FAIR Voice Biomarker Data for Safe-Assured Embodied Health AI

Ishaan Mahapatra
Haslett High School
Haslett, MI, USA
imahapatra@icloud.com

Nihar R. Mahapatra[*]
Electrical and Computer Engineering
Michigan State University
East Lansing, MI, USA
nrm@egr.msu.edu

Abstract

Voice-based signals such as speech, breathing, and coughing form a primary perceptual channel for embodied health AI systems—from assistive robots and diagnostic companions to multimodal therapeutic agents. The safety and trustworthiness of such systems depend not only on model behavior but on the quality, traceability, and interoperability of the sensory data that drive their perception. Yet, current voice biomarker resources exhibit uneven data governance and limited adherence to the FAIR principles—Findability, Accessibility, Interoperability, and Reusability—creating hidden risks for safety assurance and regulatory compliance. In this work, we perform a comprehensive FAIRness evaluation of publicly available voice biomarker datasets spanning five major disease domains. Using a priority-weighted rubric aligned with the FAIR Data Maturity Model, we benchmark expert manual assessments against three automated tools (F-UJI, FAIR Evaluator, and FAIR-Checker) and quantify agreement and reliability through human–tool comparison and test–retest analysis. Results reveal strong Findability and Accessibility but persistent weaknesses in Interoperability and Reusability, notably in controlled vocabularies, qualified references, licensing, and provenance—elements essential for traceable and verifiable embodied perception pipelines. Automated tools often underestimated compliance due to rigid, domain-agnostic logic, underscoring the need for contextual evaluation. We introduce a data-level safety-assurance perspective, positioning FAIRness as a foundation for verifiable, policy-aligned embodied AI. By providing actionable recommendations—domain-specific metadata standards, machine-readable licensing, and FAIR-supportive repositories—we outline a practical pathway toward trustworthy, reproducible, and safe-assured embodied health AI systems.

1 Introduction

1.1 Generative AI and Voice in Healthcare

Generative AI (GenAI) is rapidly reshaping digital health, with foundation models enabling multimodal interaction through text, images, and increasingly speech and audio [42, 52]. Large-scale speech encoders and generative language models now support voice-based clinical assistants, conversational diagnostics, and synthetic data generation for underrepresented patient populations [42, 52, 15, 8]. Voice, breathing, and coughing signals are particularly promising because they are non-invasive, inexpensive to collect, and rich in clinical information [23, 27, 30]. Applications of voice biomarkers have been demonstrated across psychiatric disorders, neurodegenerative

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Workshop on Embodied and Safe-Assured Robotic Systems.

diseases, respiratory illness, and cardiometabolic conditions [23, 27, 18, 30]. Clinical perspectives emphasize the potential of *audiomics*—voice-based phenotyping through systematic extraction of acoustic and linguistic features—as a scalable approach to health monitoring [15]. Within embodied health AI systems—such as assistive robots, diagnostic companions, and conversational therapeutic agents—voice serves as a primary perceptual channel enabling human–machine interaction. The reliability of these embodied agents therefore depends not only on model behavior but also on the quality, traceability, and interoperability of the underlying voice data that train and validate their perception. As GenAI systems begin to integrate voice both as input (symptom elicitation, screening) and output (therapeutic dialogue, conversational monitoring), the robustness and reliability of underlying voice datasets become decisive for clinical validity, regulatory acceptance, and data-driven safety assurance in embodied contexts.

1.2 Challenges in Voice Biomarker Datasets

Despite their potential, publicly available voice biomarker datasets often lack consistent metadata, standardized acquisition protocols, and clear licensing terms [21, 23, 53]. These shortcomings undermine reproducibility and limit lawful reuse, thereby constraining the development of trustworthy and safe GenAI models. In particular, poor alignment with the FAIR principles—Findability, Accessibility, Interoperability, and Reusability—slows the translation of promising voice biomarkers into clinical and embodied settings [11]. Incomplete or inconsistent adherence to FAIR principles hinders scaling of foundation models that rely on large, diverse datasets, and creates risks for bias, hallucination, or non-compliance in downstream GenAI systems—factors that can manifest as unpredictable or unsafe behavior in embodied applications.

1.3 FAIR Assessment of Voice Biomarker Datasets

The FAIR principles [64, 29] are widely endorsed as a framework for ensuring dataset transparency, interoperability, and lawful reuse. FAIR assessments have been applied in several biomedical domains, including genomics, proteomics, and institutional repositories [11, 58], where automated tools such as F-UJI [24], FAIR Evaluator [65], and FAIR-Checker [28] are increasingly used to provide scalable, automated scoring. Within voice biomarkers, however, systematic FAIR assessments remain scarce: the only structured attempt to date considered a limited subset of datasets from two disease domains [41]. To our knowledge, no study has evaluated FAIRness comprehensively across all major disease categories where voice biomarkers are applied, nor has any work benchmarked automated FAIRness tools against expert human judgment in this context. By positioning FAIRness as a data-level safety verification framework, our analysis connects dataset governance with the broader goals of reliability, validation, and assurance that underpin safe and predictable embodied AI behavior.

1.4 Motivation and Scope

To our knowledge, the only prior structured evaluation of voice dataset FAIRness considered a small subset of mental health and neurodegenerative datasets [41]. In contrast, our study provides the first comprehensive assessment spanning five major disease categories. Specifically, we (1) apply a refined scoring rubric aligned with the FAIR Data Maturity Model, (2) compare expert manual assessments with outputs from three widely used automated FAIR tools, and (3) introduce human–tool agreement and test–retest reliability analyses to quantify consistency. These contributions establish a domain-wide evidence base and methodological framework for benchmarking dataset quality, tool consistency, and evaluation robustness as a foundation for safe and trustworthy embodied GenAI systems in health.

1.5 Key Contributions

This paper makes four primary contributions:

1. We conduct the most comprehensive FAIRness evaluation to date of voice biomarker datasets, spanning five clinically relevant disease domains.
2. We perform the first systematic comparison of manual FAIR assessments with three automated evaluation tools in this domain, highlighting discrepancies and tool limitations.

