A Survey on Public Data Sets Related to Chronic Diseases

Ioannis Bilionis
Adhera Health SL, Spain
Universitat Pompeu Fabra, Barcelona, Spain
0000-0002-1435-4358

Luis Fernandez-Luque
Adhera Health SL, Spain
0000-0001-8165-9904

Carlos Castillo
ICREA and Universitat Pompeu Fabra
Barcelona, Spain
0000-0003-4544-0416

***Abstract*—This paper presents an extensive survey of publicly available biomedical datasets, revealing four dozen databases connected to chronic diseases, such as cancer, diabetes, heart diseases, and COVID-19, among others. Our main objective is to describe these datasets, highlighting commonalities and best practices among them, and to raise awareness about the wealth of data available to study chronic diseases, focusing on the importance of the sociodemographic data in biomedical research.**

***Index Terms*—databases, health care, machine learning, chronic diseases**

I. Introduction

The incidence of chronic diseases is rapidly increasing not only among the elderly but also in young children. Cancer, diabetes, and cardiovascular diseases are a scourge of everyday life as they undermine people's health and degrade the quality of life and well-being of society. According to the World Health Organization, the long duration of chronic diseases can be attributed to a combination of genetic, physiological, environmental and behavioral factors. Chronic diseases are responsible for the death of 41 million people each year, equivalent to 74% of all deaths globally [1].

Research in the field of biomedicine is increasingly using data-driven methods including artificial intelligence and machine learning. Therefore, it is critical to implement protocols on fairness, transparency, and explainability of algorithms to mitigate the potential risks posed by, for instance, not considering social-environmental factors (e.g., sex and age). Hence, in this paper we present a new multidimensional collection of datasets that contribute to the research of chronic conditions, by giving an overview and highlighting commonalities.

II. Methodology

The target of our data survey are publicly available datasets disseminated through scientific studies, and created either for studying a chronic disease, or including chronic disease patients as the majority of the study subjects. The datasets matching these conditions were identified through extensive research in publicly available health information repositories.

The initial step in our survey was to consider the most common and critical chronic diseases. Initially, in addition to heart disease, cancer and diabetes, which are the leading causes of death and disability in the US [2], we added multiple sclerosis, hypertension, arthritis, chronic kidney disease, mental and sleep disorders (e.g. chronic depression, anxiety), chronic obstructive pulmonary disease (COPD), dementia, Alzheimer's and Parkinson's diseases, which belong to the most common chronic conditions spectrum in adults [3]- [5]. Likewise, due to COVID-19 pandemic incidence, a challenging chronic illness, long COVID, is identified through the persistent symptoms upon recovery from COVID-19.

TABLE I
Summary of the data collection methodology and results.

Chronic Conditions	**Search Engines**	**Results**
Alzheimer's		
Cancer	Google	**n=48** Publicly
Chron. Kidney Dis.	Scholar	Avalaible Databases
COPD	Zenodo [6]	
COVID-19	Kaggle [7]	**Embedded Features**
Diabetes	UCI ML [8]	Socio-Demographics
Heart Diseases	PhysioNet [9]	Physical Activity
Mental Disorders	NIH-Mendeley [10]	Clinical Data
Multiple Sclerosis	EMBL-EBI [11]	Sleep Data
Parkinson's		Biometric Data
Rheumatoid Arthritis		Psychometrics

The next step, as depicted in Table I, was to perform a series of searches in well-known search engines (e.g., Google Scholar, PubMed, Nature) and online data repositories (e.g., Kaggle, Zenodo). We considered clusters of queries where in each cluster various combinations of words, related to chronic diseases and data science, lead to the identification of relevant information. We combined these queries with the names of the chronic conditions we identified as targets. The first cluster of queries focuses on retrieving information related to biomedical research based on time series. This cluster's queries included: "chronic diseases", "symptoms", "comorbidities", "wearable", "biometric", "smart", "vital signs", "clinical data", "time series", "cancer", "heart disease", "diabetes","degenerative", "multivariate data". The second cluster includes keywords related to the investigation of mental disorders and psychological symptoms through machine learning algorithms based on the biometric profile and daily activity (physical activity, sleep, wearable sensors' data) of chronic disease patients. Specifically, the following keywords were used: "machine learning", "sensor", "smartwatch", "physical activity", "sleep", "mental

This work is funded by Spanish Ministry of Science and Innovation (REF:DIN2021-011865).

