Hearing examinations in Southern Denmark (HESD): Database description and preprocessing

MANUELLA LECH CANTUARIA[1,2,*], ELLEN RABEN PEDERSEN[3], METTE SØRENSEN[2], FRANS BOCH WALDORFF[4], JESPER HVASS SCHMIDT[1,5,6]

[1] *Institute of Clinical Research, Faculty of Health, University of Southern Denmark, DK-5000 Odense, Denmark*

[2] *Department of Environment and Cancer, Danish Cancer Society Research Center, DK-2100 Copenhagen, Denmark*

[3] *The Mærsk Mc-Kinney Møller Institute, Faculty of Engineering, University of Southern Denmark, DK-5230 Odense, Denmark*

[4] *Research Unit of General Practice, Department of Public Health, University of Southern Denmark, DK-5000 Odense, Denmark*

[5] *Department of Audiology, Odense University Hospital, DK-5000 Odense, Denmark*

[6] *OPEN, Odense Patient data Explorative Network, Odense University Hospital, DK-5000 Odense, Denmark*

All hearing examinations from the public health system of the Region of Southern Denmark have been electronically recorded from 1996 to 2018 and merged into a single database, named the Hearing Examinations of Southern Denmark (HESD) database. This database contains hearing information for more than 143,000 adults, totaling 271,575 valid pure-tone audiograms. The use of this dataset, however, needs to be preceded by an intensive preprocessing procedure in order for the data to be used for research purposes. This study is aimed at describing the HESD database, as well as the preprocessing steps and rules used to classify different types of hearing loss. An initial overview of the different types of hearing profiling and their distribution among our sample is also provided.

INTRODUCTION

The World Health Organization has pointed out hearing impairment as one of the most frequent sensory disabilities worldwide and a leading cause of disease burden. The global prevalence of this disorder for males and females older than 15 years was estimated to be 9.8 % and 12.2 %, respectively (Stevens *et al.*, 2011). Given the aging of the population in many countries, it is likely that the prevalence of hearing loss continues to increase (Cunningham and Tucci, 2017).

Besides aging, there are several other risk factors that can potentially result in hearing loss (HL), such as noise exposure, genetic mutations, cardiovascular diseases (CVD) and ototoxic drugs (Agrawal *et al*., 2009; Cunningham and Tucci, 2017). On the other hand, difficulties in hearing may have critical impacts on the

*Corresponding author: mlca@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

individual's ability to navigate in life (e.g., communication), which can increase the risk for other health outcomes. As an example, different studies have consistently found associations between hearing loss and dementia, suggesting hearing loss as an important modifiable risk factor for this disease (Thomson *et al*., 2017).

Even though different hypotheses linking hearing loss and health outcomes, as well as the biological mechanisms behind it, already exist, there is still much to be explored in this regard. Large sample-sized epidemiological data on hearing performance are therefore essential in this context. Within this scope, the HESD (Hearing Examinations in Southern Denmark) database has arisen. The HESD database establishment was based on the data electronically recorded in AuditBase, which is a data capture system used in the public medical system of the Region of Southern Denmark since 1996. However, as this dataset was not originally implemented for research purposes, its use needs to be preceded by an intensive cleaning and preprocessing procedure.

The purpose of this study is to describe the establishment and preprocessing of the HESD database, as well as the information thereby available. We further describe the rules used to classify different types of HL, in order to assess associations between HL characteristics and different diseases. An overview of the different types of hearing profiling and their distribution among our sample is also provided.

METHODS

Database establishment

The HESD database is based on the data electronically recorded in AuditBase from February 1996, March 1998 and June 2003 to 2018 in the public clinics of Vejle, Odense and Sønderborg, respectively. The large majority of the examinations are from patients that had a previous complaint about hearing or any suspected HL (e.g., older patients). The clinics had the software gradually implemented the duration needed for testing and adaptation varied. AuditBase, which was developed by the company Auditdata, is used for collecting and managing auditory clinical data, and therefore consists of a large source of documented data over the years. The clinical data collection is based on the process as defined in ISO 14155 (ISO, 2011) for medical device trials, whereas the audiometric measurements conducted in all of the clinics are based on ISO 8253-1 (ISO, 2010), which addresses the procedure for pure-tone air conduction (AC) and bone conduction (BC) threshold audiometry. The use of the clinical records for research purposes has been authorized by the Danish Patient Safety Authority.

AuditBase contains recorded information of the most important auditory tests, such as: (i) pure-tone audiometry (AC and BC thresholds); (ii) acoustic reflexes (ipsilateral and contralateral stapedius); (iii) speech audiometry (speech reception thresholds and discrimination scores determined using monosyllabic words (Elberling *et al.*, 1989); (iv) tympanometry; and (v) Weber test. Additionally, the system stores person-related information on the patients (e.g. name, date of birth, sex), as well as data on all the people who have performed the testing. The patients

are identified by a unique ID, and each ID is associated to the patient's personal identification number, which can be used to link the HESD data with data from all health registries in Denmark.