3. We incorporate human–tool agreement and test–retest reliability analyses, establishing methodological transparency and robustness for FAIR assessments.
4. We translate our findings into actionable recommendations for safety-assured GenAI and embodied health systems, emphasizing metadata standards, repository selection, and compliance-ready practices to enable reliable, reproducible, and lawful use of voice data.

Taken together, these contributions address the critical issues of *trust, safety, and policy compliance* in embodied and generative health AI. By analyzing FAIRness across disease domains, we expose structural barriers that currently limit the safe use of voice data in GenAI models. By benchmarking automated FAIR tools against expert judgment, we provide evidence on the strengths and limitations of scalable compliance mechanisms. Finally, by introducing reliability metrics, we enhance transparency and reproducibility, contributing to both trustworthy evaluation and regulatory readiness. Our findings thus inform how GenAI systems in healthcare can responsibly integrate voice data to realize their clinical potential. Our findings position FAIRness as a practical data-assurance mechanism for embodied AI systems, linking dataset quality directly to safety, predictability, and assurance in real-world health applications.

2 Related Work

2.1 FAIR Principles and Biomedical Data Sharing

The FAIR guiding principles—Findable, Accessible, Interoperable, and Reusable—provide a widely adopted framework for making scientific datasets discoverable, traceable, and reusable across institutions and tools [64, 29]. To operationalize FAIR, the FAIR Data Maturity Model (FDMM) specifies subprinciples and indicators that enable structured assessment and comparison across resources [11]. In biomedicine, FAIR has become closely tied to *AI-readiness*: initiatives emphasize persistent identifiers, machine-actionable metadata, provenance, and licensing as prerequisites for trustworthy AI development and evaluation [21]. Regulatory and translational contexts reinforce these expectations—for example, the European Health Data Space (EHDS) Regulation codifies interoperability, access conditions, and governance with explicit references to FAIR, while clinical device oversight highlights documentation and transparency for AI-enabled tools [7, 14]. These principles also underpin emerging safety and assurance frameworks for embodied AI systems, where dataset traceability and provenance directly affect the verifiability of model behavior and post-deployment accountability.

2.2 Automated FAIR Assessment Tools

Multiple automated validators implement parts of the FDMM to provide scalable checks: F-UJI assesses identifier persistence, license detectability, and selected interoperability indicators [24]; FAIR Evaluator offers configurable tests aligned to community metrics [65]; and FAIR-Checker leverages semantic web standards to probe findability and reuse [28]. Empirical examinations show these tools can under- or mis-estimate compliance when evidence is encoded in valid but non-canonical forms, or when domain context is needed for interpretation, underscoring the continuing role of expert review [58]. Collectively, automated validators are useful for screening and regression testing, but they cover only a subset of indicators and require domain-aware complements for comprehensive evaluation. In the context of safety assurance, these automated evaluations can be viewed as early-stage verification tools that assess data integrity—the first link in a broader assurance pipeline extending from dataset governance to embodied system reliability.

2.3 Voice Biomarkers: Promise and Dataset Challenges

Voice biomarkers have been investigated across psychiatric, neurodegenerative, respiratory, and cardiometabolic conditions, with reviews and systematic analyses highlighting clinical promise and methodological gaps [23, 27, 30, 53]. Public datasets have proliferated, yet documentation quality, licensing clarity, interoperability (controlled vocabularies, qualified links), and provenance metadata remain uneven—limitations that directly impact reproducibility and lawful reuse in clinical AI. These gaps translate into practical safety risks for embodied health agents whose auditory perception depends on such datasets: missing or inconsistent metadata can propagate uncertainty through sensing, reasoning, and actuation layers.

2.4 FAIR Assessments for Voice Datasets: Evidence to Date

Systematic FAIR assessments tailored to voice biomarker datasets remain scarce. To our knowledge, the only structured attempt evaluated a limited subset of datasets focused on mental health and neurodegenerative disorders, without benchmarking automated tools or quantifying the reliability of manual scoring [41]. No prior work has assessed FAIRness comprehensively across all major disease categories where voice biomarkers are applied, and none has compared expert judgments against multiple automated validators in this domain. The present study therefore connects FAIR data evaluation with embodied AI assurance, positioning dataset-level FAIRness as a measurable foundation for safe and trustworthy embodied perception. It provides a domain-wide analysis and a human–tool comparison with reliability quantification, offering an evidence base and methodology directly aimed at GenAI development and governance in health.

3 Methodology

3.1 Study Design and Scope

Our study was designed to provide a comprehensive, reproducible, and domain-specific evaluation of FAIR alignment in publicly accessible voice biomarker datasets. We focused on five major disease categories that collectively represent the majority of voice biomarker applications: *mental health, neurodegenerative, respiratory, cardiometabolic/cardiovascular, and autoimmune*. These categories were selected to ensure broad coverage across physiological and psychological domains where voice is increasingly investigated as a diagnostic or monitoring tool. Because voice is a core perceptual signal for embodied health AI systems, these datasets were treated as representative of the sensory foundations that inform perception, reasoning, and action in embodied agents.

3.2 Dataset Identification and Selection

Candidate datasets were systematically identified through comprehensive searches on Google Dataset Search conducted between January and May 2025, complemented by targeted queries of major repositories including PhysioNet, Zenodo, Hugging Face, and institutional data archives. Additional datasets were located through targeted follow-up queries using ChatGPT, which helped identify resources potentially missed in initial searches.

For each clinical domain, we formulated targeted queries combining disease-specific terms with voice or acoustic descriptors. Representative examples include: (a) **Autoimmune diseases:** "multiple sclerosis voice dataset," "rheumatoid arthritis speech dataset"; (b) **Cardiometabolic and cardiovascular diseases:** "hypertension speech dataset," "diabetes voice dataset"; (c) **Neurodegenerative diseases:** "Parkinson's disease speech dataset," "Alzheimer's voice dataset"; (d) **Mental health:** "depression voice dataset," "bipolar disorder speech dataset," "anxiety voice dataset"; (e) **Respiratory diseases:** "COVID-19 cough dataset," "pneumonia voice dataset," "tuberculosis cough dataset."

