laws is accepted and used by clinicians. Accordingly, designing, implementing, and establishing a streamlined process for conducting hypothesis testing in public health by using secondary data appears possible. The initial version presented in this study involves the development of an analytical platform with a data protection-compliant infrastructure and a comprehensive ethical mandate, which will be extended with respect to semantic and syntactic interoperability found in the literature. This innovation has culminated in the establishment of a tool in clinical reality, which occupies a unique niche within the national health care landscape.

Given the illustrative use case, our findings indicate that routine data can facilitate the creation of data sets on scales comparable to prospective studies within significantly shorter time frames than those. This observation carries profound implications for diverse hospital roles: Patients gain transparency and trust in research processes, as the platform serves as a reliable authority for consent, enhancing confidence in the hospital's practices. Clinicians find empirical support for hypothesis testing, aiding in evidence-based decision-making and simplifying time-consuming replication studies. The Data Protection Officer benefits from automated queries, reducing project-related risk management burdens and minimizing infringement risks through a secure, tested architecture. Research data infrastructure experts receive structured support for handling researchers' queries. Finally, the hospital itself benefits from the efficient use of routine clinical data, offering potential cost savings, increased efficiency, and enhanced competitiveness.

Beyond the scope of our study, there is a discernible increase in interest in the tool within clinical reality. Over the last 90 days, the number of successful queries has doubled, and in total, clinical data from 600,000 patients or 1.6 million cases were retrieved from the platform. Although the absolute figures remain constrained, there is evident adoption by clinical researchers, indicating active use of the new tool for their hypothesis tests.

Limitations

Our investigation underscores that the long execution times of queries on general-purpose databases in clinical systems, which are not inherently designed for the queries executed by the research platform, can limit the interactivity of researchers with the clinical data. Similar systems in other hospitals may likely face comparable issues. The complexity of supporting various query formulations through SQL query adapters further complicates optimization, often resulting in less efficient query statements compared with those that are meticulously crafted by hand. To enhance our platform and achieve shorter execution times, further development is essential. By now, we established a time limitation for queries, terminating excessively large ones. In the future, we plan to use strategies such as horizontally scaling the data sources, using alternative data stores or data caches, or using FHIR search or the Clinical Query Language as the query mechanism instead of traditional SQL [31].

Furthermore, our observations indicate a lack of universal intuitiveness among clinical users in our hospital regarding the Jupyter Notebooks used for analysis. Despite the formulation of data queries, the execution of analyses experienced a notable decline. The participation of clinicians in platform design underscores a potential gap in data literacy among individual

physicians. To mitigate this, we advocate for an additional reduction in the entry barrier through the introduction of user-friendly, broadly applicable dashboards and visualizations tailored to each data query.

An additional aspect that holds potential for enhancing usability in the future is the ability to share access to analysis spaces. This feature would enable users with limited statistical expertise to invite statistical or biomedical experts into their analysis space, gradually receiving support throughout the analysis process. By allowing collaborative access, inexperienced users can benefit from the guidance and assistance of domain experts, facilitating their learning and development in statistical analysis. Accordingly, this feature is currently under development.

Conclusions

With the presented research platform, we were able to establish a valuable tool for hypothesis testing and secondary use of clinical data. By automating the retrieval process of pseudonymized clinical data and providing a clear legal framework, the platform contributes to the facilitation of the research process. The practical usability of the platform was demonstrated through the replication of a scientific study using the example of PD, confirming the validity of the concept. In further development stages and through the integration of additional clinical data sources, we aim to continuously increase the quantity of data and the usability of the platform. In the long term, through further modularization and standardization, the platform should be made usable for additional national and European sites, significantly facilitating the secondary use of clinical data.

Acknowledgments

The authors received no direct funding for this work. However, we gratefully acknowledge the financial support provided to the Institute for Applied Medical Informatics by the Dean’s Office of the University Medical Center Hamburg Eppendorf, which made this work possible. Furthermore, the work was made feasible through the assistance of the SMITH team at the Hamburg Eppendorf site, which is part of the SMITH consortium funded by the German Federal Ministry of Education and Research (BMBF; grant 01ZZ1803O). We also acknowledge financial support from the Open Access Publication Fund of UKE (Universitätsklinikum Hamburg-Eppendorf) and DFG (German Research Foundation).

During the preparation of this manuscript, the authors used (generative) artificial intelligence–powered tools like DeepL Translator, Grammarly (Grammarly Inc), and ChatGPT (Open AI) to improve the language and style of some parts of the paper. These tools were not used to generate any content of the paper itself.

Data Availability

The anonymized data sets regarding the usage behavior analyzed during this study are available in the ZFDM repository [32]. The clinical data sets are not publicly available due to the missing legal foundation for the export of routine data but are available from the corresponding author on reasonable request and given permission from the required access committees.

Authors' Contributions

CG and KG wrote the manuscript. CG and AJW conducted the analysis. MA, MW, and JG led the technical development. FÜ, MW, and JG edited the manuscript. CG is the corresponding author of this paper.

Conflicts of Interest

None declared.

References

1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34-49. [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
2. Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med. 2018;57(S 01):e50-e56. [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
3. Gehring S, Eulenfeld R. German medical informatics initiative: unlocking data for research and health care. Methods Inf Med. 2018;57(S 01):e46-e49. [FREE Full text] [doi: 10.3414/ME18-13-0001] [Medline: 30016817]
4. Medizin Informatik Initiative. Der kerndatensatz der medizininformatik-initiative. URL: https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative [accessed 2023-11-23]
5. Nerius M, Fink A, Doblhammer G. Parkinson's disease in Germany: prevalence and incidence based on health claims data. Acta Neurol Scand. 2017;136(5):386-392. [FREE Full text] [doi: 10.1111/ane.12694] [Medline: 27726128]
6. Choe C, Petersen E, Lezius S, Cheng B, Schulz R, Buhmann C, et al. Association of lipid levels with motor and cognitive function and decline in advanced Parkinson's disease in the Mark-PD study. Parkinsonism Relat Disord. 2021;85:5-10. [doi: 10.1016/j.parkreldis.2021.02.007] [Medline: 33636481]
7. Soto GE, Spertus JA. EPOCH® and ePRISM®: a web-based translational framework for bridging outcomes research and clinical practice. 2007 Computers in Cardiology. 2007:205-208. [FREE Full text] [doi: 10.1109/cic.2007.4745457]

8. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. KETOS: clinical decision support and machine learning as a service - a training and deployment platform based on docker, OMOP-CDM, and FHIR web services. PLoS One. 2019;14(10):e0223010. [FREE Full text] [doi: 10.1371/journal.pone.0223010] [Medline: 31581246]
9. Valenzuela W, Balsiger F, Wiest R, Scheidegger O. Medical-blocks-a platform for exploration, management, analysis, and sharing of data in biomedical research: system development and integration results. JMIR Form Res. 2022;6(4):e32287. [FREE Full text] [doi: 10.2196/32287] [Medline: 35232718]
10. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform. 2015;53:162-173. [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]
11. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-S37. [FREE Full text] [doi: 10.1097/MLR.0b013e31829b1dbd] [Medline: 23774517]
12. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/ [accessed 2023-06-01]
13. Gradient MLOps Platform. URL: https://www.paperspace.com/gradient [accessed 2023-06-01]
14. Google Colaboratory. URL: https://colab.research.google.com/ [accessed 2023-06-01]
15. CodaLab. URL: https://codalab.org/ [accessed 2023-06-01]
16. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart medical information technology for healthcare (SMITH). Methods Inf Med. 2018;57(S 01):e92-e105. [FREE Full text] [doi: 10.3414/ME18-02-0004] [Medline: 30016815]
17. Prokosch H-U, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical informatics in research and care in university medicine. Methods Inf Med. 2018;57(S 01):e82-e91. [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
18. § 12 HmbKHG Hamburgisches Krankenhausgesetz (HmbKHG), Gesetze des Bundes und der Länder. 1991. URL: https://www.landesrecht-hamburg.de/bsha/document/jlr-KHGHArahmen [accessed 2023-06-01]
19. Kadioglu D, Breil B, Knell C, Lablans M, Mate S, Schlue D, et al. Samply.MDR – A metadata repository and its application in various research networks. In: German Medical Data Sciences: A Learning Healthcare System. Amsterdam, Netherlands. IOS Press; 2018:50-54.
20. DeMaagd G, Philip A. Parkinson's disease and its management: part 1: disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis. P T. 2015;40(8):504-532. [FREE Full text] [Medline: 26236139]
21. Rizek P, Kumar N, Jog MS. An update on the diagnosis and treatment of Parkinson disease. CMAJ. 2016;188(16):1157-1165. [FREE Full text] [doi: 10.1503/cmaj.151179] [Medline: 27221269]
22. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society UPDRS Revision Task Force. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord. 2008;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
23. Bhatia KP, Bain P, Bajaj N, Elble RJ, Hallett M, Louis ED, et al. Tremor Task Force of the International Parkinson and Movement Disorder Society. Consensus statement on the classification of tremors. from the task force on tremor of the international Parkinson and movement disorder society. Mov Disord. 2018;33(1):75-87. [FREE Full text] [doi: 10.1002/mds.27121] [Medline: 29193359]
24. Lenka A, Jankovic J. Tremor syndromes: an updated review. Front Neurol. 2021;12:684835. [FREE Full text] [doi: 10.3389/fneur.2021.684835] [Medline: 34381412]
25. Gupta DK, Marano M, Zweber C, Boyd JT, Kuo S-H. Prevalence and relationship of rest tremor and action tremor in Parkinson's disease. Tremor Other Hyperkinet Mov (N Y). 2020;10:58. [FREE Full text] [doi: 10.5334/tohm.552] [Medline: 33384882]
26. Louis ED, Levy G, Côte LJ, Mejia H, Fahn S, Marder K. Clinical correlates of action tremor in Parkinson disease. Arch Neurol. 2001;58(10):1630-1634. [doi: 10.1001/archneur.58.10.1630] [Medline: 11594921]
27. Parkinson Progression Marker Initiative. The Parkinson progression marker initiative (PPMI). Prog Neurobiol. 2011;95(4):629-635. [FREE Full text] [doi: 10.1016/j.pneurobio.2011.09.005] [Medline: 21930184]
28. Kang U, Alcalay R, Goldman J, Henchcliffe C, Hogarth P, Tuite P, et al. The BioFIND study (Fox investigation For new discovery of biomarkers In Parkinson's disease): design and methodology (P4.043). Neurology Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology; 2014. URL: https://n.neurology.org/content/82/10_Supplement/P4.043 [accessed 2023-06-01]
29. Rosenthal LS, Drake D, Alcalay RN, Babcock D, Bowman FD, Chen-Plotkin A, et al. PDBP consortium. The NINDS Parkinson's disease biomarkers program. Mov Disord. 2016;31(6):915-923. [FREE Full text] [doi: 10.1002/mds.26438] [Medline: 26442452]
30. Jagadeesan AJ, Murugesan R, Vimala Devi S, Meera M, Madhumala G, Vishwanathan Padmaja M, et al. Current trends in etiology, prognosis and therapeutic aspects of Parkinson's disease: a review. Acta Biomed. 2017;88(3):249-262. [FREE Full text] [doi: 10.23750/abm.v88i3.6063] [Medline: 29083328]

31. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch H, et al. The architecture of a feasibility query portal for distributed COVID-19 fast healthcare interoperability resources (FHIR) patient data repositories: design and implementation study. JMIR Med Inform. 2022;10(5):e36709. [FREE Full text] [doi: 10.2196/36709] [Medline: 35486893]
32. Gundler C. Dataset for unlocking the potential of secondary data for public health research: a retrospective study with a novel clinical platform. ZFDM Repository; 2023. URL: https://www.fdr.uni-hamburg.de/record/13839 [accessed 2024-09-04]

Abbreviations

BioFIND: Fox Investigation for New Discovery of Biomarkers
CSV: comma-separated values
DIC: data integration center
EHR4CR: Electronic Health Records for Clinical Research
ETL: Extract, Transform, Load
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
***ICD-10-GM*:** *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, German Modification*
MDR: Metadata Repository
MDS: International Parkinson and Movement Disorder Society
MII: Medical Informatics Initiative
PD: Parkinson disease
PDBP: Parkinson's Disease Biomarkers Program
PPMI: Parkinson Progression Marker Initiative
REST: Representational State Transfer
SQL: Structured Query Language
UKE: University Medical Center Hamburg-Eppendorf
UPDRS: Unified Parkinson's Disease Rating Scale

Edited by T de Azevedo Cardoso; submitted 03.08.23; peer-reviewed by C Gulden, T Karen, K Fultz Hollis; comments to author 05.10.23; revised version received 01.12.23; accepted 17.07.24; published 01.10.24

Please cite as:
Gundler C, Gottfried K, Wiederhold AJ, Ataian M, Wurlitzer M, Gewehr JE, Ückert F
Unlocking the Potential of Secondary Data for Public Health Research: Retrospective Study With a Novel Clinical Platform
Interact J Med Res 2024;13:e51563
URL: https://www.i-jmr.org/2024/1/e51563
doi: 10.2196/51563
PMID:

sensors

MDPI

Article

Opportunities and Limitations of Wrist-Worn Devices for Dyskinesia Detection in Parkinson's Disease

Alexander Johannes Wiederhold [1,*], Qi Rui Zhu [1], Sören Spiegel [1], Adrin Dadkhah [2,3], Monika Pötter-Nerger [4], Claudia Langebrake [2], Frank Ückert [1] and Christopher Gundler [1]

[1] Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[2] Hospital Pharmacy, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[3] Department of Stem Cell Transplantation, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
[4] Institute of Neurology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany
* Correspondence: a.wiederhold@uke.de; Tel.: +49-176-70073562

Highlights

- Sensor-driven measurements can support the assessment of dyskinesia fluctuations in clinical practice.
- Dimensional reduction of accelerometer data using PCA improves machine learning model performance.
- Semantic feature extraction enhances model generalization and predictive capability for dyskinesia detection.
- Integrating standardized neurological assessments can further improve the reliability of sensor-based monitoring.

Abstract

During the in-hospital optimization of dopaminergic dosage for Parkinson's disease, drug-induced dyskinesias emerge as a common side effect. Wrist-worn devices present a substantial opportunity for continuous movement recording and the supportive identification of these dyskinesias. To bridge the gap between dyskinesia assessment and machine learning-enabled detection, the recorded information requires meaningful data representations. This study evaluates and compares two distinct representations of sensor data: a task-dependent, semantically grounded approach and automatically extracted large-scale time-series features. Each representation was assessed on public datasets to identify the best-performing machine learning model and subsequently applied to our own collected dataset to assess generalizability. Data representations incorporating semantic knowledge demonstrated comparable or superior performance to reported works, with peak F_1 scores of 0.68. Generalization to our own dataset from clinical practice resulted in an observed F_1 score of 0.53 using both setups. These results highlight the potential of semantic movement data analysis for dyskinesia detection. Dimensionality reduction in accelerometer-based movement data positively impacts performance, and models trained with semantically obtained features avoid overfitting. Expanding cohorts with standardized neurological assessments labeled by medical experts is essential for further improvements.

check for updates

Academic Editor: Lorenzo Scalise

Received: 16 June 2025
Revised: 12 July 2025
Accepted: 17 July 2025
Published: 21 July 2025

Citation: Wiederhold, A.J.; Zhu, Q.R.; Spiegel, S.; Dadkhah, A.; Pötter-Nerger, M.; Langebrake, C.; Ückert, F.; Gundler, C. Opportunities and Limitations of Wrist-Worn Devices for Dyskinesia Detection in Parkinson's Disease. *Sensors* **2025**, *25*, 4514. https://doi.org/10.3390/s25144514

Keywords: Parkinson's disease; dyskinesia; wearable devices; semantic data analysis; tsfresh; machine learning; accelerometer data; supportive decision making; feature extraction; principal component analysis

1. Introduction

The optimization and fine-tuning of therapy in response to the progression of Parkinson's disease (PD) necessitate in-hospital diagnostic assessments and evaluations by means of clinical rating scales [1,2]. The overdosing of dopaminergic drugs often evokes levodopa-induced dyskinesia (LID), an uncomfortable side effect that is characterized by uncontrolled, involuntary muscle movements of all body parts [3]. While many symptoms associated with PD may be more effectively identified through observational means, dyskinesias in the upper limbs hold significant potential for detection using wrist-worn sensors. The direct recording of health parameters provides vital insights into the ongoing progression of motor symptoms. Unlike a clinician who can only assess a snapshot of a moment, body-mounted systems are worn continuously throughout the day, offering a temporal recording of events. The acquired data is confined to the specific body part where the sensor is worn, yet it is tailored precisely to its designated tasks, thereby facilitating the capture of therapy-dependent motor fluctuations [4,5]. Subsequently, wearable devices such as smartwatches provide accessible monitoring aids, enabling the recording and tracking of movement data for understanding motor symptom variations and optimizing therapy plans [6–8].