Cleaning and preprocessing steps

The raw data extracted from the AuditBase system demands an intensive preprocessing procedure, so that it can be used for research purposes. This is mainly because of the huge size of the dataset, high vulnerability to missing data and lack of consistency between audiograms (e.g. audiograms may vary in terms of the number of frequencies and ears tested, as well as the type of measurements obtained for each ear). Figure 1 shows a simplified flowchart describing the most relevant preprocessing steps used to prepare and transform the data to a suitable form.

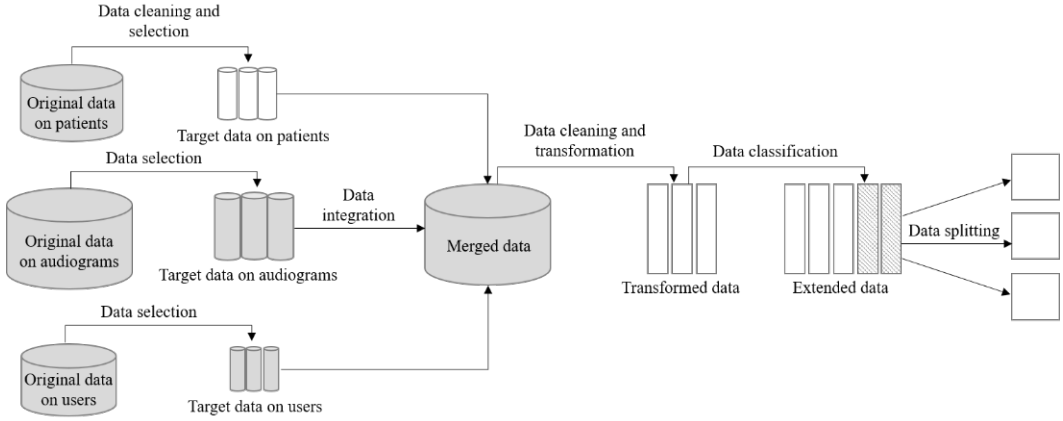

Fig. 1: Steps involved in the HESD database preprocessing.

The preprocessing procedure starts with the original datasets for patients and users' information, as well as the original dataset containing hearing thresholds (measured by, for example, AC, BC and soundfield) and acoustic reflexes thresholds. The latter dataset is organized in a stacked (i.e. narrow) format, meaning that the thresholds obtained for each of the curves present in one specific audiogram were disposed in separate rows. For all the original datasets, the relevant variables were selected and renamed. Additionally, the patients' dataset was cleaned, so that all nonexistent patients (i.e., patients with invalid identification numbers) were removed from the data. In the data integration step, all datasets were merged by the patient's ID.

The merged dataset was further cleaned. In this stage, which was the most demanding one, we have looked for, for example, blank curves and curves with large amount of missing, audiograms obtained for the same patient and at the same examination date, identical audiograms obtained at different dates and audiograms that seem to have been obtained for testing purposes. After this step, we made sure that, for each patient and examination date, we had data for only one audiogram. The transformation stage consisted in: (i) selecting solely the data for the most relevant and frequent audiogram-related measurements (i.e., AC, BC and acoustic reflexes);

and (ii) unstacking the data, so that there is only one row of data for each patient and examination date. It is worth mentioning that a fraction of patients has had their hearing measured more than once along the years in which the data were collected. The data from all these audiograms (i.e. visits) were kept in the dataset.

With the dataset transformed accordingly, we have created categorical variables that indicate:

1. Threshold status, which captures whether the thresholds were masked, out-of-range (i.e., exceeded the maximum output level), crossed-over (i.e., when the sound presented to the test ear was heard by the non-test ear) or uncertain.
2. Data missingness, which captures whether data are available for only one ear and whether BC thresholds were measured. A numerical variable indicating the amount of missing data (for both the total frequency range and the most relevant frequencies) was also created.
3. Data reliability, which captures whether BC thresholds are substantially higher (>10 dB) than the AC threshold measured at the same frequency and masking was done correctly.

Database organization

As the HESD database assembles audiogram data for more than 20 years, differences in the amount of data collected may exist. The lack of data for some specific measurements can also be due to variations on the patients' hearing and symptoms reported, as well as time limitations during the examination. Among all measurements available in the AuditBase system, AC thresholds are certainly the most frequent and the most applicable for research purposes. Given that, the HESD datasets organization is centralized on the AC data. Therefore, all audiograms with AC thresholds available at the most relevant octave frequencies (0.5-4 kHz) for at least one of the ears and obtained for patients older than 18 years were included in the final HESD database.