Inclusion criteria required that datasets:

1. Contain human voice or acoustic signals directly linked to health conditions.
2. Be publicly available without requiring restricted access or special permissions.
3. Provide sufficient metadata or documentation to enable FAIR evaluation.
4. Clearly pertain to at least one of the five target disease categories.

Datasets were excluded if they:

- Provided only summary statistics without underlying audio data.
- Consisted entirely of synthetic or non-human audio.
- Lacked adequate clinical annotation or clear linkage to a disease condition.

Each dataset meeting inclusion criteria was assigned an alphanumeric code by disease domain (A, C, M, N, R) for consistent reference in results and figures. While not all datasets were fully multimodal, most paired acoustic signals with structured clinical labels or demographic metadata, forming bimodal inputs analogous to those used in embodied robotic perception pipelines.

3.2.1 Final Dataset Selection

Our initial selection included 51 datasets focused on voice biomarkers of disease: **Autoimmune disease datasets (A1–A3)** included A1 [10], A2 [45], and A3 [35]. **Cardiometabolic and cardiovascular datasets (C1–C7)** included C1 [38], C2 [6], C3 [56], C4 [2], C5 [4], C6 [38], and C7 [3]. **Mental Health datasets (M1–M10)** included M1 [12], M2 [59], M3 [20], M4 [5], M5 [9], M6 [19], M7 [34], M8 [25], M9 [66], and M10 [37]. **Neurodegenerative datasets (N1–N17)** included N1 [30], N2 [33], N3 [63], N4 [62], N5 [40], N6 [31], N7 [12], N8 [13], N9 [26], N10 [36], N11 [50], N12 [39], N13 [16], N14 [46], N15 [61], N16 [55], and N17 [43]. **Respiratory disease datasets (R1–R14)** included R1 [57], R2 [17], R3 [17], R4 [1], R5 [48], R6 [47], R7 [44], R8 [49], R9 [54], R10 [32], R11 [51], R12 [60], R13 [67], R14 [22]

3.3 FAIR Data Maturity Model and Rubric Development

We adopted the FAIR Data Maturity Model [11], which operationalizes the FAIR principles into 15 subprinciples and 41 indicators. To adapt the framework to voice biomarker datasets, we applied the rubric first introduced in prior voice-specific FAIR assessment [41] and extended it by:

- Incorporating domain-specific clarifications for indicators requiring context (e.g., acceptable controlled vocabularies, audio format standards).
- Using a priority-weighted system (Essential, Important, Useful) to reflect the relative importance of indicators for dataset reuse in GenAI pipelines and for embodied safety verification.

Each indicator was assessed as a binary variable—1 if the indicator was satisfied and 0 otherwise. For each subprinciple, the binary indicator scores were mapped to a subprinciple score of 0 (no indicators satisfied), 0.5 (some but not all indicators satisfied), or 1 (all indicators satisfied). Priority weights (Essential, Important, Useful) were then used to aggregate these subprinciple scores into principle-level and composite FAIR scores. This structured rubric serves as a quantitative proxy for data-level assurance, where higher FAIR alignment corresponds to stronger traceability and lower risk in downstream embodied AI applications.

3.4 Automated FAIR Assessment Tools

To benchmark manual evaluations, we applied three automated FAIR assessment tools: (a) F-UJI [24], which tests for identifiers, licensing, and selected interoperability criteria; (b) FAIR Evaluator [65], which offers configurable tests against community metrics; and (c) FAIR-Checker [28], which integrates semantic web standards to evaluate findability and reusability.

Each tool was run using the publicly accessible dataset landing page or metadata URL, and results were normalized into the same indicator structure used in manual scoring. This allowed direct comparison between automated and expert evaluations at indicator, subprinciple, and principle levels. We interpret these automated evaluations as analogous to automated verification modules within safety assurance pipelines, useful for scalable compliance screening of data assets that feed embodied AI systems.

3.5 Test–Retest Reliability Protocol

To establish intra-rater reliability, we reassessed a representative subset of 12 datasets, balanced across all five disease domains, four weeks after the initial scoring. Subprinciple- and principle-level scores were collected at both time points, enabling test–retest reliability analysis using RMSEs, as detailed in the Results (Section 4). Stable retest performance demonstrates methodological robustness, a property parallel to validation consistency required in safety-critical robotic evaluation.

3.6 Inter-Annotator Agreement: Human versus Tool

We framed automated FAIR assessment tools as independent annotators, allowing direct comparison of automated and manual annotations. Automated tool outputs were mapped onto the same scoring structure as manual scores, facilitating human–tool agreement analysis using RMSE. This provided quantitative insights into the alignment between automated assessments and expert human judgment, highlighting where automated safety-verification analogs align—or diverge—from expert assurance review.

3.7 Transparency and Reproducibility

All scoring decisions were documented in a structured rubric to minimize ambiguity and ensure reproducibility. Detailed justifications for indicator-level scores were logged, and tool outputs were archived to allow independent verification. By adopting transparent, version-controlled scoring and openly sharing evaluation artifacts, we align FAIR practice with core principles of explainable and auditable safety assurance for embodied systems.

4 Results

4.1 Overall FAIRness Patterns

Figure 1 summarizes manual FAIR scores at the subprinciple, principle, and composite levels across all datasets. Findability (F) is consistently high, with persistent identifiers and basic metadata broadly satisfied. Accessibility (A) is also relatively strong, but systematic weaknesses emerge in A1.2 (metadata accessibility independent of data) and, to a lesser extent, A2 (long-term metadata persistence). Interoperability (I) and Reusability (R) remain the most limiting principles, with pronounced deficiencies in I2 (controlled vocabularies), I3 (qualified references), and R1.x (provenance, licensing, and standards). These appear as vertical red/yellow bands in the heatmap, cutting across disease domains. The overall pattern indicates that while most datasets are findable and retrievable, many remain difficult to integrate, machine-read, or lawfully reuse. From an embodied AI perspective, such deficiencies translate into reduced traceability and uncertain sensory integration, which undermine assurance and post-deployment verifiability in robotic or interactive health systems. Addressing missing vocabularies, qualified references, and explicit licensing/provenance would directly improve dataset quality and readiness for safety-critical embodied GenAI pipelines.