Wearables offer potential benefits for observing daily fluctuations in a hospital setting, allowing an expert-guided labeling of symptoms. While various sensor-driven approaches for capturing tremor and bradykinesia have been explored, there is a limited focus on dyskinesia detection despite its potential as a key biomarker for therapy responses [9,10]. Alongside other reports, studies conducted by Hssayeni et al. and Pfister et al. aim to assess dyskinesia in a free-living environment to be able to monitor the disease condition at home [11,12]. Hssayeni et al. seeks to estimate dyskinesia using accelerometer data during daily activities, reporting a Pearson correlation within the range of 0.70 to 0.84. In a similar vein, Pfister et al. reports the capability to detect dyskinesias in a free-living environment, achieving a sensitivity/specificity of 0.64/0.89. Acknowledging the importance of understanding the symptoms impact on everyday activities, there is clinical relevance in observing and estimating its occurrence during a hospital stay. When patients visit the clinic, neurologists adjust dopaminergic pharmaceuticals, leading to the frequent appearance of LID [3,13]. Monitoring these side effects during hospital admission could provide insights into the symptomatic fluctuations of patients, potentially aiding in the determination of the optimal drug dosage for individuals and supporting clinicians in addressing LIDs before patients are discharged. Sieberts et al. conducted a study referred to as the DREAM Challenge, which aligns with this suggestion by identifying biomarkers linked to tremor, bradykinesia, and dyskinesia in PD. Utilizing public datasets, the DREAM Challenge aimed to predict the severity of PD symptoms, resulting in an AUPR (area under the precision-recall curve) of 0.48 specifically for dyskinesia [9].

Subsequently, we encourage the implementation of a novel monitoring setup specifically designed for dyskinesias emerging during hospital admission. Additionally, we propose an evaluation of the generalizability of movement data from publicly accessible datasets to our internally collected hospital data.

Thus, this paper employs two innovative approaches for movement data representation: one is a purely semantic technique utilizing principal component analysis (PCA) in

combination with biomechanical feature extraction, and the other is an automatic, serial feature representation. Following the training of these methods on publicly available datasets from the Michael J. Fox Foundation (MJFF) [14], our objective was to assess the performance of the resulting models on our own collected movement data and thus the models' generalizability. To fulfill our objective, we formulated two intents: (1) investigating the impact of various movement data representations on model performance and (2) evaluating the generalizability of machine learning (ML) models from publicly available datasets [14] onto the PACMAN (Parkinson's Clinical Movement Assessment) dataset.

2. Materials and Methods

2.1. Data

2.1.1. Public Datasets

We used two distinct and publicly available datasets from the MJFF: the Levodopa Response Study (LRS) and the Clinician Input Study (CIS-PD) [14,15]. Both studies aim to measure movement symptoms and their fluctuations of PD by means of accelerometers and according to the Unified Parkinson's Disease Rating Scale (UPDRS). These datasets were always retrieved together and referred to as MJFF dataset.

The LRS includes 28 patients diagnosed with PD that were monitored both in-clinic, where they engaged in a battery of standard activities, and at home while performing their daily activities. While the primary focus of the study is to comprehend motor fluctuations, patients were measured over four consecutive days with accelerometers. On the day of admission, UPDRS assessments were conducted at the clinic while patients were still on regular dopaminergic medication. Over the next two days, patients were released home, where they could carry out regular activities. On the last day, patients returned to the clinic and underwent similar UPDRS tests without any dopaminergic treatment [14].

The CIS-PD was a 6-month longitudinal investigation involving wearable tracking for 51 PD patients. The study encompassed clinic visits and at-home monitoring using smartwatches. Following the baseline assessment, in-clinic visits were scheduled at 2 weeks, 1 month, 3 months, and 6 months. During these visits, clinicians conducted standard clinical assessments and reviewed data recorded at home. Between the hospital visits, the patients were asked to continue wearing their smartwatches and regularly report symptom severity and medication intake using a mobile phone app [15].

2.1.2. Data Collection of Our Own Movement Data (PACMAN)

PD patients admitted at the Department of Neurology at the University Medical Center Hamburg-Eppendorf (UKE) stay a minimum of two weeks during an inpatient care program, known as the Parkinson-Komplexbehandlung (PKB). A multidisciplinary team, including neurologists, physiotherapists, neuropsychologists, and other paramedical specialists, is dedicated to a patient-centered and individualized clinical approach in search of the optimal therapy [2,15]. This setting presents a distinctive opportunity for the continuous gathering of accelerometer data, accompanied by task-coupled severity scoring.

During our own 4-month clinical data collection, we raised 7 different standardized clinical examinations according to the third part of the Unified Parkinson Disease Rating Score (UPDRS III) per visit. An experienced neurologist decided on the examination criteria relevant for our hypothesis to detect dyskinesias during hospital admittance. The neurological task was required to not only be significant for the detection of upper limb dyskinesia but also measurable by a wrist-worn device. Hence, we decided on alternating hand movements, which include the supination and consequent pronation of the most affected hand (UPDRS 3.6). Further, we assumed that this periodic movement is projectable by semantic data representation and thus offers potential for clinically relevant feature extraction.

Each patient underwent a minimum of two daily visits over a span of up to two weeks, mirroring the duration of the PKB program [2]. During the initial consultation, the physician provided each patient with the watches and conducted the necessary setup each morning. Subsequently, the UPDRS 3.6 task was assessed, and the corresponding timestamp was annotated for accurate measurement. To ensure consistent data quality, the physician verified that the watch was worn tightly and in the correct orientation during each labeled assessment. The watch remained on the patient's most affected side, continuously capturing accelerometer data until the physician's return in the afternoon.

This non-interventional prospective cohort study used the accelerometer of an Apple Watch Series 6. The resulting dataset, further referred to as PACMAN, comprises movement data along with timestamped labels indicating symptom severity. The entire data collection transpired within the framework of clinical routine at an inpatient unit of the UKE and was carried out in accordance with relevant guidelines and regulations (Declaration of Helsinki). Written informed consent for the study was obtained from all participants and an approval from the Ethics Commission of the Ärztekammer Hamburg under the ID 2022-100846-BO-ff was granted beforehand.

2.2. Uniform Data Infrastructure

To achieve our goal of robustly detecting motor fluctuations, it was imperative to establish an infrastructure that could seamlessly integrate into clinical settings. This need was met by implementing a database adhering to the Fast Healthcare Interoperability Resources (FHIR) standard, ensuring the consistent storage and retrieval of movement data [16]. All acquired movement data is systematically deposited into the FHIR database, serving as an objective for subsequent comparative analyses. Once stored, a custom-designed data loader facilitates the retrieval of all measurements, allowing for tailored specifications and consistent loading of the requested data. Our team devised this uniform data architecture in a preceding project, with comprehensive details published elsewhere [16].

The process of obtaining all stored movement data involves a crucial consideration regarding the length of measurement samples. In the realm of sequential health data analysis, the term window size refers to the duration in which the data is examined. Given that different measurements possess unique recording lengths, the sample's window size can be of standardized length or dynamic. Opting for a predefined length ensures straightforward cross-database compatibility but may result in an exclusion of shorter measurements. Thus, setting the window size to a smaller value may omit a necessary task characteristic in some samples [17]. Therefore, we retain the full duration of each task as observed in the clinical setup and store them as variable-length samples. All subsequently chosen methods were selected to be compatible with this kind of temporal data with varying lengths. To further ensure comparability across sessions and participants, we applied a global rotation to compensate for the different coordinate systems of the chosen wearables. Since each sample was labeled in its entirety by a clinician, it is important to preserve the full temporal context. Subsequent classification directly relies on these clinical labels and thus benefits from maintaining the integrity of the original task as far as possible.

2.3. Data Representations

Representing movement data is crucial for its adequate analysis. A data representation is a way of encoding information into a format interpretable by an algorithm without losing significant meaning. Therefore, the selected representation format should align with the inherent characteristics of the measured data. Employing an appropriate data representation facilitates the extraction of meaningful features during preprocessing. The resulting features are, in turn, used to train our models for the detection of dyskinesias. We

opted for two distinct techniques of data representation that maintain the structure of the acquired accelerometer recordings.

2.3.1. Semantic Representation: PCA and Biomechanical Features

As movement data is recorded in a three-dimensional space by accelerometers, a reduction in dimensionality is amenable to visualization and analysis through human interaction. Given our focus on the alternating hand movements, we assumed that this periodic movement should be prominent in the data. This unique characteristic provides an opportunity for the representation of multi-dimensional movement data using semantic approaches. To reduce the complexity of our data, while still respecting the activity on each axis, we utilized a PCA.

PCA was applied to project the tri-axial accelerometer data into a single principal component, yielding a one-dimensional representation of movement over time. This projection facilitates semantic feature extraction, where distinct movement characteristics can be interpreted in a way that mirrors clinical assessment. As depicted in Figure 1, this allows clinicians and researchers alike to identify meaningful signal patterns, such as oscillation symmetry or amplitude modulation, using a reduced yet expressive representation.

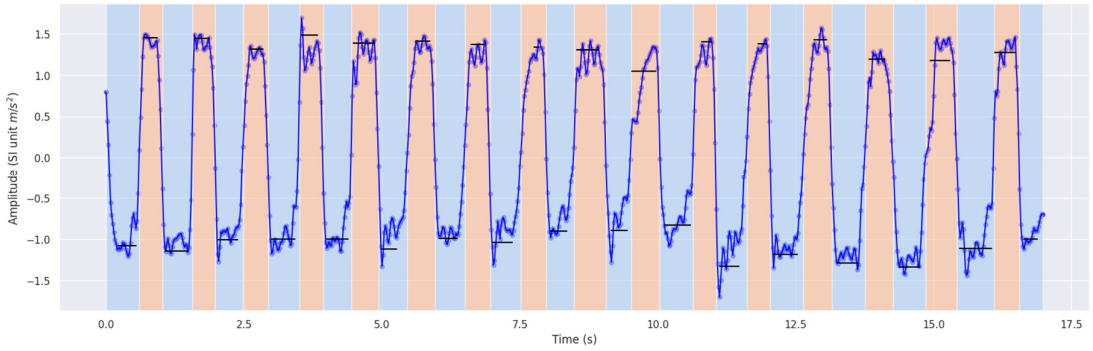


Figure 1. An exemplary segmented measurement. The blue segments have their extremum smaller than 0 and the orange segments show an extremum larger than 0. This segmentation is used for feature analysis.

Our semantic feature extraction follows the methodology introduced by Sánchez-Fernández et al., who identified 20 biomechanical features as particularly relevant for characterizing the alternating hand movement task (UPDRS 3.6) [18]. Their analysis, however, relied on multimodal sensor input (accelerometer, gyroscope, magnetometer). Since our dataset comprises only accelerometry data, we selected a subset of 9 clinically and technically interpretable features that can be derived robustly from this modality. These were selected through expert-driven, domain-specific relevance. The selected features are as follows:

- Feature for entire measurement:
  - Absolute mean of extremum
- Features for each part:
  - Number of segments comprising each part
  - Duration ratio of each part
  - Mean duration of each part
  - Interquartile range of each part
  - Relative maximum mean of each part
  - Relative maximum interquartile range of each part
  - Relative minimum mean of each part
  - Relative minimum interquartile range of each part

Next, to respect characteristic events of each individual measurement, we divided each sample into three distinct parts for the beginning, middle, and end. These parts are further divided into equally enduring segments. Figure 1 shows an example measurement with segments marked in orange or blue, defined by either the segment's maxima or minima, respectively. This segmentation was crucial for capturing temporal variations within the movement, such as oscillatory or acceleration changes, which are diagnostically relevant in Parkinson's disease.

2.3.2. Automatic Feature Extraction

In our alternative representation method, we utilized an automatic time-series feature extraction tool, further referred to as tsfresh, to automate the complex process of time-series engineering. Instead of human guidance, the Python library (Version: 0.20.1 on Python 3) identifies features by considering different algorithms of signal processing and time-series analysis to extract over 3000 features from temporal structured data [19]. This vast amount of mostly interpretable features is then systematically reduced through statistical tests.

As the automatically feature extraction can be applied on both the original three-dimensional as well as the reduced data, we considered three ways of movement data representation: (1) a fully semantic technique consisting of PCA and biomechanical feature extraction, (2) the PCA combined with the automatic feature extraction, and (3) automatic feature extraction only.

2.4. Training of the ML Models

Each of the three representation techniques yields distinctive features for the measured UPDRS task. A vast amount of these features, especially the numerous automatically extracted features, are not relevant to our research question. However, to identify an optimal ML model that maximizes performance based on only meaningful features, we aimed to determine the best combination of each representation with available models. The core component of the resulting data pipeline is the classifier, a predefined algorithm that assigns labels based on the provided features. To identify the classifier's highest performance, we integrated every possible combination into a grid search that predicts on the MJFF dataset.

A grid search is a systemic hyperparameter optimization that finds the optimal configuration of meaningful features, the classifier, and its hyperparameters by training each model separately [20]. Our implemented grid search used a stratified 10-fold cross validation prepared for imbalanced data [21]. We considered (1) a feature selector to find the ideal count of relevant features, (2) an oversampling technique to equalize for underrepresented labels (SMOTE) [22], and (3) different classifiers with numerate possibilities of their hyperparameters. As for the selection of parametric and non-parametric classification algorithms, we compared the performance of a logistic regression, a k-nearest neighbors' classifier, a random forest classifier, a support vector machine, and a gradient-boosting classifier.

2.5. Evaluation of the Resulting ML Models

Next, we determined the ten best-performing combinations of classifiers, selectors, and samplers for each of the three methods of feature representation on the MJFF dataset. The grid search results were ranked by the unweighted F_1 score, calculated as the arithmetic mean of all per-class F_1 scores. By combining precision and recall into a single metric, the F_1 score offers a balanced view of the model's performance across both classes. Although no consensus exists on metric selection for clinical relevance in this particular task [11,12] a high F_1 score suggests greater reliability in capturing label fluctuations, a critical requirement for potential therapeutic decision support. In addition to the F_1 score, the accuracy of each model's performance was also computed.

To ensure robustness of our findings, we conducted a conventional t-test, incorporating Welch's modification to accommodate potential variations and to assess significant performance disparities among different models. We expressed our outcomes as mean ± standard deviation, with a predefined threshold for statistical significance set at $p < 0.001$.

As a final step, we employed the top 10 models per representation, based on their unweighted F_1 performance, and implemented them on our own collected PACMAN movement data. The assessment of their performance utilized the same metric analysis and statistical significance to ensure comparability.

3. Results

3.1. The Patient Cohort

Our paper incorporated a total of 27 patients from the LRS dataset, spanning an age range of 50 to 84 years, with an average age of 67 years (±9 years). We included an average of 51 dyskinesia measurements (±9 measurements) per patient. Additionally, we included 24 patients from CIS-PD, with an average age of 63 years (±10 years), ranging from 36 to 75 years. Here, we used 12 dyskinesia measurements (±3 measurements) per patient on average. Finally, our in-clinic data collection contributed 25 patients, with an average age of 65 years (±8 years) and a range from 49 to 84 years. The PACMAN dataset incorporates an average of 3 dyskinesia measurements (±2 measurements) per patient.

3.2. Data Integrity

The presented data sources originate from distinct sites and were designed for different purposes. Nevertheless, they share comparability in terms of sensor type, demographics, neurological assessments, and disease-related intention of recording. The populations depicted in all MJFF studies fall within the same age range and undergo recurring hospital care for therapy adjustments. The types of sensors employed are consistent, as all studies integrate accelerometers, with both CIS-PD and PACMAN even utilizing Apple Watch devices.

Given that the MJFF studies involve a considerable number of ambulatory accelerometer recordings without precise physician annotations, our exclusive reliance on supervised labels was necessary to achieve our goal of identifying the most accurate representation of movement data for dyskinesia detection in a clinical setting. Consequently, all retained movement data and its corresponding labels are derived from hospital admittance and adhere to UPDRS standards.