TABLE II

Summary of datasets related to chronic conditions along. The number of subjects is denoted by N. A dark circle means the dataset includes sociodemographic data (SD), clinical data (CD), physical activity (PA), time-series data (TSD), and/or psychometric data (PSY). We indicate whether the dataset is hosted at an archival repository and thus has a Digital Object Identifier (DOI), and if it is accompanied by a publication (Paper). We also include whether there is an ethics statement (ETH), and/or details on the data anonymization procedure (AN).

				Features					Publication			Details	
#	**Dataset**	**Therapeutic Area**	**N**	**SD**	**CD**	**PA**	**TSD**	**PSY**	**Year**	**DOI**	**Paper**	**ETH**	**AN**
1	Karoly et al. [12]	Brain Disease	31	●	○	○	○	●	2021	●	●	●	●
2	Andrzejak et al. [13]	Brain Disease	500	○	●	○	●	○	2001	●	●	○	○
3	Sada et al. [14].	Cancer	35	●	●	●	○	●	2021	●	●	●	○
4	Stump et al. [15]	Cancer	50	●	●	○	○	○	2020	●	●	●	○
5	U. Hosp of Coimbra [16]	Cancer	165	●	●	○	○	○	2015	●	●	○	○
6	U. Hosp. of Caracas [17]	Cancer	858	●	●	○	○	○	2017	●	●	○	○
7	Islam et al. [18]	Chron. Kidney Dis.	202	●	●	○	○	○	2020	●	●	○	○
8	Rogan et al. [19]	Chron. Kidney Dis.	226	●	●	○	○	○	2017	●	●	●	○
9	Soundarapandian et al. [20]	Chron. Kidney Dis.	400	●	●	○	○	○	2015	○	○	○	○
10	PTB Diagnostic ECG [21]	Common Aging Dis.	290	●	●	○	●	○	2020	●	●	○	○
11	Anne Arundel Med. Center [22]	COVID-19	117	●	●	○	○	○	2020	●	●	●	●
12	Welltory [23]	COVID-19	186	●	○	●	●	●	2020	●	●	●	●
13	Hajifathalian et al. [24]	COVID-19	664	●	●	○	●	○	2020	●	●	●	○
14	Alavi et al. [25]	COVID-19	3,318	○	○	●	●	○	2021	●	●	●	●
15	Mishra et al. [26]	COVID-19	5,262	○	○	●	●	○	2020	●	●	●	●
16	COVID-19 focus patients [27]	COVID-19	4,5M	●	●	○	○	○	2020	○	○	○	○
17	COV19 Open Data Mexico [28]	COVID-19	6,6M	●	●	○	○	○	2021	○	○	○	○
18	BIG IDEAs [29]	Diabetes	16	●	●	●	○	○	2021	●	●	●	●
19	D1NAMO [30]	Diabetes	29	●	●	●	●	○	2018	●	●	●	●
20	Washington U. [31]	Diabetes	70	○	●	●	○	○	1994	○	○	○	○
21	Smith et al. [32]	Diabetes	768	●	●	○	○	○	1988	●	●	○	○
22	Strack et al. [33]	Diabetes	70K	●	●	○	○	○	2014	●	●	●	●
23	St. Petersburg INCART [34]	Heart Disease	32	●	●	○	●	○	2008	●	○	○	○
24	U. of Creighton [35]	Heart Disease	35	○	●	○	●	○	1986	●	○	○	○
25	MIT-BIH Arrhythmia [36]	Heart Disease	47	●	○	○	●	○	2005	●	●	○	○
26	European ST-T [37]	Heart Disease	78	●	●	○	●	○	1992	●	●	○	○
27	SHAREE [38]	Heart Disease	139	●	●	○	●	○	2015	●	●	●	○
28	Shen et al. [39]	Heart Disease	200	●	●	○	●	○	2020	●	●	●	●
29	Detrano et al. [40]	Heart Disease	920	●	●	○	●	○	1989	●	●	●	○
30	Guvenir et al. [41]	Heart Disease	452	●	●	○	●	○	1997	●	●	○	○
31	Golovenkin et al. [42]	Heart Disease	1700	●	●	○	○	○	2020	●	●	●	●
32	Framingham heart study [43]	Heart Disease	4,240	●	●	○	○	○	2010	●	○	●	○
33	Zheng et al. [44]	Heart Disease	10K	●	●	○	●	○	2020	●	●	●	●
34	Cardiovascular Disease [45]	Heart Disease	70K	●	●	○	○	○	N/A	○	○	○	○
35	Aziz et al. [46]	Hypertension	160	●	○	○	○	●	2020	●	●	●	○
36	MMASH [47]	Mental Disorder	22	●	○	●	●	●	2020	●	●	●	○
37	SWELL-KW [48]	Mental Disorder	25	○	○	●	○	●	2014	●	●	●	●
38	YAAD [49]	Mental Disorder	25	○	○	○	●	●	2021	●	●	●	●
39	Depresjon [50]	Mental Disorder	55	●	○	●	●	●	2018	●	●	○	○
40	Ihmig et al. [51]	Mental Disorder	57	○	○	○	○	●	2020	●	●	●	●
41	U. of Michigan [52]	Mental Disorder	62	○	○	●	○	●	2019	●	●	●	○
42	Rekeland et al. [53]	Myalgic Encephalomyelitis	27	●	○	●	●	●	2022	●	●	●	●
43	Fuller et al. [54]	None	46	○	○	●	○	○	2019	●	●	●	○
44	Sakar et al. [55]	Parkinson	1040	○	○	○	●	○	2013	●	●	●	○
45	Apnea-ECG [56]	Sleep Disorder	21	○	○	○	●	○	2000	●	●	○	○
46	Luo et al. [57]	Sleep Disorder	28	●	○	●	●	●	2020	●	●	●	○
47	T. U. of Darmstadt [58]	Sleep Disorder	42	●	○	●	○	○	2014	●	●	●	●
48	Thyroid Disease Data Set [59]	Thyroid Disease	7200	○	○	○	○	○	1987	○	○	○	○
				73%	62%	31%	48%	25%		87%	81%	60%	33%