4.2 FAIRness by Disease Category

Figure 2 shows mean FAIR scores across disease domains. As with overall patterns, Findability (F) and Accessibility (A) are consistently stronger than Interoperability (I) and Reusability (R). Cardio datasets perform weakest on I and R, resulting in the lowest composite FAIRness. Autoimmune datasets achieve mid-range composite scores, while mental health, neurodegenerative, and respiratory datasets cluster at the top with the highest overall alignment. These disparities point to uneven maturity across research communities: cardiometabolic voice datasets in particular would benefit most from targeted improvements in interoperability (controlled vocabularies, qualified references) and reusability (provenance and licensing). In embodied or sensor-driven systems, such domain-level variability implies unequal reliability of auditory perception models across clinical contexts, which poses safety and trust challenges when models are deployed in heterogeneous user populations.

4.3 FAIRness by Repository/Hosting Venue

Figure 3 groups composite FAIRness by repository (dataset counts in parentheses). The effect of venue is clear: curated or institutionally managed platforms—Synapse and PhysioNet (both $\approx$100), IEEE Data Port and Mendeley Data ($\approx$95), and Zenodo ($\approx$90)—sit at the top, reflecting stronger enforcement of metadata fields, persistent identifiers, versioning, and license declarations. Publication/code portals are more mixed: Papers with Code performs relatively well in this sample ($\approx$90), while ScienceDirect and PLOS One are mid–high ($\approx$85–88) and GitHub is lower ($\sim$79) with broad variability (large error bar). Venues with minimal repository-level requirements (Kaggle $\approx$68, OSF $\approx$55, Hugging Face $\approx$50) occupy the bottom of the chart. Error bars (e.g., GitHub, Zenodo, UCI ML Repository, Mendeley Data) indicate within-venue spread when multiple datasets are present. Overall, repository choice directly influences FAIRness; steering deposits toward FAIR-supportive venues—or mirroring datasets there—can yield immediate, practical gains in dataset quality. In practical terms, repositories with enforced metadata policies act as "safety anchors," ensuring consistent provenance and licensing—both prerequisites for compliance and traceability in embodied health systems.

4.4 FAIRness Over Time

Figure 4 plots FAIRness scores against dataset release year with a best-fit regression line. No strong temporal trend is evident: the slope is slightly negative and the explained variance is negligible ($R^2 = 0.0128$). Newer datasets are not consistently more FAIR, indicating that improvements over time have been uneven. This finding challenges the assumption that FAIR alignment will naturally increase as data sharing practices evolve. Instead, effective gains will likely require structural interventions such as repository-level mandates, standardized metadata templates, and submission-time compliance checks. The wide dispersion of scores within each year—ranging from low to very high—further underscores that age alone does not determine FAIRness. The lack of temporal improvement reinforces the need for institutionalized data-governance frameworks analogous to ongoing safety certification in embodied AI—continuous assurance rather than one-time compliance. The main implication is that FAIRness in voice biomarker datasets is not improving organically; deliberate measures for dataset creation and curation are therefore necessary, motivating the recommendations introduced later in this paper.

4.5 Manual–Automated Agreement, Reliability, and Illustrative Example

Figure 5 reports pairwise RMSE of FAIR subprinciple scores between manual and automated assessments (F-UJI, FAIR Evaluator, FAIR-Checker), as well as among the automated tools. Agreement is uneven across principles. For **manual–tool comparisons**, Findability (F) RMSEs are relatively high for F-UJI and FAIR Evaluator ($\sim$0.67–0.68) but lower for FAIR-Checker (0.45). Accessibility (A) shows the widest spread, with low disagreement for FAIR Evaluator and FAIR-Checker (0.27–0.30) but much higher for F-UJI (0.66). Interoperability (I) and Reusability (R) are consistently challenging: RMSEs span 0.31–0.57 for I and 0.31–0.61 for R, with the largest mismatches occurring for manual–FAIR Evaluator on I and manual–F-UJI on R. Composite RMSEs range from 0.37–0.59, reflecting the accumulation of partial divergences across principles. For **tool–tool comparisons**, alignment is somewhat tighter but still principle-dependent. F-UJI and FAIR Evaluator agree closely on F (0.14) but diverge strongly on R (0.72). F-UJI and FAIR-Checker show moderate disagreement across all principles (0.31–0.49), while FAIR Evaluator and FAIR-Checker are closer on F and A (0.22–0.30) but diverge on I (0.40) and R (0.53). Overall, tools tend to converge on easier-to-detect elements (F, some aspects of A) while diverging most on I and R, where metadata standards, provenance, and licensing are less consistently machine-readable. This divergence between automated and expert judgment mirrors gaps seen in safety assurance pipelines, where domain-agnostic verification tools may overlook contextual evidence that human evaluators recognize as critical for embodied system integrity.

To confirm that these discrepancies do not reflect instability in manual scoring, we conducted a test–retest reliability check. A representative subset of 12 datasets spanning all five domains was rescored four weeks after the initial evaluation. Subprinciple- and principle-level scores were identical across test and retest, yielding perfect agreement (RMSE = 0). This establishes that the manual rubric is stable and reproducible, and that the observed disagreements in Figure 5 are due to tool limitations rather than evaluator inconsistency. Reliability at this level ensures that the FAIR evaluation procedure itself meets assurance expectations comparable to repeatability requirements in safety certification workflows.

An illustrative case is the Bridge2AI-Voice v1.1 dataset on PhysioNet, which achieved perfect manual scores across all subprinciples owing to persistent identifiers, detailed metadata, explicit licensing, provenance, and adherence to community standards. Automated validators, however, underestimated its FAIRness. FAIR-Checker flagged insufficient discoverability despite embedded DOIs, version history, and repository-level search. F-UJI failed to recognize a repository-specific license URL and did not credit human-readable vocabularies. FAIR Evaluator marked down identifier persistence and searchability even though DOIs and repository indexing were present. These discrepancies arose from narrow pattern-matching and schema-specific expectations: elements expressed in valid but non-standard ways were overlooked. This example underscores that automated tools, while useful for scalable screening, are incomplete proxies for FAIR compliance. Domain-aware manual evaluation remains essential for ensuring dataset readiness and trustworthy integration into safety-critical embodied AI pipelines.