3.3. Performance on Training Datasets

Figure 2 illustrates the confusion matrix of the leading model for each data representation on the MJFF studies. This visualization simultaneously presents the assigned dyskinesia label and the model's prediction. Hence, the confusion matrix provides a reliable means to identify instances when a model incorrectly labeled a class. All depicted models adequately identified the absence of dyskinesia. However, they encountered challenges in accurately recognizing dyskinesia.

When it comes to the full evaluation of performance, we ranked the predictive outcomes for each representation by the unweighted F_1 score. Here, we achieved peak performances of 0.68 with both data representations, employing solely automatically extracted features and the automatically extracted features on the one-dimensional representation. The semantic representation using PCA and biomechanical features yielded a slightly lower performance, registering 0.63 for unweighted F_1. Accuracies of all three extraction methods

were as high as 0.89 for each representation. The top five models per representation are listed in Tables 1–3.

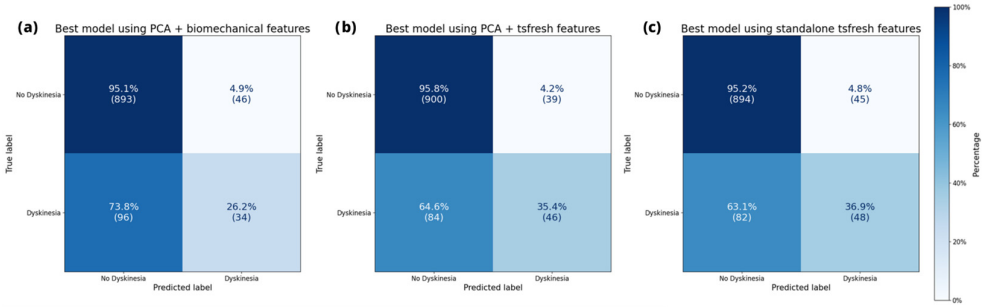


Figure 2. Confusion matrix of the best-performing models for each representation method on the MJFF dataset.

Table 1. Results of top 5 models for representation: PCA and biomechanical features on the MJFF dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	gradient boosted trees (lr = 1.0)	10	none	0.63	0.87
2	gradient boosted trees (lr = 1.0)	none	none	0.62	0.88
3	random forest	5	SMOTE	0.61	0.89
4	gradient boosted trees (lr = 1.0)	none	SMOTE	0.60	0.83
5	k-nearest neighbors (nn = 5)	5	none	0.60	0.88

lr = learning rate, nn = number of neighbors, SMOTE = Synthetic Minority Over-sampling Technique.

Table 2. Results of top 5 models for representation: PCA and tsfresh features on the MJFF dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	random forest	none	SMOTE	0.68	0.88
2	random forest	10	none	0.68	0.88
3	random forest (depth = 5)	10	none	0.67	0.89
4	random forest (depth = 5)	none	SMOTE	0.67	0.84
5	gradient boosted trees (lr = 1.0)	10	none	0.66	0.86

lr = learning rate, SMOTE = Synthetic Minority Over-sampling Technique.

Table 3. Results of top 5 models for representation: PCA and standalone tsfresh features on the MJFF dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	random forest	10	none	0.68	0.88
2	random forest (depth = 5)	10	none	0.67	0.89
3	random forest (depth = none)	none	SMOTE	0.67	0.88
4	random forest (depth = 5)	none	SMOTE	0.67	0.83
5	random forest (depth = 5)	5	none	0.66	0.88

SMOTE = Synthetic Minority Over-sampling Technique.

Neither the purely semantic technique, nor the automatic feature extraction, alone or combined, achieved a significant difference compared to each other. Across all performed combinations, the average unweighted F_1 yielded scores of 0.53 ± 0.05, 0.57 ± 0.08, and 0.57 ± 0.08 for PCA with biomechanical features, PCA with feature extraction, and standalone feature extraction, respectively (mean $\pm$ standard deviation).

3.4. Generalization into Clinical Setting

While the performance on the public datasets though cross validations might approximate the generalizability on unseen data, we further tested this claim by utilizing the novel clinical data collected for this study. Application of the best 10 models per representation method on the PACMAN validation set yielded nuanced findings. The confusion matrix of the best model per representation, depicted in Figure 3, shows a similar classification pattern as on the MJFF studies.

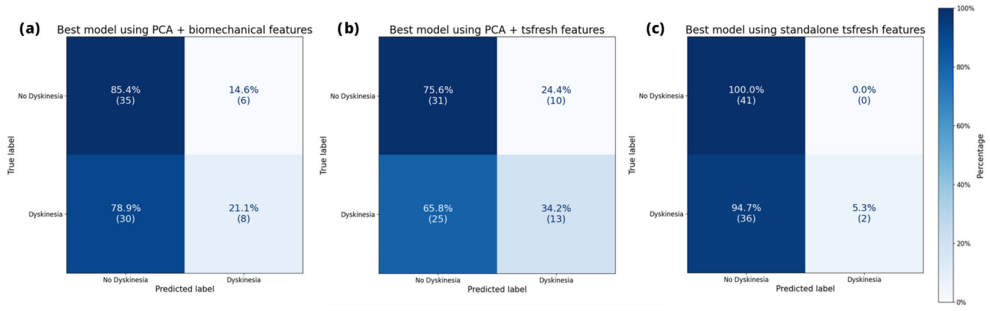


Figure 3. Confusion matrix of the best-performing models for each representation on the PACMAN dataset.

The top-performing representation for unweighted F_1 achieved a score of 0.53, utilizing PCA in conjunction with automatically extracted features. Following this was the semantic approach with a score of 0.48, and the standalone automatically extracted features ranked the lowest with a score of 0.40. The top five outcomes per representation of our generalization efforts are detailed in Tables 4–6, including all the parameters employed.

Table 4. Results of top 5 models for representation: PCA and biomechanical features on PACMAN dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	gradient boosted trees (lr = 1.0)	none	SMOTE	0.48	0.54
2	gradient boosted trees (lr = 1.0)	10	SMOTE	0.47	0.52
3	random forest	10	SMOTE	0.45	0.53
4	gradient boosted trees (lr = 1.0)	5	none	0.44	0.53
5	gradient boosted trees (lr = 1.0)	10	none	0.41	0.51

lr = learning rate, SMOTE = Synthetic Minority Over-sampling Technique.

Table 5. Results of top 5 models for representation: PCA and tsfresh features on PACMAN dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	gradient boosted trees (lr = 1.0)	10	none	0.53	0.56
2	gradient boosted trees (lr = 1.0)	none	none	0.44	0.56
3	random forest	5	none	0.42	0.54
4	random forest (depth = 5)	5	none	0.40	0.54
5	random forest (depth = 5)	none	SMOTE	0.38	0.51

lr = learning rate, SMOTE = Synthetic Minority Over-sampling Technique.

Overall, the three methods yielded average unweighted F_1 scores of 0.42 ± 0.04, 0.39 ± 0.06, and 0.36 ± 0.02 for PCA with biomechanical features, PCA with automatically extracted features, and automatically extracted features alone, respectively (mean $\pm$ standard deviation). Performances of the purely semantic representation were significantly higher than the automated technique ($p < 0.001$).

Table 6. Results of top 5 models for representation: standalone tsfresh features on PACMAN dataset.

Rank	Model			Macro F_1 Score	Accuracy
	Classifier	Selector	Sampling		
1	k-nearest neighbors (nn = 5)	10	none	0.37	0.54
2	random forest	10	none	0.37	0.53
3	random forest (depth = 5)	5	none	0.37	0.53
4	gradient boosted trees (lr = 1.0)	10	none	0.37	0.49
5	random forest (depth = 5)	none	SMOTE	0.36	0.52

lr = learning rate, nn = number of neighbors, SMOTE = Synthetic Minority Over-sampling Technique.

4. Discussion

This paper explores ML models for movement data representation, utilizing two distinct approaches and their combination. We first determined the top-performing models on the MJFF datasets and then evaluated their performance on our collected test dataset. Thereby, we were focusing on the impact of movement data representations and the generalizability of our resulting models for dyskinesia detection as our central hypotheses.

The results demonstrate how a dimensional reduction in movement data informed by the nature of the task has a positive impact on performance. Irrespective of whether combined with semantic or automatic extraction methods, the top 10 performing models incorporate this transformation when applied on our PACMAN dataset. The optimal performance is observed when the transformation is combined with automatic features. Nevertheless, the transformation combined with biomechanical features yielded comparable performance. This finding supports the idea that human-interpretable features enhance the ability of ML techniques to generalize across movement datasets. This finding confirms that semantically grounded preprocessing can serve as a safeguard against overfitting while supporting interpretability, a crucial factor in clinical implementation. The models using unselected features from a single automatic feature extraction generally show better results during training on the MJFF datasets but fail on our collected PACMAN data. Likely due to overfitting, it almost entirely fails to detect dyskinesia labels and assigns only two labels correctly. This underlines the importance of aligning feature representations with domain knowledge to improve robustness in real-world deployment.

Comparing the two approaches in detail, we observe that automated feature extraction (tsfresh) offers a wide array of statistical descriptors that may capture subtle signal characteristics, which contributes to strong performance on the structured and relatively homogeneous MJFF dataset. However, this approach appears to be less robust when applied to the clinically diverse PACMAN dataset, suggesting high sensitivity to variability in sensor noise, wearing conditions, and patient behavior. In contrast, the semantic representation consistently yields more stable results, even with limited and heterogeneous data. This suggests that the semantic approach not only improves interpretability for clinicians but also enhances robustness against real-world variability, which is critical for generalization across datasets. Therefore, while automated features may excel in high-data or controlled settings, semantically grounded representations prove more effective in noisy, low-data clinical environments, where reliability and explainability are essential.

Further, our analysis aligns closely with the performance reported in the DREAM Challenge, which aimed to identify LIDs on MJFF datasets resulting in an AUPR of 0.48. However, the latter study did not evaluate distinct data representations nor test generalization on an independent dataset [9]. Previously, the mentioned studies by Hssayeni et al. and Pfister et al. report higher performances in dyskinesia detection, but their non-clinical setups are incomparable in terms of the sensor types used or a non-standardized assessment of dyskinesia [11,12]. Moreover, almost all the presented papers reveal distinct

metrics, which are incomparable to each other. As each metric analysis is favorable for the individual intention, these metrics are a challenge to evaluate in terms of comparability. To this end, we chose the F_1 score as the principal metric for performance evaluation, as it balances precision and recall, two properties particularly important in the clinical context where both false positives and false negatives can impact therapeutic decisions.

While there is currently no universally accepted threshold for the F_1 score in clinical ML applications, our observed values (MJFF: 0.68; PACMAN: 0.53) indicate a level of consistency and reliability that supports potential real-world use.

In order to lay the groundwork for ML generalization to work, standardization is required. First, the evaluative framework of studies working with supervised ML on movement data should use comparable and clinically relevant metrics. Analyses of medical data must account for class-specific performance, as predictions for each class need to be evaluated separately. Overall accuracy, for example, is insufficient in clinical settings, as it can obscure the detection of relevant disease phenotypes, particularly in imbalanced datasets. Secondly, the application of standardized neurological assessments, such as the suggested UPDRS, should be used. While some publications evaluate activities of daily living [11], these activities are not sufficiently reproducible and only play a minor role in therapeutic adjustments. These assessments lay the foundation of the task-specific data labeling and thus are essential for generalizability.

On the contrary, the data collection approach presented in this paper provides a unique opportunity to acquire standardized clinical assessments of movement distortions over a two-week period per patient. The dense data quality obtained per patient facilitates the detection of dyskinesias in a hospital setting enabling early identification of LIDs for medication adjustments before the patient is discharged.

Regarding clinical implementation of the presented movement data methodology, the generalizability of the fully semantic representation suggests great opportunities for future applications of wearables to detect LIDs during clinical stays. Our primary assumption, that the periodic alternation between supination and pronation of the hand are well projectable by a simple dimensional reduction, turns out to be valid. Adhering to clinical expertise and translating it directly into straightforward data representations suitable for machine learning algorithms significantly influences the final performance outcomes. Although real-time analysis was not the focus of this study, the short inference time of our trained models suggests feasibility for future real-time applications, such as adaptive therapy monitoring during inpatient stays. The dimensionally reduced representations demonstrated better generalization on the PACMAN dataset compared to the automatic features, which overfitted on the MJFF dataset. This supports the value of embedding domain knowledge into preprocessing pipelines, particularly when data availability is limited, a common challenge in real-world clinical contexts. The results on the PACMAN dataset further suggest the enormous potential for a combination of both techniques. Combining task-specific semantic dimensional reduction with automatic feature extraction may offer the best of both worlds, as this hybrid approach performed best on the PACMAN dataset. Both techniques are rooted in the assumption that the UPDRS examination, a clinically validated neurological scoring standard for over 30 years, provides a robust foundation for interpreting dyskinesia. Accordingly, the semantic pipeline partially mimics a clinician's process of evaluating movement patterns. A multidisciplinary approach between clinical expertise and data science is imperative for a successful application of this technology into routine.

However, this technology has its limitations for the detection of dyskinesia due to its unreliable predictive power. The relatively small size of our PACMAN dataset constrains the statistical power of our findings and likely contributes to variability in model

performance. This study was designed as an empirical step towards estimating minimal data requirements under clinical constraints, but future work must include larger, statistically powered datasets. Particularly, an expanded data collection of clinically annotated measurements is essential for tracking LIDs and assisting clinicians in optimizing therapy. Perhaps even synthetically generated movement data could provide a foundation to train and optimize models without the extensive need of gathering patients. Furthermore, comparability between datasets and standardization plays a pivotal role for the generalizability and its coherent ability to detect dyskinesias, as stated previously. Also, it remains unclear if other neurological assessments that are diagnostically relevant for the dyskinesia detection can be projected by semantic knowledge. This suggestion is also undermined by the smartwatch's limitation to detect movement of fingers or the hand. Nevertheless, the outcomes of this study demonstrate potential for the development of robust decision-support systems grounded in semantic principles, particularly in the analysis of clinically recorded movement data.

5. Conclusions

This paper presents a unique opportunity to gather standardized clinical assessments of movement distortions over a two-week period per patient, facilitating enhanced dyskinesia detection within a hospital setting. The utilization of inbuilt accelerometers in smart watches provides a stable and convenient solution to track movement distortions in hospitalized patients. Our investigation has identified objective movement data representations conducive to dyskinesia recognition and highlighted the semantic impact on sensor data analysis, especially when generalizing onto foreign datasets. Notably, semantic feature representations demonstrated more robust performance than automated features when applied to real-world clinical data. This robustness stems from their interpretability and alignment with established clinical rating schemes. The results suggest that semantically grounded preprocessing may offer a critical advantage in small-data or high-variability scenarios. Yet, further research with a larger cohort and standardized labeling protocols is essential to optimize therapy for levodopa-induced dyskinesias. The results of this work provide evidence of feasibility, suggesting that technology-based measurements have the potential to serve as supportive tools for comprehending symptom fluctuations during clinical practice.

Author Contributions: Conceptualization, A.J.W., Q.R.Z., S.S., A.D., M.P.-N., C.L., F.Ü. and C.G.; methodology, A.J.W. and C.G.; software, Q.R.Z., S.S., C.G. and A.J.W.; validation, A.J.W. and C.G.; formal analysis, A.J.W. and C.G.; investigation, A.J.W.; resources, A.J.W.; data curation, A.J.W.; writing—original draft preparation, A.J.W.; writing—review and editing, A.J.W., Q.R.Z., S.S., A.D., M.P.-N., C.L., F.Ü. and C.G.; visualization, A.J.W.; supervision, C.G.; project administration, A.J.W.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: No financial support was received for the preparation of this manuscript. Apple Inc. provided the Apple Watch devices for the research. Apple was not involved in the design of the research, nor was it involved in the collection, analysis, or interpretation of the research data, or the content of this or any related publication. We acknowledge financial support from the Open Access Publication Fund of UKE—Universitätsklinikum Hamburg-Eppendorf.

Institutional Review Board Statement: Participants in the study (PACMAN) granted informed consent. All procedures involving human participants adhered to the ethical standards of the institutional research committee and were in accordance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards. The collection of patient measurements received approval from the Ethics Commission of the Ärztekammer Hamburg under the ID 2022-100846-BO-ff.

Informed Consent Statement: Written informed consent was obtained from all subjects involved in the study (PACMAN).

Data Availability Statement: The code generated during the current study is publicly available in the repository of the University of Hamburg, (http://doi.org/10.25592/uhhfdm.14189 (accessed on 16 July 2025) since 18 July 2025.