disorder", "stress", "anxiety", "psychometrics", "emotion", "depression", "mhealth". The third cluster aims to collect information and data about COVID-19 through queries consisted of the keywords: "COVID-19", "symptoms", "risk factors", "detection", "heart rate", "patients", "mortality", "prediction", "severe", "illness", "hospitalization".

These queries resulted in a large number of results, which we manually reviewed using the following inclusion criteria: must be a publicly available dataset, must refer to a chronic condition, must include clinical information or physical activity data, and must contain at least 10 patients. Out of 110 potential sources of data related to chronic diseases, the presented collection consists of 48 publicly and readily available health datasets; the majority of the excluded datasets did not provide direct free access to the data, and instead require a formal data request to be obtained.

III. Datasets Description and Comparison

In this section, we present the characteristics and descriptive statistics extracted by examining the metadata and publications releasing each dataset. Table II provides summary information about each dataset's characteristics. We include the therapeutic areas, sample size, year of publication, as well as information about the available features: sociodemographic data, clinical data, physical activity, time-series data, and psychometric data. In addition, we describe whether the dataset is hosted at an archival repository and thus has a Digital Object Identifier number, and whether it is accompanied by a publication.

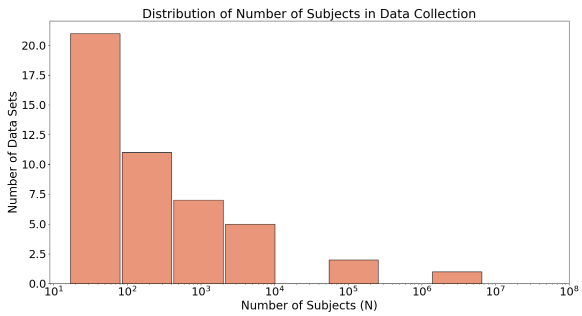

Fig. 1. Histogram indicating the distribution of number of subjects (N) throughout the component databases of the introduced data collection.

Finally, we include whether the data release includes a detailed ethics statement and/or details on the anonymization process.

Firstly, with respect to the therapeutic area, human data from 13 different therapeutic areas have been spotted whilst the 14th one, marked as "None", represents wearable data of healthy subjects in a study focused on physical activity [54]. Because of the beneficial impact of the physical activity and exercise on the prevention and treatment of the chronic diseases [60], potentially relevant insights could be extracted of that data set. Furthermore, 18% of the surveyed datasets focus on COVID-19, having the largest sample sizes compared to other chronic diseases, followed by cancer, diabetes, heart diseases and mental disorders. A histogram of sample sizes is included in Figure 1.