Practical Recommendations for GenAI-Ready Voice Datasets. Our results point to several high-yield interventions for dataset creators and curators: (i) **Machine-readable licensing and provenance**—declare SPDX-encoded licenses and structured provenance (e.g., DataCite `relatedIdentifiers`) to address R1.1–R1.3; (ii) **Controlled vocabularies and qualified references**—use community ontologies and `schema.org`/JSON-LD links to related resources to raise I2/I3; (iii) **Repository choice**—deposit or mirror in FAIR-supportive venues (e.g., PhysioNet, Zenodo, Synapse) with PIDs, versioning, and enforced metadata, which lifts F/A and composite scores; (iv) **Discoverability**—ensure keywords, topical tags, version history, and repository indexing so both automated and human search succeed on F2; (v) **Validator-aware metadata**—where feasible, provide license URIs and vocabulary encodings that current validators recognize, while retaining human-readable documentation for domain context. These actions directly target the systematic weaknesses observed in I2/I3 and R1.x and improve readiness for safety-assured embodied GenAI pipelines.

Limitations. This study evaluates publicly accessible datasets, so findings may not generalize to proprietary clinical repositories. Automated validators emphasize machine-readable, domain-agnostic indicators and cannot assess many voice-specific quality signals (e.g., device/channel metadata, acquisition environment, acoustic feature vocabularies). Even within their scope, tool logic sometimes fails to detect valid evidence—for example, repository-specific license URIs, embedded DOIs, or provenance recorded in human-readable sections. These limitations explain why automated outputs often underestimated FAIRness relative to domain-aware manual scoring. They do not alter the core finding: voice biomarker datasets are generally findable and accessible but remain limited in interoperability and reusability, and automated tools alone are insufficient to capture their FAIRness comprehensively. Future work should integrate FAIR evaluation with embodied AI testing frameworks to enable unified data-to-deployment assurance pipelines for trustworthy and safe health robotics.

5 Conclusion

This study provides the first comprehensive FAIRness evaluation of voice biomarker datasets spanning all major disease categories, benchmarked through both expert assessment and automated tools, with added reliability analysis to ensure robustness. Our findings reveal that while datasets are generally findable and accessible, critical weaknesses in interoperability and reusability remain, limiting their integration into GenAI pipelines and their lawful reuse in clinical research. Automated FAIRness tools, though promising for scalability, frequently underestimated compliance due to rigid detection logic, underscoring the continuing need for domain-aware evaluation. By pinpointing systematic gaps and offering concrete recommendations—such as adopting controlled vocabularies, ensuring machine-readable licensing, and depositing in FAIR-supportive repositories—we provide a roadmap for improving dataset quality, compliance, and readiness for generative health applications. Within embodied health AI systems, these improvements extend beyond data management to serve as a data-level safety-assurance mechanism, ensuring that perception modules built on voice input remain transparent, verifiable, and compliant. Addressing these challenges is essential for building trustworthy, reproducible, and policy-aligned GenAI systems that leverage voice to advance diagnosis, monitoring, and equitable healthcare delivery.

Beyond technical evaluation, these findings have broader implications for safety and assurance in embodied AI. Foundation models and generative applications increasingly rely on large, heterogeneous voice datasets, and weaknesses in interoperability and reusability directly translate into risks of biased outputs, poor reproducibility, and regulatory non-compliance. For embodied or interactive systems—such as assistive robots, conversational companions, and diagnostic agents—these weaknesses manifest as uncertainty in sensory interpretation and unpredictable behavior, highlighting the need for auditable data pipelines. By showing where FAIR principles break down, our study provides an evidence base for interventions at multiple levels: dataset creators can improve metadata, licensing, and provenance practices; repositories can enforce structured templates and machine-readable standards; and policymakers can align guidance with FAIR-supportive infrastructures, as envisioned in initiatives such as the European Health Data Space. Embedding such improvements upstream ensures that voice biomarker data feeding GenAI pipelines are not only technically usable but also clinically credible and legally defensible. In this way, FAIRness becomes a cornerstone of assurance for embodied AI—linking transparent data governance to safe, reliable, and trusted behavior in real-world health systems.

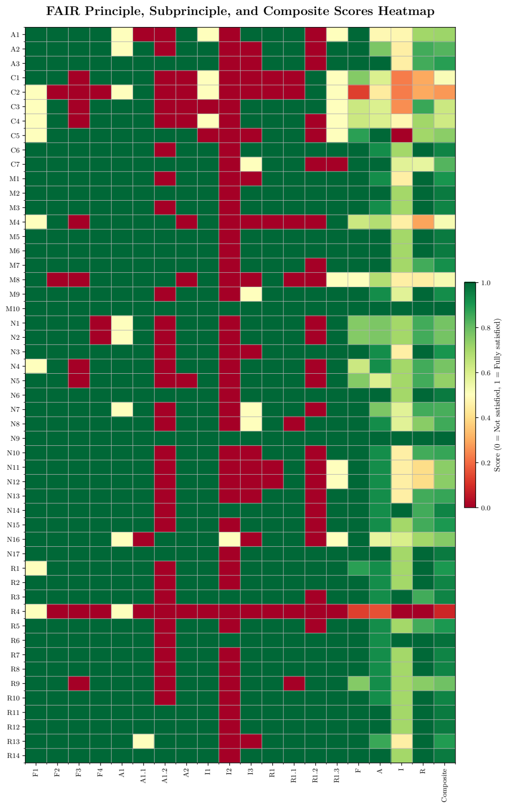

Figure 1: Heatmap of manual FAIR principle, subprinciple, and composite scores across disease domains.

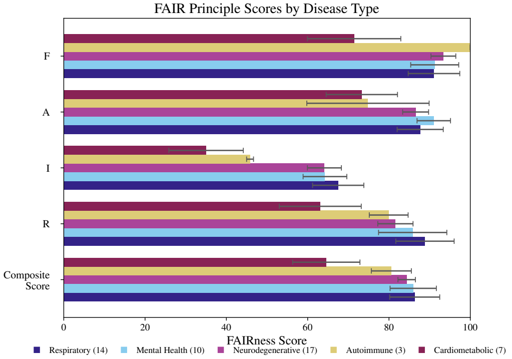

Figure 2: Mean FAIR principle scores by disease domain.