Acknowledgments: The authors extend their appreciation to the Michael J. Fox Foundation for Parkinson's Research for generously providing access to the datasets associated with the Levodopa Response Study and the Clinician Input Study to the scientific community. A.J.W. expresses his gratitude to Sara Tiedemann for her valuable contributions during the manuscript preparation. During the preparation of this work the authors used DeepL Translator and ChatGPT 4o in order to enhance the language and style of specific sections. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

UKE	University Medical Center Hamburg-Eppendorf
PCA	Principal Component Analysis
tsfresh	Time-Series Feature Extractors
LID	Levodopa-Induced Dyskinesia
PD	Parkinson's Disease
MJFF	Michael J. Fox Foundation
UPDRS	Unified Parkinson's Disease Rating Scale
LRS	Levodopa Response Study
CIS-PD	Clinician Input Study-Parkinson's Disease
PKB	Parkinson-Komplexbehandlung
PACMAN	Parkinson's Clinical Movement Assessment
FHIR	Fast Healthcare Interoperability Resources
ML	Machine Learning
SMOTE	Synthetic Minority Over-Sampling Technique
AUPR	Area Under the Precision-Recall Curve
DREAM	Digital Biomarker Evaluation and Analysis for Mobile Health Challenge

References

1. Post, B.; Merkus, M.P.; de Bie, R.M.; de Haan, R.J.; Speelman, J.D. Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov. Disord.* **2005**, *20*, 1577–1584. [CrossRef]
2. Richter, D.; Bartig, D.; Muhlack, S.; Hartelt, E.; Scherbaum, R.; Katsanos, A.H.; Müller, T.; Jost, W.; Ebersbach, G.; Gold, R.; et al. Dynamics of Parkinson's Disease Multimodal Complex Treatment in Germany from 2010–2016: Patient Characteristics, Access to Treatment, and Formation of Regional Centers. *Cells* **2019**, *8*, 151. [CrossRef]
3. Kwon, D.K.; Kwatra, M.; Wang, J.; Ko, H.S. Levodopa-Induced Dyskinesia in Parkinson's Disease: Pathogenesis and Emerging Treatment Strategies. *Cells* **2022**, *11*, 3736. [CrossRef]
4. Maetzler, W.; Domingos, J.; Srulijes, K.; Ferreira, J.J.; Bloem, B.R. Quantitative wearable sensors for objective assessment of Parkinson's disease. *Mov. Disord.* **2013**, *28*, 1628–1637. [CrossRef]
5. Jalloul, N. Wearable sensors for the monitoring of movement disorders. *Biomed. J.* **2018**, *41*, 249–253. [CrossRef]
6. Bloem, B.R.; Post, E.; Hall, D.A. An Apple a Day to Keep the Parkinson's Disease Doctor Away? *Ann. Neurol.* **2023**, *93*, 681–685. [CrossRef]
7. Giannakopoulou, K.-M.; Roussaki, I.; Demestichas, K. Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review. *Sensors* **2022**, *22*, 1799. [CrossRef] [PubMed]
8. Tabatabaei, S.A.H.; Pedrosa, D.; Eggers, C.; Wullstein, M.; Kleinholdermann, U.; Fischer, P.; Sohrabi, K. Machine Learning Techniques for Parkinson's Disease Detection using Wearables during a Timed-up-and-Go-Test. *Curr. Dir. Biomed. Eng.* **2020**, *6*, 376–379. [CrossRef]

9. Sieberts, S.K.; Schaff, J.; Duda, M.; Pataki, B.Á.; Sun, M.; Snyder, P.; Daneault, J.-F.; Parisi, F.; Costante, G.; Rubin, U.; et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: The Parkinson's Disease Digital Biomarker DREAM Challenge. *npj Digit. Med.* **2021**, *4*, 53. [CrossRef] [PubMed]
10. Van Gerpen, J.A.; Kumar, N.; Bower, J.H.; Weigand, S.; Ahlskog, J.E. Levodopa-Associated Dyskinesia Risk Among Parkinson Disease Patients in Olmsted County, Minnesota, 1976-1990. *Arch. Neurol.* **2006**, *63*, 205–209. [CrossRef]
11. Hssayeni, M.D.; Jimenez-Shahed, J.; Burack, M.A.; Ghoraani, B. Dyskinesia estimation during activities of daily living using wearable motion sensors and deep recurrent networks. *Sci. Rep.* **2021**, *11*, 7865. [CrossRef]
12. Pfister, F.M.J.; Um, T.T.; Pichler, D.C.; Goschenhofer, J.; Abedinpour, K.; Lang, M.; Endo, S.; Ceballos-Baumann, A.O.; Hirche, S.; Bischl, B.; et al. High-Resolution Motor State Detection in Parkinson's Disease Using Convolutional Neural Networks. *Sci. Rep.* **2020**, *10*, 5860. [CrossRef]
13. Hughes, A.J.; Daniel, S.E.; Kilford, L.; Lees, A.J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **1992**, *55*, 181–184. [CrossRef]
14. Synapse.org. MJFF Levodopa Response Study [Internet]. 2019. Available online: https://www.synapse.org/Synapse:syn20681023/wiki/594678 (accessed on 16 July 2025).
15. Elm, J.J.; Daeschler, M.; Bataille, L.; Schneider, R.; Amara, A.; Espay, A.J.; Afek, M.; Admati, C.; Teklehaimanot, A.; Simuni, T. Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson's disease data. *npj Digit. Med.* **2019**, *2*, 95. [CrossRef] [PubMed]
16. Gundler, C.; Zhu, Q.R.; Trübe, L.; Dadkhah, A.; Gutowski, T.; Rosch, M.; Langebrake, C.; Nürnberg, S.; Baehr, M.; Ückert, F. A Unified Data Architecture for Assessing Motor Symptoms in Parkinson's Disease. *Stud. Health Technol. Inform.* **2023**, *307*, 22–30. [CrossRef] [PubMed]
17. Banos, O.; Galvez, J.-M.; Damas, M.; Pomares, H.; Rojas, I. Window Size Impact in Human Activity Recognition. *Sensors* **2014**, *14*, 6474–6499. [CrossRef] [PubMed]
18. Sánchez-Fernández, L.P.; Garza-Rodríguez, A.; Sánchez-Pérez, L.A.; Martínez-Hernández, J.M. A Computer Method for Pronation-Supination Assessment in Parkinson's Disease Based on Latent Space Representations of Biomechanical Indicators. *Bioengineering* **2023**, *10*, 588. [CrossRef]
19. Christ, M.; Braun, N.; Neuffer, J.; Kempa-Liehr, A.W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh—A Python package). *Neurocomputing* **2018**, *307*, 72–77. [CrossRef]
20. Liashchynskyi, P.; Liashchynskyi, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS [Internet]. *arXiv* **2019**, arXiv:1912.06059. Available online: http://arxiv.org/abs/1912.06059 (accessed on 27 March 2024).
21. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
22. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]

Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240499

Continuous, Learned Imputation of Missing Values in Parkinson's Disease

Christopher GUNDLER[a,1] and Monika PÖTTER-NERGER[b]
[a] *Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Germany*
[b] *Department of Neurology, University Medical Center Hamburg-Eppendorf, Germany*
ORCiD ID: Christopher Gundler https://orcid.org/0000-0001-9301-8872

Abstract. Parkinson's disease management requires accurate clinical scores but suffers from missing data. Leveraging self-supervised learning, we demonstrate superior generalization capabilities across populations compared to other well-established imputation techniques (MIWAE, MissForest, MICE). With the ability to employ the method already during the data collection and not afterward, the technology allows more robust data collection in clinical reality.

Keywords. Parkinson's disease, imputation, deep learning, self-supervised learning, missing value

1. Introduction

As a neurodegenerative disorder without a cure and significant impact on individuals and societies, managing Parkinson's disease requires careful and ongoing treatment to mitigate symptoms and maintain quality of life. Measuring motor and non-motor symptoms is required for tracking disease progression and treatment efficacy. Despite promising results for applying sensor technologies for movement analysis [1,2], the current state of the art in the clinical reality remains standardized assessment forms.

One example within neurology is the revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [3]. With sections for patients respective their caregivers and physicians, the form provides ordinal scaled assessments of motor and non-motor symptoms, commonly rated between 0 as "no symptom" to 4 for "significant symptoms". Despite its widespread and validated use, the application of the MDS-UPDRS is not without its challenges. As an example, missing values are common in ratings due to different clinical guidelines in hospitals or human error in filling out the sheets. The considerable effects of missing values when deriving scores are well documented [4].

In this work, we focus on an evaluation of imputation strategies for the missing values of MDS-UPDRS forms already during the data collection procedure. While imputation itself is a well-researched area, the algorithms are often employed just before the data analysis. Utilizing the imputation of missing predictors for scores as a supporting tool already during treatment is a newer yet promising area [5]. For the specific task of imputing missing MDS-UPDRS codes, a variety of methods from multivariate

[1] Corresponding Author: Christopher Gundler, Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany; E-mail: c.gundler@uke.de.

imputation by chained equations [3] (MICE) over copulas [6] and deep learning-based autoencoders [7] have been proposed and evaluated. However, the usage within the data collection process itself has rarely been evaluated to the best of our knowledge.

Given the current popularity of self-supervised learning in modern machine learning, considering corresponding models for continuous imputations appears attractive. Within this training paradigm, latent representations are obtained without explicit labels by deliberately corrupting or masking the training data [8]. The latter training paradigm resembles the imputation tasks considerably. On the one hand, some authors have shown evidence regarding the presence of a latent structure behind the individual UPDRS codes [9]. On the other hand, the additional complexity of deep learning methods is often considered not beneficial for classical imputation tasks [10]. Within the following, we will evaluate a deep learning architecture in comparison with well-established alternatives for their generalizability over scores from different populations as the central requirement for continuous imputation of missing items on previously unseen subjects.

2. Methods

2.1. Data

We employed a combination of observational and clinical datasets to develop and evaluate different imputation models. For training purposes, we primarily utilized data from the Parkinson's Progression Markers Initiative (PPMI), a large observational study providing comprehensive clinical assessments of Parkinson's disease progression [11]. For validation and hyperparameter optimization, we employed routine data automatically extracted from the electronic health records of the University Medical Center Hamburg-Eppendorf (UKE), comprising a substantial number of samples while containing the inherent noise associated with clinical reality. As a surrogate for the datasets found in smaller hospitals, we utilized scores from the DREAM challenge [12] as final test data.

2.2. Model architecture

As an existing foundation for our imputation model, we utilized the Masked Encoding for Tabular Data (MET) architecture [13]. This architecture represents a graphical model that explicitly considers the influences of variables on each other through attention and was optimized to learn linearly separable representations from datasets. Trained through self-supervision, the model reconstructed masked input during training. Treating missing values as masked, we mainly used standard hyperparameters utilized already by the authors, while reducing the number of masked values to 10%. The details of the architecture and hyperparameters can be found in the corresponding publication.

2.3. Training process

All models and baseline methods were trained on the PPMI data. The MET model utilized the UKE dataset for validation. We utilized a learning rate of 0.001 with a batch size of 64 and trained for 1000 epochs, implementing early stopping after 20 epochs without improvement. We conducted automatic hyperparameter optimization of the network architecture and the value of dropout.

2.4. Evaluation

For the evaluation, we utilized a subset of the imputation methods and their default hyperparameters implemented within the HyperImpute package [14]. More precisely, we choose MIWAE [15] as a deep learning approach, MICE and MissForest [16] as well-known and explainable methods, and a simple median imputation as the baseline. For the evaluation of performance, we generated 30 artificial datasets with 10% values missing at random from the test dataset. We choose the unweighted mean F_1-score to avoid unmeaningful results resulting from the class imbalance. For testing for significant performance differences, we employed a t-test with Welch's modification for potential changes in variances and set the threshold of statistical significance at $p < 0.01$.

3. Results

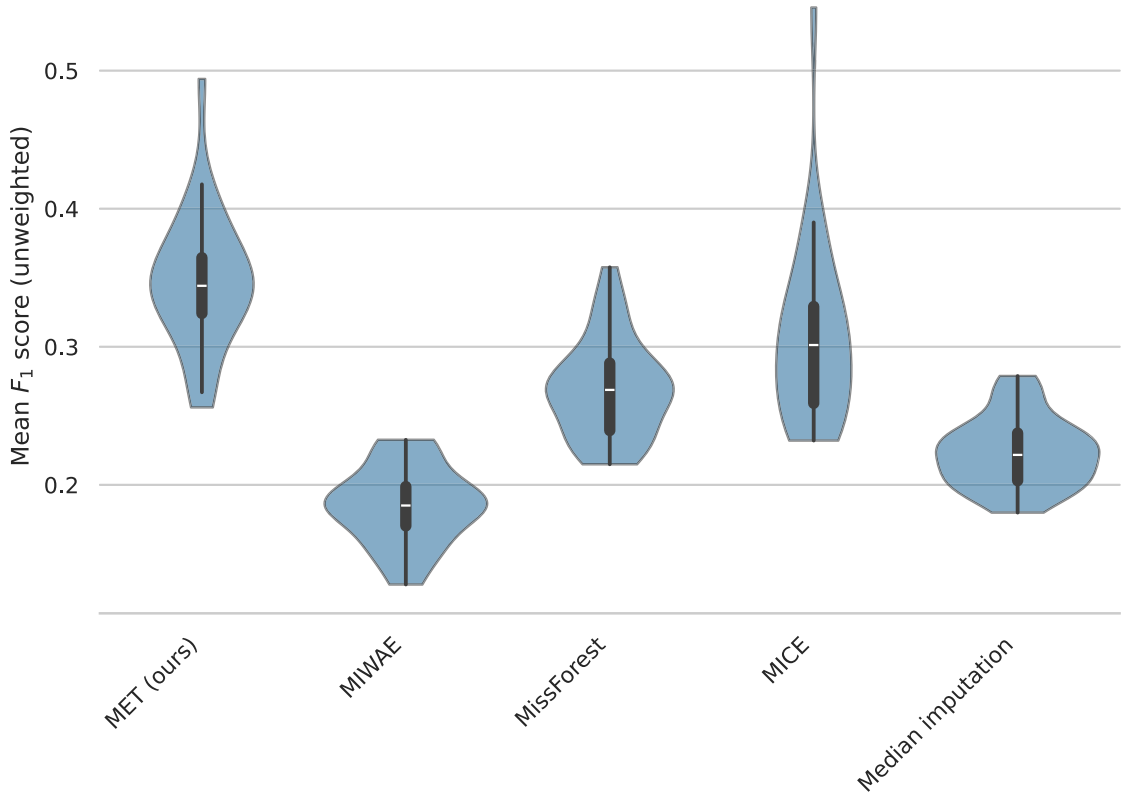


Figure 1. Performance of the different imputation methods measured through the unweighted mean F_1-score on the 30 artificially generated datasets from the test dataset. Our model shows significantly better performance while classic methods represent a tough baseline. Higher values are better.

In our comparative analysis of the imputation methods (Figure 1), we found notable variations in performance across the five approaches evaluated. Our model demonstrated the highest imputation accuracy, achieving a mean accuracy of 0.35 with a standard deviation of 0.05. Importantly, this performance was significantly better than that of the well-established MICE method, which achieved a mean accuracy of 0.31 ± 0.06. Even median imputation as the simplest baseline yielded a mean accuracy of 0.22 ± 0.03 and

was significantly better than the deep learning-based MIWAE with its mean accuracy of 0.18 ± 0.03. In general, the linear models emerged as unexpectedly powerful contenders in our analysis. We observed that MICE significantly outperformed MissForest, with a mean accuracy of 0.27 ± 0.04 compared to 0.23 ± 0.03, respectively.

4. Discussion

Our study demonstrates the potential of deep learning architecture learning latent representations for imputing missing MDS-UPDRS values during the data collection procedure itself. Leveraging self-supervised learning offers a robust strategy even across patient populations and diverse cohorts [8]. While the evaluated method statistically outperformed other imputation techniques in our evaluation setup, those alternative methods remain valid and offer distinct advantages. For example, MICE demonstrated impressive performance even in the generalization setup, showcasing both interpretability and significantly faster training times compared to our deep learning approach. Nevertheless, MICE does not provide an explicit and potentially linearly separable representation that could be leveraged for downstream tasks, a unique advantage of generative models [7].

When comparing our findings to those within the existing literature on imputation [4,5,10], it is crucial to recognize the differences in evaluation paradigms. While classical imputation methods are typically fitted on fully collected datasets where missing values appeared during collection, our methodology focuses on an arguably more challenging paradigm. We evaluate algorithms based on their ability to generalize to novel observations during the data collection process itself. These methodological differences likely contribute to the novel results observed in our study.