Secondly, sociodemographic information (such as sex and age) linked to the participants is present in 73% of the datasets. Figure 2 illustrates the presence of three features that we find across many of the surveyed datasets: sex, age and Body Mass Index (BMI). We can observe that sex and age are more frequent than BMI, whereas at least one of these features appears in 34 of the surveyed datasets. Moreover, clinical features are provided in more than half of the data sets, whilst psychological information is present in one quarter of the proposed data collection's components, mostly linked to physical activity information.

Thirdly, most of the data sets are identified by a DOI number and have been published either in public data repositories or as supplementary material for a scientific paper. Regarding the type of the data files, common formats (plain text, comma-separated values, Microsoft Excel) are used in more than 40 data sets, whereas the rest of the data are stored in less common, usually propietary formats (.hea, .qrs, .mat), for which nevertheless free/open source libraries are available in most popular programming languages. These libraries are often created by reverse engineering, hence correctness when reading the data using them cannot be always guaranteed.

Finally, about 60% of the datasets include an ethics declaration statement. None of the datasets we examined includes direct identifiers of people, but only about half of the studies indicate the specific process done for anonymization.

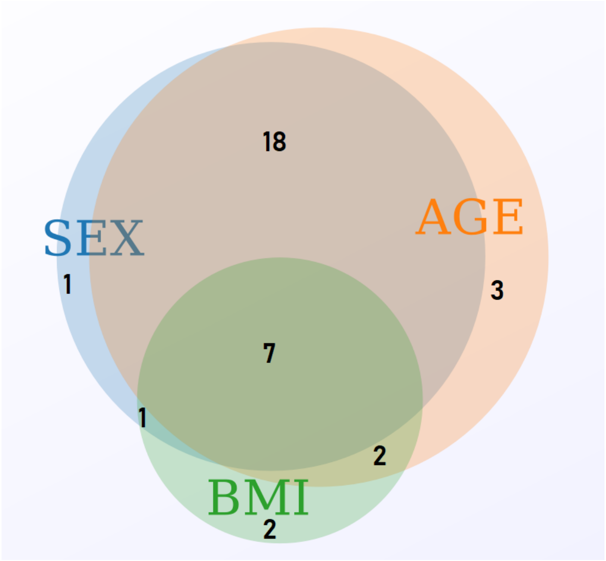

Fig. 2. Schematic representation of Sex-Age-BMI variables through Venn diagram.

IV. Discussion

We uncover that there are at least 48 publicly available datasets to study chronic diseases, and that these datasets cover a wide range of diseases including cancer, diabetes, and recently long COVID. A common element we find in these datasets is the contribution of clinicians and biomedical researchers to their creation. In most cases, a detailed description of the process from collection to data release is provided, providing credibility to the clinical annotations and ensuring the medical relevance of the features. The datasets have a wide range of sample sizes, which seem to follow an exponential distribution (Figure 1 has logarithmic scale in the X axis) and include from a few tens to several million people. Most of the datasets we found were created in the last five years, which suggests an accelerated process and that we will see more dataset creation in the coming years. The datasets we surveyed include relevant sociodemographic, clinical, biometric and psychometric information. It was common to find sex and age of the patients in the datasets, as well as BMI.

Data privacy protocols and regulations have been established to control the excessive collection and illegitimate disclosure of human subjects data. At the same time, data about age and sex can be useful for clinical research and to detect and mitigate unwanted biases in the dataset, or in the models built from it. Therefore, the inclusion of sociodemographic information in healthcare databases is a positive trend if personal data protection protocols are in place and may lead to better statistical models with good algorithmic fairness properties.

Future work includes the further extension of the proposed Data Collection by including underrepresented diseases (hypertension, brain diseases, among others) and other common neurodegenerative diseases (such as multiple sclerosis, Alzheimer's). Furthermore, we are investigating the potential unique characteristics and mechanisms of dependencies between symptoms and outcomes in diverse subgroups. The identification of intrinsic differences between, for instance, statistical model performance for women and men, or younger and older patients, could contribute to detecting and mitigating possible algorithmic discrimination risks.

References

[1] World Health Organization. (2022). *Noncommunicable diseases. Fact Sheets. Newsroom.* Retrieved from: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[2] Sherry L. Murphy, Kenneth D. Kochanek, Jiaquan Xu, Elizabeth Arias: Mortality in the United States, 2020. NCHS Data Brief No. 427, December 2021.