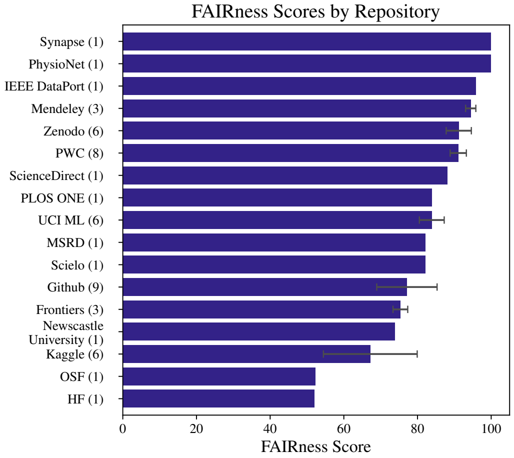

Figure 3: FAIRness scores (0–100) by repository/hosting venue, with dataset counts in parentheses.

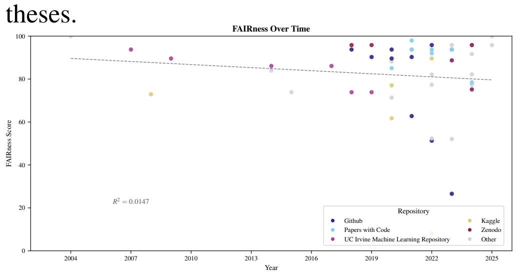

Figure 4: FAIRness scores plotted by dataset release year with best-fit regression line.

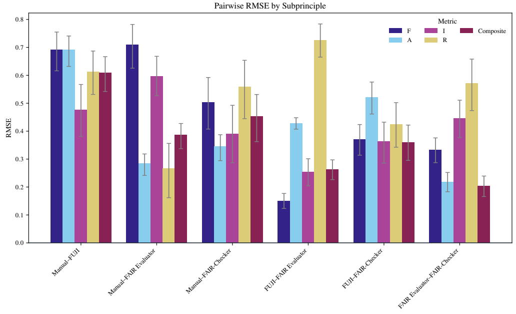

Figure 5: Pairwise RMSE of FAIR subprinciple scores between manual and automated tools (F-UJI, FAIR Evaluator, FAIR-Checker), summarized by principle and composite; error bars show variability.

References

[1] COVID-19 cough audio melspectrograms. https://www.kaggle.com/datasets/cuongdo/covid19-cough-audio-melspectrograms.

[2] Diabetes dataset. https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

[3] SciELO brazil - speech perception performance of subjects with type I diabetes mellitus in noise speech perception performance of subjects with type I diabetes mellitus in noise. https://www.scielo.br/j/bjorl/a/LtHn7JSyTNQJpmHnYWVqggs/?lang=en.

[4] Respiratory modulation of OscP and KorS, January 2016.

[5] Amod/mental_health_counseling_conversations · datasets at hugging face, May 2024.

[6] LIHVOICE/voice-and-diabetes-VOCADIAB. LIH - Projects from the Deep Digital Phenotyping Research Unit, June 2025.

[7] Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union, L 2025/327, 5 March 2025, 2025. URL `https://eur-lex.europa.eu/eli/reg/2025/327/oj/eng`.

[8] Scott J Adams, Julián N Acosta, and Pranav Rajpurkar. How generative AI voice agents will transform medicine. *npj Digital Medicine*, 8(1):353, 2025.

[9] Muhammad Fahreza Alghifari, Teddy Surya Gunawan, and Mira Kartiwi. Development of sorrow analysis dataset for speech depression prediction. In *2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 01–06, May 2023. doi: 10.1109/I2MTC53148.2023.10176040.

[10] Per A. Alm. Streptococcal infection as a major historical cause of stuttering: Data, mechanisms, and current importance. *Frontiers in Human Neuroscience*, 14, November 2020. ISSN 1662-5161. doi: 10.3389/fnhum.2020.569519.

[11] Christophe Bahim, Carlos Casorrán-Amilburu, Makx Dekkers, Edit Herczog, Nicolas Loozen, Konstantinos Repanas, Keith Russell, and Shelley Stall. The FAIR data maturity model: An approach to harmonise FAIR assessments. *Data Science Journal*, 19:41, October 2020. ISSN 1683-1470. doi: 10.5334/dsj-2020-041.

[12] Andrew Bailey and Mark D. Plumbley. Gender bias in depression detection using audio features, August 2021.

[13] Kirrie J. Ballard, Sharon Savage, Cristian E. Leyton, Adam P. Vogel, Michael Hornberger, and John R. Hodges. Logopenic and nonfluent variants of primary progressive aphasia are differentiated by acoustic measures of speech production. *PLOS ONE*, 9(2):e89864, February 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0089864.

[14] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digital Medicine*, 3(1):118, 2020.

[15] Yaël Bensoussan, Olivier Elemento, and Anaïs Rameau. Voice as an AI biomarker of health—introducing audiomics. *JAMA Otolaryngology–Head & Neck Surgery*, 150(4):283–284, 2024.

[16] Gorkem Serbes C. Sakar. Parkinson's disease classification, 2018.

[17] Gunvant Chaudhari, Xinyi Jiang, Ahmed Fakhry, Asriel Han, Jaclyn Xiao, Sabrina Shen, and Amil Khanzada. Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough, November 2020.

[18] Siye Chen, Linghan Li, Shuyu Han, Wei Luo, Wenxia Wang, Yufan Yang, Xiaomeng Wang, Wenmin Zhang, Mo Chen, and Zhiwen Wang. Review of voice biomarkers in the screening of neurodegenerative diseases. *Interdisciplinary Nursing Research*, 3(3):190–198, 2024.

[19] Ahmadul Karim Chowdhury, Saidur Rahman Sujon, Md. Shirajus Salekin Shafi, Tasin Ahmmad, Sifat Ahmed, Khan Md Hasib, and Faisal Muhammad Shah. Harnessing large language models over transformer models for detecting bengali depressive social media text: A comprehensive study. *Natural Language Processing Journal*, 7:100075, June 2024. ISSN 2949-7191. doi: 10.1016/j.nlp.2024.100075.