Despite the demonstrated strengths of our approach, certain uncertainties require future work. Firstly, further evaluation of our model's performance on tasks involving fine-tuning on fully collected datasets with missing values and, accordingly, the comparison with imputation algorithms in their "native setup." Additionally, addressing conditions where values are not missing at random but rather dependent on unobserved covariates, appears attractive to access the robustness and be comparable to the classical evaluation of imputation algorithms. Lastly, exploring the utility of the learned representations for downstream tasks, such as treatment efficacy assessments, holds promise for its practical application in clinical settings.

5. Conclusions

In conclusion, our study demonstrates the effectiveness and generalizability of learned latent representations for imputing missing clinical scores related to Parkinson's research across different cohorts. The technology might help in designing better systems to obtain reliable assessments, datasets with minimal missing useful for secondary usage, and representations useful for downstream tasks without additional effort.

Declarations

Conflict of Interest: The authors declare that there is no conflict of interest.
Contributions of the authors: CG planned and conducted the study. CG and FÜ were involved in writing and revision of the manuscript.

Acknowledgements: The DREAM Challenge data were generated by participants of The Michael J. Fox Foundation for Parkinson's Research Mobile or Wearable Studies. They were obtained as part of the Biomarker & Endpoint Assessment to Track Parkinson's Disease DREAM Challenge (through Synapse ID Synapse:syn20825169) made possible through partnership of The Michael J. Fox Foundation for Parkinson's Research, Sage Bionetworks, and BRAIN Commons. Additional data used in the preparation of this article were obtained on January 20 2024 from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR 006431. For up-to-date information and the funding partners of the study visit www.ppmi-info.org.

References

[1] Del Din S, Kirk C, Yarnall AJ, Rochester L, Hausdorff JM. Body-Worn Sensors for Remote Monitoring of Parkinson's Disease Motor Symptoms: Vision, State of the Art, and Challenges Ahead. Mirelman A, Dorsey ER, Brundin P, Bloem BR, editors. JPD. 2021;11(s1):S35–47.

[2] Sigcha L, Borzì L, Amato F, Rechichi I, Ramos-Romero C, Cárdenas A, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review. Expert Systems with Applications. 2023;229:120541.

[3] Luo S, Goetz CG, Choi D, Aggarwal S, Mestre TA, Stebbins GT. Resolving Missing Data from the Movement Disorder Society Unified Parkinson's Disease Rating Scale: Implications for Telemedicine. Movement Disorders. 2022;37(8):1749–55.

[4] Goetz CG, Luo S, Wang L, Tilley BC, LaPelle NR, Stebbins GT. Handling missing values in the MDS-UPDRS: HANDLING MDS-UPDRS MISSING VALUES. Mov Disord. 2015;30(12):1632–8.

[5] Nijman SWJ, Hoogland J, Groenhof TKJ, Brandjes M, Jacobs JJL, Bots ML, et al. Real-time imputation of missing predictor values in clinical practice. European Heart Journal - Digital Health. 2021;2(1):154–64.

[6] Houari R, Bounceur A, Kechadi T, Tari AK, Euler R. Missing Data Analysis Using Multiple Imputation in Relation to Parkinson's Disease. In: Proceedings of the International Conference on Big Data and Advanced Wireless Technologies [Internet]. Blagoevgrad Bulgaria: ACM; 2016 [cited 2024 Mar 12]. p. 1–6. doi: 10.1145/3010089.3010117.

[7] Peralta M, Jannin P, Haegelen C, Baxter JSH. Data imputation and compression for Parkinson's disease clinical questionnaires. Artificial Intelligence in Medicine. 2021;114:102051.

[8] Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng. 2022;6(12):1346–52.

[9] Gottipati G, Karlsson MO, Plan EL. Modeling a Composite Score in Parkinson's Disease Using Item Response Theory. AAPS J. 2017;19(3):837–45.

[10] Sun Y, Li J, Xu Y, Zhang T, Wang X. Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications. 2023;227:120201.

[11] Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, et al. The Parkinson Progression Marker Initiative (PPMI). Progress in Neurobiology. 2011;95(4):629–35.

[12] Sieberts SK, Schaff J, Duda M, Pataki BÁ, Sun M, Snyder P, et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. NPJ Digit Med. 2021;4(1):1–12.

[13] Majmundar K, Goyal S, Netrapalli P, Jain P. MET: Masked Encoding for Tabular Data. 2022 [cited 2024 Mar 16]; doi: 10.48550/ARXIV.2206.08564.

[14] Jarrett D, Cebere B, Liu T, Curth A, van der Schaar M. HyperImpute: Generalized Iterative Imputation with Automatic Model Selection. 2022 [cited 2024 Mar 14]; doi: 10.48550/ARXIV.2206.07769.

[15] Mattei PA, Frellsen J. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In: Proceedings of the 36th International Conference on Machine Learning [Internet]. PMLR; 2019 [cited 2024 Mar 15]. p. 4413–23.

[16] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.

sensors

MDPI

Article

Assessing the Generalizability of Foundation Models for the Recognition of Motor Examinations in Parkinson's Disease

Christopher Gundler [1,*], Alexander Johannes Wiederhold [1] and Monika Pötter-Nerger [2]

[1] Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany
[2] Department of Neurology, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany
* Correspondence: c.gundler@uke.de

Abstract

Current machine learning approaches focusing on motor symptoms in Parkinson's disease are commonly trained on small datasets and often lack generalizability from developmental setups to clinical applications. Foundation models using large, unlabeled datasets of healthy participants through self-supervised learning appear attractive for such setups with limited samples, despite the potential impact of motoric symptoms. Acting as an exemplar, this study aims to evaluate the robustness of fine-tuned models in recognizing movements related to motor examinations across datasets and recording setups. Accelerometer data of 51 participants with Parkinson's disease in three different training and fine-tuning setups were used to tailor the general model to the disease. Training the model on pre-trained weights, both partially (F_1 = 0.70) and fully (F_1 = 0.69), statistically significantly outperformed training the model from scratch (F_1 = 0.55) in a nested cross-validation. For evaluation, the model's ability to process data recorded from 24 patients in clinic was tested. The models achieved lower mean F_1 scores of 0.33 (train from scratch), 0.43 for full, and 0.48 for partial fine-tuning, but demonstrated improved generalizability and robustness regarding the orientation of sensors compared to training from scratch. Utilizing foundation models for accelerometer data trained on healthy participants and fine-tuned for clinical applications in movement disorders appears as an effective strategy for optimized generalizability with small datasets.

Keywords: Parkinson's disease; movement data; self-supervised learning; foundation model; generalizability; human activity recognition

Academic Editor: Hsin-Yi Kathy Cheng

Received: 8 July 2025
Revised: 29 August 2025
Accepted: 1 September 2025
Published: 4 September 2025

Citation: Gundler, C.; Wiederhold, A.J.; Pötter-Nerger, M. Assessing the Generalizability of Foundation Models for the Recognition of Motor Examinations in Parkinson's Disease. *Sensors* **2025**, *25*, 5523. https://doi.org/10.3390/s25175523

1. Introduction

Parkinson's disease (PD) is the most rapidly growing neurological disease with a doubling of its prevalence from 1990 to 2016 [1]. This progressive, neurodegenerative disorder is characterized by specific motor symptoms such as bradykinesia, muscle rigidity, tremor, and postural instability as well as various non-motor symptoms [2]. The disease progresses with a consecutive change in symptom load, urging a need for thorough clinical reassessment and constant therapy adjustments for effective management [3]. Consequently, for adequate adjustment of therapy, long-term monitoring is advisable. In clinical routine, the physician usually assesses the patient in the outpatient clinic every 3–6 months and receives only a single moment impression, but no objective measures over

the last weeks. Using movement data from smartphones, wearables, and related devices for motor assessments in PD could fill that information gap. Additionally, increasing research activities focus on on-demand, closed-loop stimulation systems to improve PD symptoms such as deep brain stimulation or biochemical sensing for pharmacological analytics [4,5] on a moment-to-moment basis. There, external wearables appear attractive as a biomarker of the motor state, e.g., to detect freezing of gait episodes [6]. Accordingly, device-aided movement tracking has become widespread in research and discussed with its challenges and opportunities [7].

Optimized for activity recognition, sensor data from accelerometers, gyroscopes, and magnetometers are commonly used in related studies [8]. Some studies focus on experimental setups in which patients perform specific tasks designed to elicit symptoms, and these tasks are then assessed automatically [9,10]. Other authors try to derive disease-relevant indicators for monitoring sensor assessments in real-world settings, capturing a more naturalistic representation of daily activities but without rater-based matching [11,12]. In both cases, the continuous sensor signal must often be segmented into different parts of activity, depending on the research question of interest. For example, detecting freezing of gait only makes sense when a patient is moving, and assessing some items of the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) only makes sense when the related movement is performed. These segmentations themselves may have different origins; physicians, scientists, or even patients often annotate them manually during studies. In real-world data, extracting those segments automatically for healthy participants is possible and is still improved in current research under the name of human activity recognition. In case of research regarding PD, related work is more limited; for example, Yue et al. [13] describe a similar setup. Some authors utilize their own data, while others use existing models trained on public data and assume they will still perform for patients despite their characteristic motoric symptoms [14].

The Issue of Generalization into Clinical Reality

The limited methodical considerations regarding the precise method for obtaining the associated activities appear amplified given the requirements regarding the robustness and generalizability of the approaches. This challenge is multifaceted. Firstly, the overall data volume is constrained. Although efforts have been made to consolidate diverse data sources for larger cohorts, the available data remains comparatively limited, particularly when compared to tasks such as general activity recognition. The problem becomes even worse as the available samples differ significantly between cohorts. Often, observational studies with multitudes of measurements recorded in home setups are based upon self-reported outcomes [15]. Mixing these unsupervised real-world data with samples recorded and annotated by experienced clinicians will likely increase the included variance significantly. Lastly, the missing standardization of sensors, their placements on the body, temporal resolutions, preprocessing methodologies applied to raw data, and similar technical factors represent a considerable challenge. Designing data-driven algorithms capable of accommodating these multifaceted challenges and establishing them in clinical reality represents a substantial hurdle [16].

Challenges regarding the amount or quality of labeled data are not limited to the clinical domain and require methods that can transfer knowledge from larger, sometimes healthier, populations and datasets. Transfer learning and, more recently, self-supervised learning [17] allow models to learn generic representations from vast pools of unlabeled data, which can then be re-used for specialized tasks with limited labeled data—an approach that has shown promise in the biomedical domain [17–19]. In the context of Parkinson's research with its limited amount of data, some related work has assessed the methodology for numerous sensory modalities suitable for the symptoms. Based on

videos, the technology has been used to obtain general gait patterns to assess the gait of patients with PD [18]. For the same objective, the technology has shown benefits for electroencephalography data [19], voice data [20], brain scans [21], or given extracted skeletons [22]. Focused on movement data, the technology proved its use for assessing disease prognosis [23], for assessing specific motor examinations [24], or for detecting anomalies of walking and freezing of gait [23,25].

Recently, Yuan et al. [26] demonstrated that self-supervised foundation models pretrained on accelerometer data could be fine-tuned to achieve significant improvement in activity recognition with a limited number of labels. Unlike similar work (i.e., [27]), they assessed the generalizability of the knowledge not only for healthy patients. While not the primarily focus of their work, they reported up to 135% performance gain when they fine-tuned the model instead of training it from scratch on a single dataset with accelerometer data of PD patients. However, the authors did not address generalizability across multiple PD datasets with varying recording conditions, sensor placements, and labeling procedures. Such cross-dataset robustness is critical for real-world clinical deployment, wherein models must cope with substantial between-dataset heterogeneity. Based upon their seminal findings, the aim was to empirically test two interconnected research hypotheses as follows:

1. Fine-tuned foundation models trained through self-supervised learning on accelerometer data of healthy participants enhance the recognition of activities associated with motor examinations conducted by PD patients across datasets and recording paradigms;
2. The fine-tuned models show increased robustness to varying recording conditions and different data preprocessing commonly observed between studies.

Through the corresponding findings, the study contributes insights into the potential usage of "general" representations of accelerometer data for use in PD research and increase robustness regarding the usage of similar systems in clinical reality.

2. Materials and Methods

2.1. Data

2.1.1. Datasets

To evaluate the potential of pre-existing knowledge regarding accelerometer data derived from healthy participants for the recognition of motor examinations in PD patients, three datasets were employed. Two publicly available datasets, namely the Clinician Input Study (CIS-PD) [28] and the Levodopa Response Trial [29], were utilized as instances of observational studies conducted mostly in an ambulant setting [28,30–32]. The third study includes data of motor disturbances from PD patients in the Parkinson's Clinical Movement Assessment (PACMAN) study, which was conducted by the Neurology Department at the University Medical Center Hamburg-Eppendorf [33].

Data from 24 participants of the CIS-PD study were included with their clinical assessments and recorded ambulatory long-term measurements with hospital visits (4 study sites in the United States), mainly at the beginning and the end of the trial. While not being at a study site, participants performed self-assessed ratings via a mobile phone application. Patients wore an Apple Watch Series 2 and only tri-axial accelerometer data were recorded and transferred to a mobile phone application, where a low-pass filter and summation of absolute acceleration in 5 and 30 s window intervals were performed [28].

The Levodopa Response Trial [29] was conducted on four consecutive days with selected items of the third part of the UPDRS on the first and the last day of the study. Participants repeated the tasks 6 to 8 times. The participants wore at least three devices: their smartphone, GeneActiv and Pebble. All of these devices were worn throughout the entire

study period, while recording daily activities during the second and third days at the participants' homes. The annotated accelerometer data of the GeneActiv device (ActivInsights Ltd., Kimbolton, United Kingdom [34]), worn on the most affected hand of the patients (upper limb), were included in the study. The data of these 27 participants were collected on two study sites in the United States. Further information regarding the devices and study setup could be found in the original publication [30].

As an example of data recorded in an entire clinical setup, sensor recordings of the PACMAN study originally collected for assessing methods of motor disturbances at the University Medical Center Hamburg-Eppendorf, Department of Neurology, Germany, were utilized [33]. In a configuration striving for similarities to clinical routine, a physician handed out an Apple Watch Series 6 to hospitalized patients with PD and conducted up to three assessments of seven selected items (3.3 rigidity, 3.4 finger tapping, 3.5 hand movements (tight fist), 3.6 rotation of hands, 3.9 arising from chair, 3.10 walking (gait), 3.17 rest tremor amplitude) of the UPDRS's third part per day. Those procedures were repeated while the therapy of the patients was adjusted for a maximum of two weeks. A total of 24 of these participants available during conduction of this study were included [33]. All procedures performed in the PACMAN study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The collection of measurements from the patients was approved by the Ethics Commission of the Ärztekammer Hamburg with the ID 2022-100846-BO-ff. Informed consent was obtained from all individual participants included in the PACMAN study.

The study encompassed 24, 27, and 24 participants of the CIS-PD, Levodopa Response Trial, and the PACMAN study, respectively. Table 1 provides an overview of the participants and available samples for each activity class, revealing a notable variation in the number of labels across datasets. Particularly, the ambulant observational studies (CIS-PD and Levodopa Response) amassed up to 17,000 samples for specific labels through self-assessments. In contrast, the clinic-based study at UKE, involving assessments by a physician, yielded considerably fewer samples.

Table 1. Patient characteristics and absolute number of annotated tasks available for every type of movement across the different cohorts. The first two studies contain far more measurements while the third dataset from the clinical domain contains far fewer variables. On the clinical set recorded at UKE, no tasks were selected while the patient was standing.

	Levodopa Response Trial	**Clinician Input Study**	**PACMAN Study**
Participants	27	24	24
Age range	50–84 years	36–75 years	49–79 years
Average age (SD)	67 (±9) years	63 (±10) years	65 (±8) years
Rotating hands (UPDRS 3.6)	4912	908	23
Other movements	15,172	3012	36
Sitting	2077	436	66
Standing	2077	418	0
Walking (UPDRS 3.10)	7010	950	135

UPDRS, Movement Disorder Society Unified Parkinson's Disease Rating Scale; PACMAN, Parkinson's Clinical Movement Assessment.

2.1.2. Data Preprocessing

The datasets represent a collection of the typical paradigms that could be found in research regarding motor symptoms of PD. To ensure interoperability at least regarding the format and labels of the accelerometer data, we utilized a unified data structure to

map all datasets to a common database structure [35]. None of the datasets were specifically designed for recognizing activities belonging to motor examinations. To obtain measurements for the different classes of activity, the raw signal was segmented into areas according to the given activity class labels. We split these areas containing only one kind of movement into segments of 10 s with a maximum overlap of 50%. This choice was guided by compatibility with the previous work and to reduce information loss at segment boundaries. No further preprocessing was conducted to maintain a representative sample of the measurements recorded in varying recording setups.