[3] Anderson, E., Durstine, J. L. (2019). Physical activity, exercise, and chronic diseases: A brief review. Sports Medicine and Health Science, 1(1), 3-10.

[4] Stein, D. J., Benjet, C., Gureje, O., Lund, C., Scott, K. M., Poznyak, V., Van Ommeren, M. (2019). Integrating mental health with other non-communicable diseases. Bmj, 364.

[5] Busse, R., Scheller-Kreinsen, D., Zentner, A. (2010). Tackling chronic disease in Europe: strategies, interventions and challenges (No. 20). WHO Regional Office Europe.

[6] European Organization For Nuclear Research, OpenAIRE. (2013). Zenodo.

[7] https://www.kaggle.com/

[8] Asuncion, A., Newman, D. (2007). UCI machine learning repository.

[9] Goldberger et al.(2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

[10] Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010;160(Pt 2):1299-303. PMID: 20841894; PMCID: PMC3408090.

[11] Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic acids research, 38(suppl_2), W695-W699.

[12] Karoly et al.(2021). Multiday cycles of heart rate are associated with seizure likelihood: An observational cohort study. EBioMedicine, 72, 103619.

[13] Andrzejak et al.(2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E, 64(6), 061907.

[14] Sada et al.(2021). Harnessing digital health to objectively assess cancer-related fatigue: The impact of fatigue on mobility performance. PloS one, 16(2), e0246101.

[15] Stump, T., Spring, B., Marchese, S., Alshurafa, N., Robinson, J. (2019). Toward a precision behavioral medicine approach to addressing high-risk sun exposure: a qualitative analysis. JAMIA open, 2(4), 547–553.

[16] Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., Carvalho, A. (2015). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of biomedical informatics, 58, 49-59.

[17] Fernandes, K., Cardoso, J. S., Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. In Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8 (pp. 243-250). Springer International Publishing.

[18] Islam et al.(2020). Risk factor prediction of chronic kidney disease based on machine learning algorithms. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 952–957).

[19] Rogan et al.(2017). Quality of life measures predict cardiovascular health and physical performance in chronic renal failure patients. PloS one, 12(9), e0183926.

[20] Soundarapandian, P., Rubini, L. J., Eswaran, P. (2015). Chronic_Kidney_Disease Data Set. UCI Mach. Learn. Repository, School Inf. Comput. Sci., Univ. California, Irvine, CA, USA.

[21] Bousseljot, R., Kreiseler, D., Schnabel, A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik / Biomedical Engineering, 40(s1), 317-318.

[22] Turcotte et al.(2020). Risk factors for severe illness in hospitalized Covid-19 patients at a regional hospital. PloS one, 15(8), e0237558.

[23] Ponomarev, A., Tyapochkin, K., Surkova, E., Smorodnikova, E., Pravdin, P. (2021). Heart rate variability as a prospective predictor of early Covid-19 symptoms. medRxiv, 2021-07.

[24] Hajifathalian et al.(2020) Development and external validation of a prediction risk model for short-term mortality among hospitalized U.S. COVID-19 patients: A proposal for the COVID-AID risk tool. PLoS ONE 15(9): e0239536.

[25] Alavi, A., Bogu, G.K., Wang, M. et al. Real-time alerting system for COVID-19 and other stress events using wearable data. Nat Med 28, 175–184 (2022).

[26] Mishra et al.(2020). Pre-symptomatic detection of COVID-19 from smartwatch data. Nature biomedical engineering, 4(12), 1208–1220.

[27] Mani, S. (2020, June). COVID-19 focus patients, Version 6. Retrieved from: https://www.kaggle.com/datasets/shirmani/characteristics-corona-patients

[28] Larasa, O. (2020, October).COV19 Open Data Mexico, Version 1. Retrieved from: https://www.kaggle.com/datasets/omarlarasa/cov19-open-data-mexico

[29] Cho, P., Kim, J., Bent, B., Dunn, J. (2022). BIG IDEAs Lab Glycemic Variability and Wearable Device Data (version 1.0.0). PhysioNet.

[30] Dubosson et al.(2018). The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management (1.2.0) [Data set]. Zenodo.