[20] Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. The turkish audio-visual bipolar disorder corpus.

[21] Timothy Clark, Harry Caufield, Jillian A. Parker, Sadnan Al Manir, Edilberto Amorim, James Eddy, Nayoon Gim, Brian Gow, Wesley Goar, Melissa Haendel, Jan N. Hansen, Nomi Harris, Henning Hermjakob, Marcin Joachimiak, Gianna Jordan, In-Hee Lee, Shannon K. McWeeney, Camille Nebeker, Milen Nikolov, Jamie Shaffer, Nathan Sheffield, Gloria Sheynkman, James Stevenson, Jake Y. Chen, Chris Mungall, Alex Wagner, Sek Won Kong, Satrajit S. Ghosh, Bhavesh Patel, Andrew Williams, and Monica C. Munoz-Torres. AI-readiness for biomedical data: Bridge2AI recommendations, October 2024.

[22] Harry Coppock, The Alan Turing Institute, UK Health Security Agency, Jobie Budd, Emma Karoune, Chris Holmes, Kieran Baker, Davide Pigoli, George Nicholson, Richard Payne, Ivan Kiskin, Josef Packham, Ana Tendero Cañadas, Selina Patel, Sabrina Egglestone, Alexander Titcomb, David Hurley, Lorraine Butler, Tracey Thornley, Jonathon Mellor, Stephen Roberts, Steven Gilmour, Björn Schuller, Vasiliki Koutra, Radka Jersakova, Peter Diggle, and Sylvia Richardson. The UK COVID-19 vocal audio dataset, May 2024.

[23] Nicholas Cummins, Alice Baird, and Bjoern W Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods (San Diego, Calif.)*, 151: 41–54, 2018.

[24] Anusuriya Devaraju and Robert Huber. An automated solution for measuring the progress toward FAIR research data. *Patterns*, 2(11):100370, November 2021. ISSN 26663899. doi: 10.1016/j.patter.2021.100370.

[25] Martin J. Dorahy, Amy Nesbit, Rachael Palmer, Bailey Wiltshire, Jacinta R. Cording, Donncha Hanna, Lenaire Seager, and Warwick Middleton. A comparison between auditory hallucinations, interpretation of voices, and formal thought disorder in dissociative identity disorder and schizophrenia spectrum disorders. *Journal of Clinical Psychology*, 79(9):2009–2022, 2023. ISSN 1097-4679. doi: 10.1002/jclp.23522.

[26] Raffaele Dubbioso, Myriam Spisto, Laura Verde, Valentina Virginia Iuzzolino, Gianmaria Senerchia, Elena Salvatore, Giuseppe De Pietro, Ivanoe De Falco, and Giovanna Sannino. Voice signals database of ALS patients with different dysarthria severity and healthy controls. *Scientific Data*, 11(1):1–14, July 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03597-2.

[27] Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: The use of vocal biomarkers from research to clinical practice. *Digital Biomarkers*, 5(1):78–88, April 2021. ISSN 2504-110X. doi: 10.1159/000515346.

[28] A. Gaignard, T. Rosnet, F. de Lamotte, V. Lefort, and MD Devignes. FAIR-checker: Supporting digital resource findability and reuse with knowledge graphs and semantic web standards. *Journal of Biomedical Semantics*, 14(1):7, 2023.

[29] P Groth, H Cousijn, T Clark, and C Goble. FAIR Data Reuse–The path through data citation. Data intelligence, 2 (1–2), 78–86, 2020.

[30] Pascal Hecker, Nico Steckhan, Florian Eyben, Björn W. Schuller, and Bert Arnrich. Voice analysis for neurological disorder recognition–a systematic review and perspective on emerging trends. *Frontiers in Digital Health*, 4, July 2022. ISSN 2673-253X. doi: 10.3389/fdgth.2022.842301.

[31] Juan Pablo Hernandez. ALS disease patient classification. 2, February 2025. doi: 10.17632/fbhc38zzm9.2.

[32] Truong V. Hoang, Quang H. Nguyen, Cuong Q. Nguyen, Phong X. Nguyen, and Hoang D. Nguyen. Sound-dr: Reliable sound dataset and baseline artificial intelligence system for respiratory illnesses, August 2023.

[33] Lihe Huang, Hao Yang, Yiran Che, and Jingjing Yang. Automatic speech analysis for detecting cognitive decline of older adults. *Frontiers in Public Health*, 12, August 2024. ISSN 2296-2565. doi: 10.3389/fpubh.2024.1417966.

[34] Rafiul Islam, Md. Taimur Ahad, Faruk Ahmed, Bo Song, and Yan Li. Mental health diagnosis from voice data using convolutional neural networks and vision transformers. *Journal of Voice*, November 2024. ISSN 0892-1997. doi: 10.1016/j.jvoice.2024.10.010.

[35] Pippa Iva, Joanne Fielding, Meaghan Clough, Owen White, Gustavo Noffs, Branislava Godic, Russell Martin, Anneke van der Walt, and Ramesh Rajan. Speech discrimination performance in multiple sclerosis dataset. *Data in Brief*, 33:106614, December 2020. ISSN 2352-3409. doi: 10.1016/j.dib.2020.106614.

[36] T. Tykalov J. Hlavnika. Early biomarkers of parkinson's disease based on natural connected speech, 2017.

[37] Alistair Johnson, Jean-Christophe Bélisle-Pipon, David Dorr, Satrajit Ghosh, Philip Payne, Maria Powell, Anais Rameau, Vardit Ravitsky, Alexandros Sigaras, Olivier Elemento, and Yael Bensoussan. Bridge2AI-voice: An ethically-sourced, diverse voice dataset linked to health information.

[38] Vidhi Khatwani. Vidhikhatwani/detecting-the-extent-of-hypertension-through-voice, January 2025.

[39] Max A. Little, Patrick E. McSharry, Stephen J. Roberts, Declan AE Costello, and Irene M. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1):23, June 2007. ISSN 1475-925X. doi: 10.1186/1475-925X-6-23.

[40] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, April 2009. ISSN 1558-2531. doi: 10.1109/TBME.2008.2005954.