In collaboration with an experienced neurologist, we selected four different classes of activity that are useful for assessing motor symptoms of PD. Besides classical human activities like standing, sitting, and walking, we included the rotation of hands, as it is part 3.6 of the UPDRS. All the individual classes of labels between the different datasets were mapped to these "dataset-independent" classes. Whenever a dataset included a mixture of tasks, like moving from sitting to standing, we excluded those. The remaining activities that were of limited interest or unavailable within the other datasets were mapped to the class "other" and contained mostly a mixture of daily activities like drinking or writing. The set of those classes represented the foundation for the subsequent analysis.

2.2. Model

2.2.1. Foundation Model

As a recent example of a foundation model for accelerometer data trained through self-supervised learning, the publicly available model by Yuan et al., based on roughly 700,000 days of movement data from more than 100,000 healthy participants, was chosen for comparison [26]. This model employs a ResNet-V2 architecture with 18 layers and one-dimensional convolutions, which processes temporal windows of 10 s at a sampling rate of 30 samples per second. The model is designed to accept tri-axis accelerometer data as input, without the necessity for supplementary contextual information. To ensure direct comparability of results, our study replicated the experimental paradigm reported by the original authors. After the "embedding layer" with 1024 neurons, a dense layer with 512 neurons and the final output layer with five neurons, each corresponding to a target class, was added. The Softmax activation function was then applied to obtain class probabilities.

2.2.2. Fine-Tuning the Foundation Model

Depending on the three setups for evaluation, the foundation model underwent training or fine-tuning within a nested cross-validation framework, utilizing combined data from the CIS-PD and the Levodopa Response Trial. This process involved five outer test folds paired with eight inner validation folds, allocating 70% of the dataset for training, 10% for validation, and 20% for testing. The stratification of folds ensured that no participant's data was included in multiple folds, thus preserving the integrity of the evaluation. The training sessions explored three distinct learning rates (0.01, 0.001, and 0.0001), covering a range of reasonable defaults with a fixed batch size of 1024 determined by the used GPU resources, prototypical for clinical workstations. An early stopping criterion was employed to halt training if no improvement was observed in the validation set over 50 epochs. To reduce the complexity of hyperparameter analysis, advanced learning rate adaptation strategies were not implemented. The Adam optimizer facilitated the optimization process, utilizing cross-entropy loss to reduce the discrepancy between predicted logits and five target classes.

For evaluation, the multiclass F_1 score, defined as the average harmonic mean of precision and recall across all classes was used. This metric was chosen because it is compatible with the work by Yuan et al. [26]. This metric ranges from 0 to 1, with the latter indicating optimal classifier performance. The F_1 score on the validation set served as the basis

for early stopping and as the criterion for selecting the best model weights during training. During the testing phase, the weights from that point were used to calculate the F_1 score on the test set.

The training was conducted using a single consumer-grade NVIDIA A100 graphics processing unit, highlighting the model's potential for straightforward replication and application in clinical settings, even with limited computational resources.

2.3. Evaluation

The evaluation was conducted in two consecutive steps guided by the research hypotheses. For the first question, the training of an algorithm was simulated as it would be conducted in a scientific routine. Given the CIS-PD and the Levodopa Response Trial, the nested cross-validation was used both to obtain a suitable model for the task and an estimate of the performance on unseen data as a measure of generalizability through the incorporation of the test sets. For the sake of assessing the influence of learned representations, the training was run in three paradigms (Figure 1):

- Training from scratch: The first evaluation was based on training the network from scratch. In this setup, the pre-trained weights of the self-supervised model were not used at all. Instead, a random initialization of the network according to the utilized PyTorch library (version 1.13) took place. Accordingly, the deep-learning architecture is trained as it would be if other data besides the data in the training set were unavailable. This condition represents the baseline and could be used to study effects such as the appropriateness of the model structure. However, the risk of overfitting is high.
- Partial fine-tuning: The second training paradigm, partial fine-tuning, was based on training only the last layers of the network for predicting the presence of motor examination. The remaining layers with their pre-trained weights serve as feature extractors and were frozen. While the number of parameters requiring training is the fewest and the risk of overfitting is reduced accordingly, the other layers may not account for a changed distribution in the input data, given the non-healthy study population.
- Full fine-tuning: The third evaluation, the full fine-tuning, consisted of training the full network while using the existing weights of the foundation model as a starting point. While the model may fully adapt to the changed input data, previously extracted representations of movements could be reused. However, overfitting might affect the performance on unseen data, given the size of the network and the few training samples.

For all the paradigms, we assessed the potential positive impact of utilizing the learned representations for the chosen task of recognizing the motor examinations.

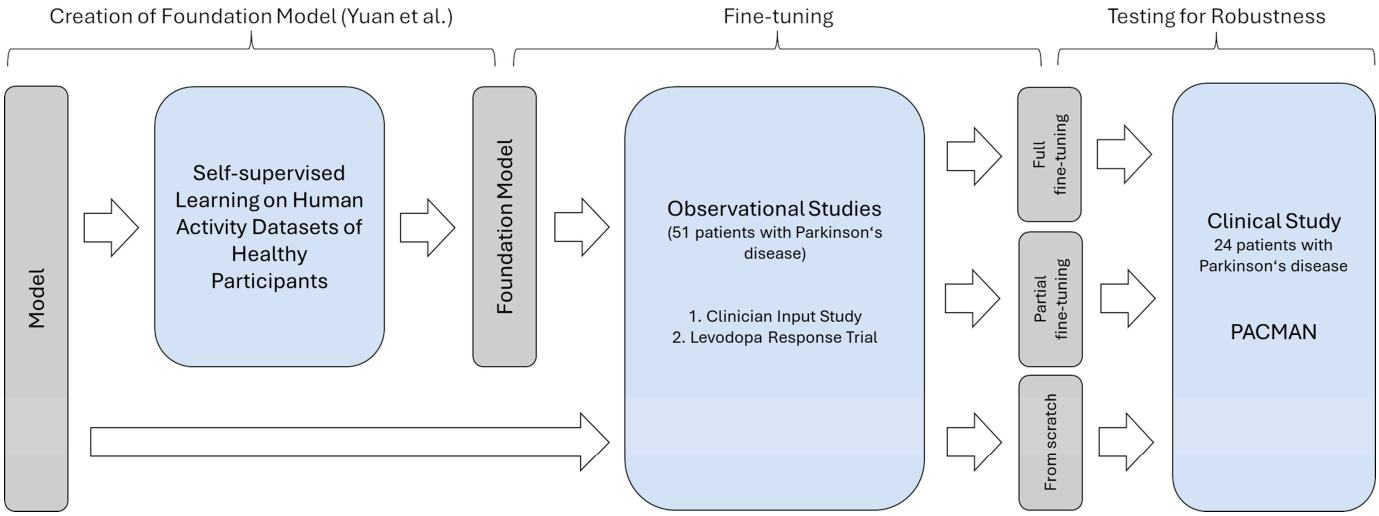


Figure 1. Overview of the evaluation procedure to test for the impact of different fine-tuning paradigms. PACMAN, Parkinson's Clinical Movement Assessment.

The second hypothesis was tested regarding increased robustness to varying recording conditions by testing the model against the clinical data (PACMAN). As one would expect given the heterogeneous datasets, the dataset did not contain measurements for the label "standing". Subsequently, the obtained performance was compared to those estimates derived through the test folds in the previous analysis. Instead of relying on additional metrics besides the F_1 score, the resulting confusion matrices were analyzed directly.

Besides this testing for generalizability between the recording environment, the impact of the recording setup, like the orientation of the sensor or scale of the data, was investigated. This was inspired by typical incompatibilities previously identified [35]. For that, the clinical PACMAN dataset was deliberately modified to generate two additional modified datasets. Through a simple linear transformation, the reported recordings according to the standardized SI unit and the common "g unit" (scaled dataset) were simulated. Additionally, the measurements from different cohorts were aligned, given the axis explaining most variance during the walking tasks, calculated through a principal component analysis [36] (rotated dataset) for measuring the impact of the orientation of the device.

2.4. Statistical Analysis

A classical *t*-test, with Welch's modification to account for potential changes in variances, was employed to test for significant performance differences between different models. Results are presented as mean ± standard deviation, and a threshold of statistical significance was set at $p < 0.001$ to ensure robustness of the findings.

3. Results

3.1. Impact of Training Procedures for Fine-Tuning

The trained models, when evaluated on the test folds during the nested cross-validation, exhibited average F_1 scores of 0.55 ± 0.06 when trained from scratch, 0.70 ± 0.02 when partially fine-tuned, and 0.69 ± 0.09 when fully fine-tuned, respectively (mean ± standard deviation). Across all experiments, models trained from scratch converged in an average of 144 ± 66 epochs, while partially and fully fine-tuned models required 128 ± 47 and 137 ± 62 epochs, respectively. Supporting the first hypothesis, training the model on pre-

trained weights, both partially and fully, statistically significantly outperformed training the model from scratch.

3.2. Influence of Learning Rate for Fine-Tuning

The impact of learning rates during evaluation differs significantly (Table 2). When comparing learning rates of 0.0001, 0.001, and 0.01, training from scratch resulted in average F_1 scores from 0.52 to 0.59. The peak performances on the test folds were obtained during full fine-tuning with scores above 0.58. In this case, the smallest learning rate was significantly better suited than the other cases. In partial fine-tuning, the effect of the learning rate was effectively mitigated and the performances were highly comparable (Figure 2). The frozen weights in the latter setup appear to let most of the models reliably converge to slightly sub-optimal values.

Table 2. F_1 scores for the three learning rates on the two observational studies combined across the different training paradigms.

Learning Rate	F_1 Score (±SD)		
	From scratch	Full fine-tuning	Partial fine-tuning
0.0001	0.52 ± 0.04	0.78 ± 0.02	0.70 ± 0.02
0.001	0.59 ± 0.04	0.70 ± 0.04	0.71 ± 0.02
0.01	0.54 ± 0.06	0.58 ± 0.07	0.70 ± 0.02

$F_1 = 1$ corresponds to the optimal performance.

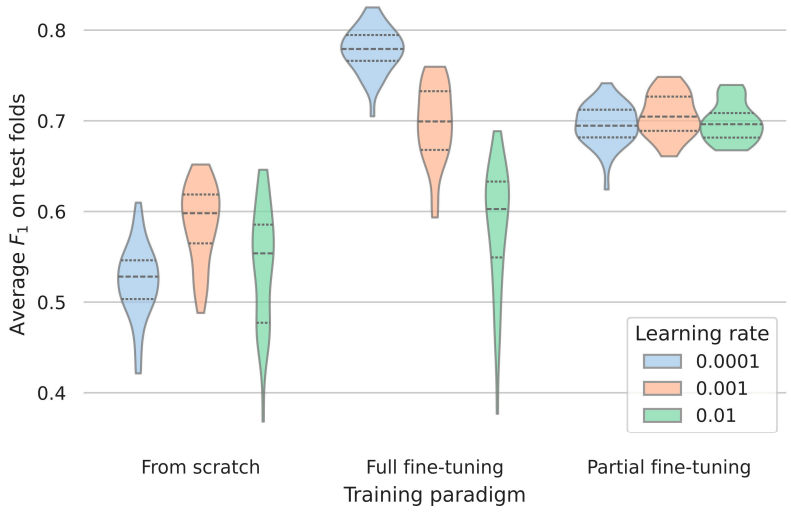


Figure 2. The impact of the learning rate on the two observational studies combined across the different training paradigms. When trained from scratch, the model failed to achieve a reasonable performance on the validation set. For the fine-tuning paradigms, the influence of the learning rate was higher (in full fine-tuning) or lower (in partial fine-tuning). $F_1 = 1$ corresponds to the optimal performance. The dotted lines indicate the quantiles of the data.

3.3. Assessing the Robustness and Generalizability of the Clinical Dataset

In the literature regarding motor symptoms, the obtained test score within (nested) cross-fold validation is commonly treated as an indicator of the proposed method's generalizability. To assess the effects of representations encoded within the foundation model for generalization in clinical setups more realistically, the models were applied to the PACMAN dataset. There was a considerable drop in absolute performance compared to the test folds' performances. The trained models achieved mean F_1 scores of 0.33 ± 0.06

when they were trained from scratch on the observational data, 0.43 ± 0.09 after being fully fine-tuned, and 0.48 ± 0.04 after partial fine-tuning (Figure 3).

In terms of relative performance difference, the fine-tuned models demonstrated considerably better generalization capabilities compared to the model trained de novo. Building upon representations learned from healthy participants resulted in statistically significantly better performance, commonly between 15 and 25%. Depending on the chosen fine-tuning paradigm, the learning rate had varying influence; a learning rate too large, when all weights could be modified, appeared to overwrite the previously learned representations completely, leading to similar performance as models only trained on the original data. When overfitting was not possible due to freezing most of the weights, the learning rate did not make a significant difference.

For a more comprehensive understanding than afforded by singular evaluation metrics, Figure 4 provides a visual depiction of the confusion matrices on the clinical dataset. For each training paradigm, the model with the highest score on the test folds of the observational dataset was chosen and applied to the data from the clinic.

While the sample size of the data recorded in a clinical context might be small compared to other similar datasets, some differences are observable. When considering the three distinct training paradigms, movements inherently present in the foundation model, such as walking, were consistently well recognized in the fine-tuned models. For movements not explicitly familiar to the model, such as hand rotation, the selected models exhibited no discernible advantages. While the dataset did not contain any samples of standing to mirror the differences in available labels across related datasets, this did not contribute to a large fraction of the wrongly classified activities.

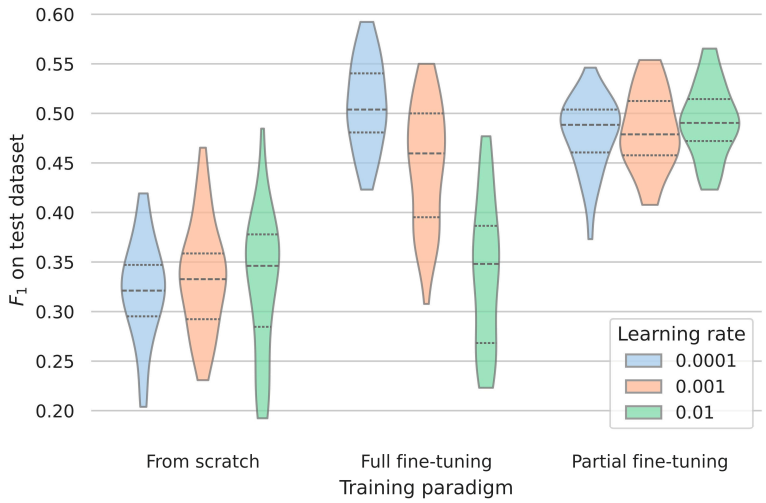


Figure 3. The performance of the models trained on cohort studies when applied to PACMAN as the test dataset for different learning rates and training paradigms. Despite the learning rate, the fine-tuned models show significantly better results than the models trained from scratch. $F_1 = 1$ corresponds to the optimal performance. The dotted lines indicate the quantiles of the data.

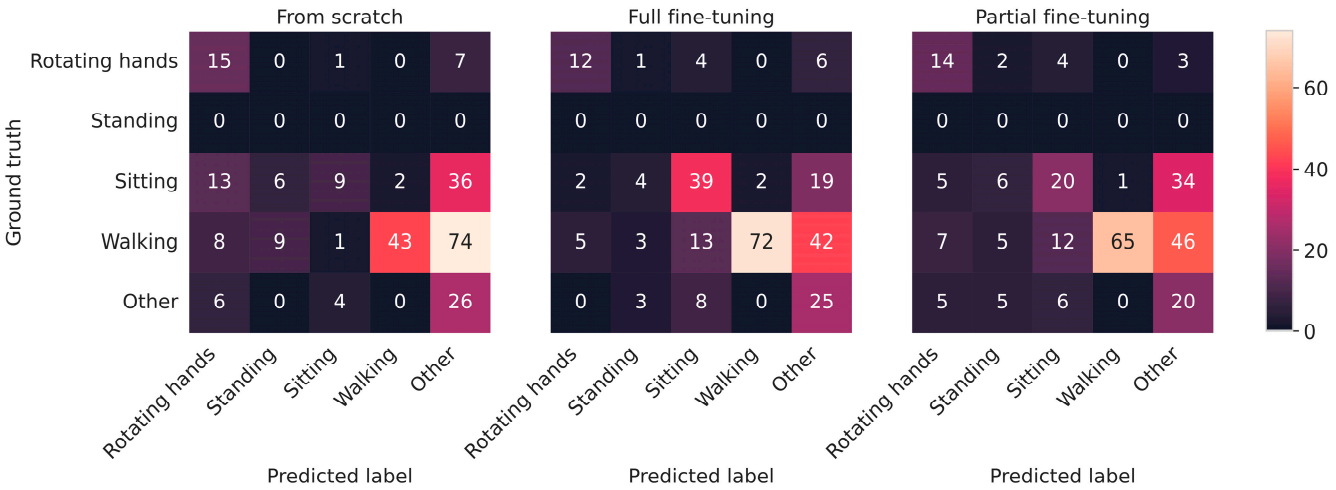


Figure 4. Confusion matrices of the best-performing models on the test folds from the three different training setups. The classes of the ground truth specify the correct class, while those below correspond to the prediction. Accordingly, an optimal classifier would only have values on the diagonal.