[31] Kahn M. (2019). Diabetes Data Set. Retrieved from: http://archive.ics.uci.edu/ml/datasets/Diabetes

[32] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.

[33] Strack et al.(2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed research international, 2014.

[34] Tihonenko, V., Khaustov, A. (2008). St Petersburg INCART 12-lead Arrhythmia Database. Retrieved from: https://physionet.org/content/incartdb/1.0.0/.

[35] Nolle FM, Badura FK, Catlett JM, Bowser RW, Sketch MH. CREI-GARD, a new concept in computerized arrhythmia monitoring systems. Computers in Cardiology 13:515-518 (1986).

[36] Moody, G. B., Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. IEEE engineering in medicine and biology magazine, 20(3), 45-50.

[37] Taddei et al.(1992). The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. European heart journal, 13(9), 1164-1172.

[38] Melillo et al.(2015). Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. PloS one, 10(3), e0118504.

[39] Shen et al.(2020). An Open-Access arrhythmia database of wearable electrocardiogram. Journal of Medical and Biological Engineering, 40, 564-574.

[40] Detrano et al.(1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304-310.

[41] Guvenir, H. A., Acar, B., Demiroz, G., Cekin, A. (1997, September). A supervised machine learning algorithm for arrhythmia analysis. In Computers in Cardiology 1997 (pp. 433-436). IEEE.

[42] Golovenkin SE, Gorban AN, Mirkes EM, et al. Myocardial infarction complications Database. 2020.

[43] Ashish Bhardwaj. (2022). Framingham heart study dataset. Retrieved from: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

[44] Zheng et al.(2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Scientific data, 7(1), 48.

[45] Ulianova, S. (N/A).Cardiovascular Disease dataset, Version 1.Retrieved from: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[46] Aziz et al.(2020). Determining hypertensive patients' beliefs towards medication and associations with medication adherence using machine learning methods. PeerJ, 8, e8286.

[47] Rossi et al.(2020). A public dataset of 24-H multi-levels psycho-physiological responses in young healthy adults. Data, 5(4), 91.

[48] Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., Kraaij, W. (2014, November). The swell knowledge work dataset for stress and user modeling research. In Proceedings of the 16th international conference on multimodal interaction (pp. 291-298).

[49] Dar, M. N., Rahim, A., Akram, M. U., Khawaja, S. G., Rahim, A. (2022, May). YAAD: Young Adult's Affective Data Using Wearable ECG and GSR sensors. In 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2) (pp. 1-7). IEEE.

[50] Garcia-Ceja et al. (2018, June). Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In Proceedings of the 9th ACM multimedia systems conference (pp. 472-477).

[51] Ihmig, F., Gogeascoechea, A., Schäfer, S., Lass-Hennemann, J., Michael, T.. (2020). Electrocardiogram, skin conductance and respiration from spider-fearful individuals watching spider video clips.

[52] McGinnis et al.(2019). Rapid detection of internalizing diagnosis in young children enabled by wearable sensors and machine learning. PloS one, 14(1), e0210267.

[53] Rekeland et al.(2022) Activity monitoring and patient-reported outcome measures in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome patients. PLoS ONE 17(9): e0274472.

[54] Fuller, D. (2020). Replication Data for: Using machine learning methods to predict physical activity types with Apple Watch and Fitbit data using indirect calorimetry as the criterion.

[55] Sakar et al.(2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics, 17(4), 828–834.

[56] Penzel, T., Moody, G. B., Mark, R. G., Goldberger, A. L., Peter, J. H. (2000, September). The apnea-ECG database. In Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163) (pp. 255-258). IEEE.

[57] Luo, H., Lee, P.A., Clay, I., Jaggi, M., De Luca, V. (2020). Assessment of fatigue using wearable sensors: a pilot study. Digital biomarkers, 4(1), 59–72.

[58] Borazio, M., Berlin, E., Kucukyildiz, N., Scholl, P., Van Laerhoven, K. (2014, September). Towards benchmarked sleep detection with wrist-worn sensing units. In 2014 IEEE International Conference on Healthcare Informatics (pp. 125-134). IEEE.

[59] Quinlan R. (1987). Thyroid Disease Data Set. Retrieved from: https://archive.ics.uci.edu/ml/datasets/thyroid+disease

[60] Anderson, E., Durstine, J. L. (2019). Physical activity, exercise, and chronic diseases: A brief review. Sports Medicine and Health Science, 1(1), 3-10.