[41] Ishaan Mahapatra and Nihar R. Mahapatra. Systematic fairness assessment of open voice biomarker datasets for mental health and neurodegenerative diseases. In Kamil Ekštein, Miloslav Konopík, Ondřej Pražák, and František Pártl, editors, *Text, Speech, and Dialogue*, pages 356–368, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-02548-7. doi: 10.1007/978-3-032-02548-7_30.

[42] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[43] Laureano Moro-Velazquez, Jorge A. Gomez-Garcia, Juan I. Godino-Llorente, Francisco Grandas-Perez, Stefanie Shattuck-Hufnagel, Virginia Yagüe-Jimenez, and Najim Dehak. Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease. *Scientific Reports*, 9(1):19066, December 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-55271-y.

[44] Ananya Muguli, Lancelot Pinto, Nirmala R, Neeraj Sharma, Prashant Krishnan, Prasanta Kumar Ghosh, Rohit Kumar, Shrirama Bhat, Srikanth Raj Chetupalli, Sriram Ganapathy, Shreyas Ramoji, and Viral Nanda. DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics, June 2021.

[45] Gustavo Noffs, Thushara Perera, Helmut Butzkueven, Scott C. Kolbe, Frederique M. C. Boonstra, Adam P. Vogel, and Anneke van der Walt. Longitudinal objective assessment of speech in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 91, November 2024. ISSN 2211-0348, 2211-0356. doi: 10.1016/j.msard.2024.105891.

[46] Betul Sakar Olcay Kursun. Parkinson's speech with multiple types of sound recordings, 2013.

[47] Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, June 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00937-4.

[48] Nemuel D. Pah, Veronica Indrawati, and Dinesh K. Kumar. Voice features of sustained phoneme as COVID-19 biomarker. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–9, 2022. ISSN 2168-2372. doi: 10.1109/JTEHM.2022.3208057.

[49] D. Trejo Pizzo and S. Esteban. IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples, December 2021.

[50] Carlos Prez. Parkinson dataset with replicated acoustic features, 2016.

[51] Alvaro Proaño, David P. Bui, José W. López, Nancy M. Vu, Marjory A. Bravard, Gwenyth O. Lee, Brian H. Tracey, Ziyue Xu, Germán Comina, Eduardo Ticona, Daniel J. Mollura, Jon S. Friedland, David A. J. Moore, Carlton A. Evans, Philip Caligiuri, Robert H. Gilman, and Tuberculosis Working Group in Peru. Data from: Cough frequency during treatment associated with baseline cavitary volume and proximity to the airway in pulmonary TB, March 2019.

[52] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022.

[53] Jessica Robin, John E Harrison, Liam D Kaufman, Frank Rudzicz, William Simpson, and Maria Yancheva. Evaluation of speech-based digital biomarkers: Review and recommendations. *Digital Biomarkers*, 4(3):99–108, 2020.

[54] B. M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, R. P. Paiva, I. Chouvarda, P. Carvalho, and N. Maglaveras. A respiratory sound database for the development of automated classification. In Nicos Maglaveras, Ioanna Chouvarda, and Paulo de Carvalho, editors, *Precision Medicine Powered by pHealth and Connected Health*, pages 33–37, Singapore, 2018. Springer. ISBN 978-981-10-7419-6. doi: 10.1007/978-981-10-7419-6_6.

[55] Milan Rusko, Róbert Sabo, Marián Trnka, Alfréd Zimmermann, Richard Malaschitz, Eugen Ružický, Petra Brandoburová, Viktória Kevická, and Matej Škorvánek. Slovak database of speech affected by neurodegenerative diseases. *Scientific Data*, 11(1):1320, December 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-04171-6.

[56] James W. Schwoebel, Joel Schwartz, Lindsay A. Warrenburg, Roland Brown, Ashi Awasthi, Austin New, Monroe Butler, Mark Moss, and Eleftheria K. Pissadaki. A longitudinal normative dataset and protocol for speech and language biomarker research, August 2021.

[57] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R, Prasanta Kumar Ghosh, and Sriram Ganapathy. Coswara – a database of breathing, cough, and voice sounds for COVID-19 diagnosis. In *Interspeech 2020*, pages 4811–4815, October 2020. doi: 10.21437/Interspeech.2020-2768.

[58] Caroline Stellmach and Michael Rusongoza Muzoora. How to assess fairness of your data – a summary of testing two FAIR validators. In *MEDINFO 2023 — The Future Is Accessible*, pages 154–158. IOS Press, 2024. doi: 10.3233/SHTI230946.

[59] Ml Tlachac, Ermal Toto, Joshua Lovering, Rimsha Kayastha, Nina Taurich, and Elke Rundensteiner. EMU: Early mental health uncovering framework and dataset. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1311–1318, December 2021. doi: 10.1109/ICMLA52953.2021.00213.

[60] Andreas Triantafyllopoulos, Anastasia Semertzidou, Meishu Song, Florian B. Pokorny, and Björn W. Schuller. Introducing the COVID-19 YouTube (COVYT) speech dataset featuring the same speakers with and without infection, September 2022.

[61] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, April 2010. ISSN 1558-2531. doi: 10.1109/TBME.2009.2036000.

[62] Maxim Vashkevich and Yulia Rushkevich. Classification of ALS patients based on acoustic analysis of sustained vowel phonations. *Biomedical Signal Processing and Control*, 65:102350, March 2021. ISSN 17468094. doi: 10.1016/j.bspc.2020.102350.

[63] Maxim Vashkevich, Alexander Petrovsky, and Yuliya Rushkevich. Bulbar ALS detection based on analysis of voice perturbation and vibrato. In *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 267–272, Poznan, Poland, September 2019. IEEE. ISBN 978-83-62065-36-3. doi: 10.23919/SPA.2019.8936657.

[64] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[65] Mark D Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Merc Crosas, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, 6(1):174, 2019.

[66] Jiaxin Ye, Xin-cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10096370.

[67] Konstantia Zarkogianni, Edmund Dervakos, Giorgos Filandrianos, Theofanis Ganitidis, Giorgos Stamou, and Konstantina Nikita. Smarty4Covid dataset, October 2022.