3.4. Evaluating the Robustness According to Recording Setups

To assess the models' robustness regarding technical changes in recording setups beyond the nature of the study, the test sets were modified as described in the methods section to simulate measurement data in different units and rotated sensors. The results are provided in Figure 5.

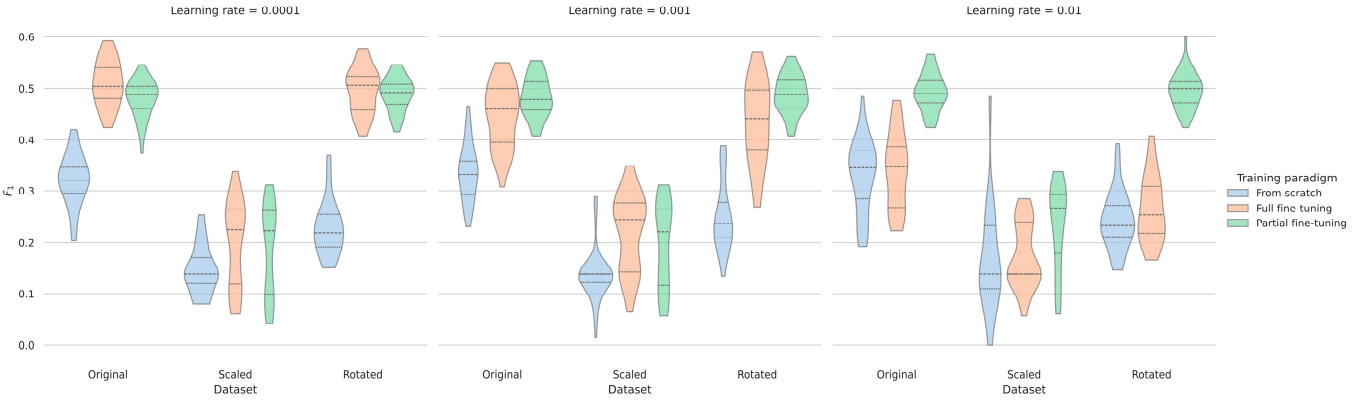


Figure 5. The performance of the models trained on the cohort studies when applied to the PAC-MAN as the test set for different learning rates and training paradigms. While all models show a considerable drop in accelerometer data of a different unit, the fine-tuned models show significantly better results when the sensor is artificially rotated. $F_1 = 1$ corresponds to the optimal performance. The dotted lines indicate the quantiles of the data.

All training paradigms exhibited significantly lower performance compared to the original data when confronted with the linearly scaled dataset, a simulation mirroring the discrepancy between the "g unit" and the SI unit for accelerometer data. However, the fine-tuned models showed similar (or sometimes even slightly better) performance when the data was rotated to simulate a sensor with a different orientation. The only exception to this was the fully fine-tuned model optimized with a high learning rate, which apparently overfitted the data and lost its advantages of the learned representations in comparison to a model trained from scratch.

4. Discussion

The limited amount of data and the disparate recording setups between studies focusing on motor symptoms of PD pose significant challenges for the sustainable development of the area of research. Enhancing the generalizability of proposed approaches is crucial for transitioning solutions from laboratories to clinical settings. The goal of this study was not to differentiate movements of PD patients from those of healthy controls, but rather to recognize specific motor examination activities in PD patients, benefitting from robust, generalizable movement representations learned on large healthy cohorts across different datasets.

4.1. Benefits of Utilizing Data from Healthy Participants for Movement Disorder Research

In line with the previous research by Yuan et al. [26], fine-tuning foundation models previously trained on accelerometer data originally collected from healthy participants significantly improved the classification of movements associated with tasks performed by PD patients. Despite the inherent differences in movement patterns between both populations, employing pre-trained weights within the deep learning framework resulted in performance improvements on test folds of up to 25% compared to training models solely on disease-specific datasets. However, this improvement over multiple training and validation datasets is significantly lower than the reported performance in the case of a single dataset. This finding highlights the danger of overfitting to the specific recording conditions and is evident in the choice of the hyperparameters, too. While the best performance was observed in the entirely fine-tuned model, the training process and its associated hyperparameters, like the learning rate, must be tightly observed to not "overwrite" the previous knowledge and overfit. Forcing the model only to adopt the parts responsible for classification efficiently reduced the danger of overfitting; however, it did not result in the best possible performance. The finding that the choice of learning rate critically affected performance, especially in full fine-tuning, likely reflects the interplay between step size and convergence landscape. Larger learning rates may have caused the optimization to miss global minima or become trapped in local minima, particularly given the limited data and early stopping strategy.

Given the computational efficiency of this approach, the additional complexity of the training setup does not pose a significant barrier for adoption in clinics. Even considering the resource-intensive nature of the meta-analyses in this study due to nested cross-validation, the process of fine-tuning a single network is manageable with a consumer-grade GPU, making it accessible and feasible for widespread use for related problems in hospitals.

4.2. Evidence for Increased Robustness Regarding Recording Setups

This study's findings suggest that the increased robustness not only holds during cross-validation but also when tested on a dataset recorded within the clinic. The observed degradation in performance was even larger than anticipated given the results on the test folds, prompting caution when interpreting results from nested cross-validation as accurate estimates of generalizability. Despite this, the generalized representations provided by fine-tuned models yielded significantly better results than models trained solely on disease-specific data. Specifically, the fine-tuned models demonstrated a 15–20% improvement over the scratch-trained models. The findings suggest that the performance "scales" between different setups. The relative differences between training paradigms and learning rates remained relatively consistent, indicating that optimization on the validation set could translate effectively to better results even across different datasets. This

property is highly advantageous, suggesting that effective cross-validation can guide better model tuning even across varying data conditions.

The model's robustness to artificially induced sensor rotations is encouraging, suggesting that representations learned by the foundation model may encode invariant features across axis permutations and device placements, a desirable property for real-world wearable deployments. By contrast, performance degradation during testing unit scaling (simulating changes from "g unit" to SI unit) likely indicates sensitivity to absolute input magnitude distribution. Accordingly, improved interoperability remains necessary. When models are trained from scratch, they do not exhibit this robust behavior, further providing evidence for fine-tuned models' capability to generalize under varied recording setups. However, it must be noted that orientation robustness was assessed using mathematically simulated data, which may not fully replicate the complexity of human-worn sensor placements in uncontrolled environments. Future studies should consider explicitly collecting datasets with known, systematically varied sensor orientations.

4.3. Utility of the Proposed Model for Recognizing Motor Examinations

To reduce the number of possible confounders for the analysis, a rather simple approach towards recognizing motor examinations was chosen. Despite the challenges of classifying specific movements such as hand rotation, which were not familiar to the model, the classification of movements like walking and sitting—prevalent in the data by the healthy participants used for training the model—was more successful despite the associated motor symptoms. This supports the use of large-scale pre-training for developing more reliable clinical monitoring tools. The absolute performance obtained through fully fine-tuning the model and testing it on the clinical data is certainly a valid starting point, but further improvements must be considered before applying such a model in the clinical context. Naturally, the obtained labels extracted from small windows should be aggregated given a meaningful temporal context, i.e., through the usage of attention mechanism. The selection of suitable targets remains important given the apparent challenges for complex movements resulting in the higher misclassification rates for the class "other", as reflected in the off-diagonal elements of Figure 4's confusion matrices. Overall, the utilization of data from healthy participants led to significantly better results and should likely be utilized in similar challenges.

4.4. Limitations and Future Work

Given the obtained insights within this study, we are certain that re-using the representation of accelerometer data will be more often used for tasks related to motor symptoms of PD. However, in the context of generalizability being the special focus of this work, some limitations should be considered in future work.

Firstly, the investigation focused on a single foundation model for accelerometer data chosen for its technical soundness, state-of-the-art performance, and computational efficiency. Future research could explore a variety of foundation models to further improve generalizability and performance. The large search space for optimizing hyperparameters, such as learning rates, batch sizes, and optimizers, poses a challenge. Optimization of the model architecture should be considered, too. While only two layers were trained in the partial fine-tuning setup to minimize overfitting, adding more layers may yield further improvements. Advanced transfer learning strategies, such as gradual unfreezing, adapter layers, or per-layer learning rates, should also be considered. Although this study addressed learning rates specifically due to their significant impact, future studies with greater computational resources might explore a broader range of hyperparameter configurations. Existing literature may already provide the first hints for such an optimized implementation [37].

Furthermore, this study concentrated on accelerometer data, a representative yet not exclusive modality for analyzing movement disorders. Incorporating additional sensor modalities, such as gyroscopes or magnetometers, in a multimodal framework, may further enhance performance. Additionally, considering temporal segments beyond fixed lengths could offer better real-world applicability, potentially combining multiple predictions into a unified score through additional post-processing steps.

5. Conclusions

In summary, the use of foundation models for accelerometer data holds the potential to significantly improve the performance on movement-related tasks in PD across recording setups. In the specific example of recognition of motor examinations under study, the additional knowledge embedded through self-supervised learning always led to significantly better results despite the presence of motor symptoms. The technology significantly improved the robustness of the approach despite discrepancies between observational and clinical datasets and recording variations. While it cannot entirely resolve all generalizability issues, it provides a meaningful step towards more robust, clinically applicable models. Continued research and development, focusing on broader datasets and diverse model architectures, will be essential for bridging the remaining gaps in research regarding motor symptoms and bringing these advancements into clinical routine.

Author Contributions: Conceptualization, C.G., A.J.W. and M.P.-N.; methodology, C.G., A.J.W. and M.P.-N.; validation, C.G.; formal analysis, C.G.; investigation, C.G.; resources, C.G. and A.J.W.; data curation, A.J.W.; writing—original draft preparation, C.G.; writing—review and editing, C.G., A.J.W. and M.P.-N.; visualization, M.P.-N.; supervision, C.G.; project administration, C.G. All authors have read and agreed to the published version of the manuscript.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article. The Apple Watches used for data acquisition were provided free of charge by Apple Inc. Apple was not involved in the design of the research, nor was it involved in the collection, analysis, or interpretation of the research data, or the content of this or any related publication.

Institutional Review Board Statement: All procedures performed in the PACMAN study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The collection of measurements from the patients was approved by the Ethics Commission of the Ärztekammer Hamburg with the ID 2022-100846-BO-ff.

Informed Consent Statement: Informed consent was obtained from all individual participants included in the PACMAN study.

Data Availability Statement: All the model weights trained during the analysis and the code required for replicating the study on own data are publicly available on GitHub (https://github.com/UKEIAM/de.uke.iam.parkinson.activity [accessed on 31 August 2025]) and the scientific data repository of the University of Hamburg (http://doi.org/10.25592/uhhfdm.13995 [accessed on 31 August 2025]).

Acknowledgments: The authors express their gratitude to the Michael J. Fox Foundation for Parkinson's Research for generously making the datasets related to the Levodopa Response Study and the Clinician Input Study available to the scientific community. We acknowledge financial support from the Open Access Publication Fund of UKE - Universitätsklinikum Hamburg-Eppendorf. During preparation of this manuscript, the authors used (generative) AI-powered tools like DeepL Translator, Grammarly, and ChatGPT 4o to improve the language and style of some parts of the paper. These tools were not used to generate any content of the paper itself.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CIS-PD	Clinician Input Study
PACMAN	Parkinson's Clinical Movement Assessment
PD	Parkinson's disease
MDS-UPDRS	Movement Disorder Society unified Parkinson's Disease Rating Scale

References

1. 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.; et al. Global, Regional, and National Burden of Parkinson's Disease, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **2018**, *17*, 939–953. https://doi.org/10.1016/S1474-4422(18)30295-3.
2. Postuma, R.B.; Berg, D.; Stern, M.; Poewe, W.; Olanow, C.W.; Oertel, W.; Obeso, J.; Marek, K.; Litvan, I.; Lang, A.E.; et al. MDS Clinical Diagnostic Criteria for Parkinson's Disease. *Mov. Disord.* **2015**, *30*, 1591–1601. https://doi.org/10.1002/mds.26424.
3. Monje, M.H.G.; Foffani, G.; Obeso, J.; Sánchez-Ferro, Á. New Sensor and Wearable Technologies to Aid in the Diagnosis and Treatment Monitoring of Parkinson's Disease. *Annu. Rev. Biomed. Eng.* **2019**, *21*, 111–143. https://doi.org/10.1146/annurev-bioeng-062117-121036.
4. Di Biase, L.; Tinkhauser, G.; Martin Moraud, E.; Caminiti, M.L.; Pecoraro, P.M.; Di Lazzaro, V. Adaptive, Personalized Closed-Loop Therapy for Parkinson's Disease: Biochemical, Neurophysiological, and Wearable Sensing Systems. *Expert Rev. Neurother.* **2021**, *21*, 1371–1388. https://doi.org/10.1080/14737175.2021.2000392.
5. Oliveira, A.M.; Coelho, L.; Carvalho, E.; Ferreira-Pinto, M.J.; Vaz, R.; Aguiar, P. Machine Learning for Adaptive Deep Brain Stimulation in Parkinson's Disease: Closing the Loop. *J. Neurol.* **2023**, *270*, 5313–5326. https://doi.org/10.1007/s00415-023-11873-1.
6. Giannakopoulou, K.-M.; Roussaki, I.; Demestichas, K. Internet of Things Technologies and Machine Learning Methods for Parkinson's Disease Diagnosis, Monitoring and Management: A Systematic Review. *Sensors* **2022**, *22*, 1799. https://doi.org/10.3390/s22051799.
7. Espay, A.J.; Bonato, P.; Nahab, F.; Maetzler, W.; Dean, J.M.; Klucken, J.; Eskofier, B.M.; Merola, A.; Horak, F.; Lang, A.E.; et al. Technology in Parkinson Disease: Challenges and Opportunities. *Mov. Disord.* **2016**, *31*, 1272–1282. https://doi.org/10.1002/mds.26642.
8. Sigcha, L.; Borzì, L.; Amato, F.; Rechichi, I.; Ramos-Romero, C.; Cárdenas, A.; Gascó, L.; Olmo, G. Deep Learning and Wearable Sensors for the Diagnosis and Monitoring of Parkinson's Disease: A Systematic Review. *Expert Syst. Appl.* **2023**, *229*, 120541. https://doi.org/10.1016/j.eswa.2023.120541.
9. Hill, E.J.; Mangleburg, C.G.; Alfradique-Dunham, I.; Ripperger, B.; Stillwell, A.; Saade, H.; Rao, S.; Fagbongbe, O.; Von Coelln, R.; Tarakad, A.; et al. Quantitative Mobility Measures Complement the MDS-UPDRS for Characterization of Parkinson's Disease Heterogeneity. *Park. Relat. Disord.* **2021**, *84*, 105–111. https://doi.org/10.1016/j.parkreldis.2021.02.006.
10. Safarpour, D.; Dale, M.L.; Shah, V.V.; Talman, L.; Carlson-Kuhta, P.; Horak, F.B.; Mancini, M. Surrogates for Rigidity and PIGD MDS-UPDRS Subscores Using Wearable Sensors. *Gait Posture* **2022**, *91*, 186–191. https://doi.org/10.1016/j.gaitpost.2021.10.029.
11. Silva De Lima, A.L.; Smits, T.; Darweesh, S.K.L.; Valenti, G.; Milosevic, M.; Pijl, M.; Baldus, H.; De Vries, N.M.; Meinders, M.J.; Bloem, B.R. Home-Based Monitoring of Falls Using Wearable Sensors in Parkinson's Disease. *Mov. Disord.* **2020**, *35*, 109–115. https://doi.org/10.1002/mds.27830.
12. Morgan, C.; Rolinski, M.; McNaney, R.; Jones, B.; Rochester, L.; Maetzler, W.; Craddock, I.; Whone, A.L. Systematic Review Looking at the Use of Technology to Measure Free-Living Symptom and Activity Outcomes in Parkinson's Disease in the Home or a Home-like Environment. *J. Park. Dis.* **2020**, *10*, 429–454. https://doi.org/10.3233/JPD-191781.
13. Yue, P.; Wang, X.; Yang, Y.; Qi, J.; Yang, P. Up-Sampling Active Learning: An Activity Recognition Method for Parkinson's Disease Patients. In *Proceedings of the Pervasive Computing Technologies for Healthcare*; Tsanas, A., Triantafyllidis, A., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 229–246.

14. Cheng, W.-Y.; Scotland, A.; Lipsmeier, F.; Kilchenmann, T.; Jin, L.; Schjodt-Eriksen, J.; Wolf, D.; Zhang-Schaerer, Y.-P.; Garcia, I.F.; Siebourg-Polster, J.; et al. Human Activity Recognition from Sensor-Based Large-Scale Continuous Monitoring of Parkinson's Disease Patients. In Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Philadelphia, PA, USA, 17–19 July 2017; pp. 249–250.
15. Denk, D.; Herman, T.; Zoetewei, D.; Ginis, P.; Brozgol, M.; Cornejo Thumm, P.; Decaluwe, E.; Ganz, N.; Palmerini, L.; Giladi, N.; et al. Daily-Living Freezing of Gait as Quantified Using Wearables in People With Parkinson Disease: Comparison with Self-Report and Provocation Tests. *Phys. Ther.* **2022**, *102*, pzac129. https://doi.org/10.1093/ptj/pzac129.
16. Shawen, N.; O'Brien, M.K.; Venkatesan, S.; Lonini, L.; Simuni, T.; Hamilton, J.L.; Ghaffari, R.; Rogers, J.A.; Jayaraman, A. Role of Data Measurement Characteristics in the Accurate Detection of Parkinson's Disease Symptoms Using Wearable Sensors. *J. Neuroeng. Rehabil.* **2020**, *17*, 52. https://doi.org/10.1186/s12984-020-00684-4.
17. Liang, Y.; Wen, H.; Nie, Y.; Jiang, Y.; Jin, M.; Song, D.; Pan, S.; Wen, Q. Foundation Models for Time Series Analysis: A Tutorial and Survey. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 6555–6565.
18. Sabo, A.; Mehdizadeh, S.; Iaboni, A.; Taati, B. Estimating Parkinsonism Severity in Natural Gait Videos of Older Adults with Dementia. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2288–2298. https://doi.org/10.1109/JBHI.2022.3144917.
19. Guo, Y.; Huang, D.; Zhang, W.; Wang, L.; Li, Y.; Olmo, G.; Wang, Q.; Meng, F.; Chan, P. High-Accuracy Wearable Detection of Freezing of Gait in Parkinson's Disease Based on Pseudo-Multimodal Features. *Comput. Biol. Med.* **2022**, *146*, 105629. https://doi.org/10.1016/j.compbiomed.2022.105629.
20. Rahman, W.; Lee, S.; Islam, M.S.; Antony, V.N.; Ratnu, H.; Ali, M.R.; Mamun, A.A.; Wagner, E.; Jensen-Roberts, S.; Waddell, E.; et al. Detecting Parkinson Disease Using a Web-Based Speech Task: Observational Study. *J. Med. Internet Res.* **2021**, *23*, e26305. https://doi.org/10.2196/26305.
21. Zhang, Y.; Lei, H.; Huang, Z.; Li, Z.; Liu, C.-M.; Lei, B. Parkinson's Disease Classification with Self-Supervised Learning and Attention Mechanism. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 4601–4607.
22. Endo, M.; Poston, K.L.; Sullivan, E.V.; Li, F.-F.; Pohl, K.M.; Adeli, E. GaitForeMer: Self-Supervised Pre-Training of Transformers via Human Motion Forecasting for Few-Shot Gait Impairment Severity Estimation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2022, Resorts World Sentosa, Singapore, 18–22 September 2022; Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S., Eds.; Springer Nature: Cham, Switzerland, 2022; pp. 130–139.
23. Jiang, H.; Bryan Lim, W.Y.; Shyuan Ng, J.; Wang, Y.; Chi, Y.; Miao, C. Towards Parkinson's Disease Prognosis Using Self-Supervised Learning and Anomaly Detection. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3960–3964.
24. Sánchez-Fernández, L.P.; Garza-Rodríguez, A.; Sánchez-Pérez, L.A.; Martínez-Hernández, J.M. A Computer Method for Pronation-Supination Assessment in Parkinson's Disease Based on Latent Space Representations of Biomechanical Indicators. *Bioengineering* **2023**, *10*, 588. https://doi.org/10.3390/bioengineering10050588.
25. Xia, Y.; Sun, H.; Zhang, B.; Xu, Y.; Ye, Q. Prediction of Freezing of Gait Based on Self-Supervised Pretraining via Contrastive Learning. *Biomed. Signal Process. Control.* **2024**, *89*, 105765. https://doi.org/10.1016/j.bspc.2023.105765.
26. Yuan, H.; Chan, S.; Creagh, A.P.; Tong, C.; Clifton, D.A.; Doherty, A. Self-Supervised Learning for Human Activity Recognition Using 700,000 Person-Days of Wearable Data. *NPJ Digit. Med.* **2023**, *7*, 91. https://doi.org/10.1038/s41746-024-01062-3.
27. Zhang, Y.; Ayush, K.; Qiao, S.; Heydari, A.A.; Narayanswamy, G.; Xu, M.A.; Metwally, A.A.; Xu, S.; Garrison, J.; Xu, X.; et al. SensorLM: Learning the Language of Wearable Sensors. *arXiv* **2025**, arXiv:2506.09108. https://doi.org/10.48550/arXiv.2506.09108.
28. Elm, J.J.; Daeschler, M.; Bataille, L.; Schneider, R.; Amara, A.; Espay, A.J.; Afek, M.; Admati, C.; Teklehaimanot, A.; Simuni, T. Feasibility and Utility of a Clinician Dashboard from Wearable and Mobile Application Parkinson's Disease Data. *NPJ Digit. Med.* **2019**, *2*, 95. https://doi.org/10.1038/s41746-019-0169-y.
29. Synapse.org. MJFF Levodopa Response Study. Available online: https://www.synapse.org/Synapse:syn20681023/wiki/594678 (accessed on 31 August 2025).
30. Daneault, J.-F.; Vergara-Diaz, G.; Parisi, F.; Admati, C.; Alfonso, C.; Bertoli, M.; Bonizzoni, E.; Carvalho, G.F.; Costante, G.; Fabara, E.E.; et al. Accelerometer Data Collected with a Minimum Set of Wearable Sensors from Subjects with Parkinson's Disease. *Sci. Data* **2021**, *8*, 48. https://doi.org/10.1038/s41597-021-00830-0.
31. Vergara-Diaz, G.; Daneault, J.-F.; Parisi, F.; Admati, C.; Alfonso, C.; Bertoli, M.; Bonizzoni, E.; Carvalho, G.F.; Costante, G.; Fabara, E.E.; et al. Limb and Trunk Accelerometer Data Collected with Wearable Sensors from Subjects with Parkinson's Disease. *Sci. Data* **2021**, *8*, 47. https://doi.org/10.1038/s41597-021-00831-z.

32. Sieberts, S.K.; Schaff, J.; Duda, M.; Pataki, B.Á.; Sun, M.; Snyder, P.; Daneault, J.-F.; Parisi, F.; Costante, G.; Rubin, U.; et al. Crowdsourcing Digital Health Measures to Predict Parkinson's Disease Severity: The Parkinson's Disease Digital Biomarker DREAM Challenge. *NPJ Digit. Med.* **2021**, *4*, 53. https://doi.org/10.1038/s41746-021-00414-7.
33. Wiederhold, A.J.; Zhu, Q.R.; Spiegel, S.; Dadkhah, A.; Pötter-Nerger, M.; Langebrake, C.; Ückert, F.; Gundler, C. Opportunities and Limitations of Wrist-Worn Devices for Dyskinesia Detection in Parkinson's Disease. *Sensors* **2025**, *25*, 4514. https://doi.org/10.3390/s25144514.
34. Activeinsights. GENEActiv with Software: Instructions for Use. Available online: https://activinsights.com/wp-content/uploads/2024/09/GENEActiv-1.2-IFU-rev-6.pdf (accessed on 26 August 2025).
35. Gundler, C.; Zhu, Q.R.; Trübe, L.; Dadkhah, A.; Gutowski, T.; Rosch, M.; Langebrake, C.; Nürnberg, S.; Baehr, M.; Ückert, F. A Unified Data Architecture for Assessing Motor Symptoms in Parkinson's Disease. *Stud. Health Technol. Inform.* **2023**, *307*, 22–30. https://doi.org/10.3233/SHTI230689.
36. Subramanian, R.; Sarkar, S. Evaluation of Algorithms for Orientation Invariant Inertial Gait Matching. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 304–318. https://doi.org/10.1109/TIFS.2018.2850032.
37. An, S.; Bhat, G.; Gumussoy, S.; Ogras, U. Transfer Learning for Human Activity Recognition Using Representational Analysis of Neural Networks. *ACM Trans. Comput. Healthc.* **2023**, *4*, 1–21. https://doi.org/10.1145/3563948.

7 Summary in English

This cumulative dissertation investigates how the opportunities presented by rapidly expanding health data can be effectively realized in clinical research and care, using Parkinson's disease as its main focus. While the volume and diversity of health data have grown exponentially, driven by digitization, novel data sources, and patient participation, these developments pose new challenges in terms of data quality, integration into existing systems, and the generation of actionable clinical knowledge.

Within the included publications, multiple approaches to interoperability are explored, including the prospective design of digital tools for standardized collection of patient-reported outcomes, retrospective structuring of clinical documents using vision-language models, and the harmonization of wearable sensor data. The implementation of a research platform further demonstrates how secure, standards-based secondary data use enables hypothesis-driven analysis on large clinical cohorts while ensuring privacy. Building on the interoperable data, the thesis explores analytic strategies to improve the robustness and generalizability of machine learning models. Both approaches, grounded in clinical knowledge and transfer learning techniques, are evaluated for their capacity to enable effective reuse of data and models across diverse datasets and tasks.

In summary, the dissertation provides concrete evidence that robust digital infrastructures, a systematic focus on interoperability, and the integration of clinical expertise with machine learning methods are crucial for developing reliable and adaptable models in healthcare. Although ongoing challenges remain, particularly in integrating multimodal data and advancing data standards, the solutions outlined here offer practical pathways toward scalable, reproducible, and clinically meaningful informatics within contemporary hospital settings.

8 Summary in German

Diese kumulative Dissertation untersucht, wie die wachsenden Menge an Gesundheitsdaten mit Bezug zu Parkinson effektiv genutzt werden kann. Während deren Volumen und die Vielfalt durch die Digitalisierung, neue Datenquellen und die Beteiligung von Betroffenen zunehmen, stellen diese Entwicklungen zugleich neue Herausforderungen bezüglich Datenqualität, Integration in bestehende Systeme und der Generierung von nutzbaren klinischem Wissen dar.

In den enthaltenen Publikationen werden technische Ansätze zur Interoperabilität untersucht, darunter die prospektive Entwicklung digitaler Werkzeuge zur standardisierten Erfassung patientenberichteter Endpunkte, die retrospektive Strukturierung klinischer Dokumente und die Harmonisierung von Sensordaten aus Wearables. Die Implementierung einer Forschungsplattform verdeutlicht darüber hinaus, wie eine sichere, standardbasierte Sekundärnutzung von Daten eine hypothesengesteuerte Analyse großer klinischer Kohorten unter Wahrung des Datenschutzes ermöglicht. Aufbauend auf den interoperablen Daten werden dann Methoden evaluiert, um die Robustheit und Generalisierbarkeit von Machine-Learning-Modellen zu verbessern. Sowohl klinisch motivierte Ansätze als auch Techniken des Transfer-Learnings werden hinsichtlich ihres Potenzials bewertet, eine effektive Wiederverwendung von Daten und Modellen über verschiedene Datensätze und Anwendungsfelder hinweg zu ermöglichen.

Zusammengefasst liefert die Dissertation konkrete Nachweise, wie durch robuste digitale Infrastrukturen, ein systematischer Fokus auf Interoperabilität und die Kombination klinischer Expertise mit Methoden des maschinellen Lernens zuverlässige Modelle im Gesundheitswesen entwickelt werden können. Obwohl weiterhin Herausforderungen bestehen, weisen die dargestellten Lösungen praxisnahe Wege zu skalierbaren, reproduzierbaren und klinisch relevanten Erkenntissen im modernen Krankenhausumfeld auf.

9 Author contributions

Throughout the various publications included in this cumulative thesis, I was primarily responsible for the planning, implementation, analysis, and management of the respective projects. For detailed information regarding my individual contributions to each publication, please refer to the author contribution statements provided within the corresponding articles.

To enhance the linguistic quality and clarity of this thesis, I utilized digital tools such as Grammarly and ChatGPT during its preparation, reviewed and edited the content as needed, and take full responsibility for the content of the published work. All intellectual content, analyses, and interpretations presented herein are my own work.

10 Acknowledgements

First, I would like to express my sincere gratitude to my supervisor, Prof. Dr. med. Frank Ückert. His unwavering trust and support during my time at the University Medical Center Hamburg-Eppendorf allowed me to grow independently as a researcher, develop resilience, and gain confidence in my abilities - much like learning to swim by being placed in the water. I would also like to thank Prof. Dr. med. Monika Pötter-Nerger for her insightful ideas and for facilitating access to the patients who played a vital role in this research. Her openness to collaboration and innovative ideas has greatly enriched this work, and I truly appreciate her support. My sincere thanks also go to Prof. Dr. med. Martin Scherer, who served as the third member of my thesis committee. I am grateful for the time, thoughtful feedback, and expertise he contributed, which have significantly strengthened this thesis.

The support of my colleagues at the University of Hamburg has been a cornerstone of my research experience. I am especially thankful to Alexander for his professionalism, skillful collaboration, and, above all, the friendship that made this academic journey both productive and enjoyable.

Finally, I wish to extend my deepest gratitude to Lisa, whose unwavering support, understanding, and encouragement have been a constant source of strength. I am also profoundly thankful to my parents for always giving me the freedom to explore independently, while providing steadfast support whenever I needed it. Their trust and encouragement have been invaluable to me.

11 Curriculum Vitae

Berufserfahrung

Universitätsklinikum Hamburg Eppendorf, Hamburg
Gruppenleitung Applied Artificial Intelligence in Healthcare **09/2022 – heute**

Wissenschaftlicher Mitarbeiter für Künstliche Intelligenz **01/2022 – 09/2022**

mindQ GmbH & Co. KG
Head of Artificial Intelligence **03/2021 – 12/2021**

Software Developer & Cognitive Computing Engineer (Werkstudent) **06/2018 – 03/2021**

Universität Osnabrück
Verantwortlicher Aussteller Hannovermesse 2019 **04/2019**

Studentische Hilfskraft Bio-inspired Computer Vision **03/2017 – 06/2017**

Ausbildung

Universität Osnabrück
Master of Science, Cognitive Science **09/2018 – 03/2021**

Bachelor of Science, Cognitive Science **08/2015 – 09/2018**

Auckland University of Technology
Computer Science & Artificial Intelligence **2017**

Ehrenamt

Studienstiftung des deutschen Volkes
Botschafter **04/2019 – 04/2021**

KZ-Gedenkstätte Neuengamme, Kirchliche Gedenkstättenarbeit
Ehrenamtlicher Guide **06/2021 – heute**

Sprachen

Deutsch	Muttersprache
Englisch	Verhandlungssicher

12 Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe, insbesondere ohne entgeltliche Hilfe von Vermittlungs- und Beratungsdiensten, verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe. Das gilt insbesondere auch für alle Informationen aus Internetquellen. Soweit beim Verfassen der Dissertation KI-basierte Tools („Chatbots“) verwendet wurden, versichere ich ausdrücklich, den daraus generierten Anteil deutlich kenntlich gemacht zu haben. Die „Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG“ aus September 2023 wurde dabei beachtet. Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe. Ich erkläre mich damit einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Datum und Unterschrift: _____