Hearing loss assessment

To extract the most important information from the pure-tone audiograms, the AC thresholds were used to describe the HL indicated by each audiogram present in the HESD database. We have created a set of rules (Table 1) to classify HL in terms of:

1. Severity, as defined by the pure-tone average (PTA) of 0.5, 1, 2 and 4 kHz.
2. Asymmetry based on interaural AC thresholds differences at octave frequencies between 0.25 to 8 kHz (Margolis and Saly, 2007).
3. Audiogram configuration based on the methods proposed by Demeester *et al*. (2009) and Hannula *et al*. (2011) involving the means of the thresholds at consecutive octave frequencies (i.e., 0.25/0.5, 1/2 and 4/8 kHz) and measurements of the poorer and better thresholds for low, mid and high frequencies.

When BC thresholds were also available, HL was also categorized in terms of:

4. Type of lesion, as defined by the number of air-bone gaps at octave frequencies between 0.25 and 2 kHz (Margolis and Saly, 2007). Air-bone gaps at 4 kHz were

not considered given uncertainties in the measurements for that specific frequency in cases of sensorineural HL (Margolis *et al*., 2013).

	Categories	**Rules**
Severity	Low or no hearing loss	PTA < 20 dB HL
	Mild	20 ≤ PTA < 40 dB HL
	Moderate	40 ≤ PTA < 70 dB HL
	Severe	70 ≤ PTA < 95 dB HL
	Profound	PTA ≥ 95 dB HL
	Not classified	PTA could not be calculated for that ear
Asymmetry	Asymmetric HL	Asymmetry is considered when there are three or more interaural differences (ID) ≥ 10 dB, two or more ID ≥ 15 dB, or one ID ≥ 20 dB (Margolis and Saly, 2007)
	Symmetric HL	
	Not classified *	
Audiogram configuration	Flat	Based on Demeester *et al*. (2009) and Hannula *et al*. (2011)
	High freq. gently sloping (HFGS)	
	High freq. steeply sloping (HFSS)	
	Low frequency ascending (LFA)	
	Mild frequency U-shape (MFU)	
	Mild freq. reverse U-shape (MFRU)	
	Unspecified *	
Type of lesion	No hearing loss	A conductive component is considered when there is a 10-dB air-bone gap (ABG) at three or more frequencies (within 0.25 – 2 kHz), or a 15-dB ABG at any one frequency (within 0.5 – 2 kHz)
	Conductive	
	Sensorineural	
	Mixed	
	Unspecified *	

* Asymmetry, type of lesion or audiogram configuration could not be defined due to data limitation.

Table 1: Hearing loss classification scheme used for the HESD database.

RESULTS

The final number of pure-tone audiograms available in the HESD dataset is 271,575 (Figure 2), which corresponds to hearing data available for 143,794 adults. The data cleaning step was characterized by a drop of 260,894 observations. This is explained by the elimination of blank audiograms, audiograms with a large amount of missing data, repeated curves and invalid measurement due to, for example, testing. The largest drop, however, was in the data transformation stage, where the number of observations was reduced by 417,142. This drop is due to the unstacking of the data, meaning that the information that was previously arranged in several rows (i.e., observations) is now organized in a single row.

Out of the total number of audiograms, 77% presents data for BC thresholds, 38% presents data for CL acoustic reflexes, 29% presents data for IL acoustic reflexes and 84% presents data for speech audiometry. The results displayed in Table 2 reveal that 68% of the audiograms were obtained for older adults (≥ 60 years) and 54% were obtained for male patients. Out of the 143,794 patients, 80,069 (i.e., 55.7%) presented data for only one audiogram in the final dataset.

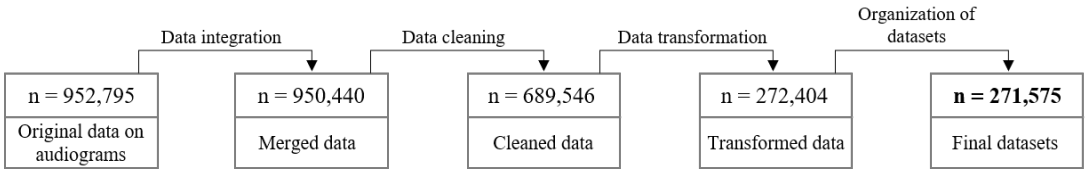

Fig. 2: Number of observations at each of the main data preprocessing stages.

Descriptive statistics on the distribution of hearing loss characteristics (Table 2) for all audiograms available in the HESD database showed asymmetric (53%) and moderate HL (47% left ear and 46% right ear) as the most frequent cases observed in the database. In terms of the type of lesion, the highest prevalence found was for sensorineural (38% left, 37% right), followed by mixed HL (12% left, 13% right).

The most prevalent configuration among audiograms was HFSS (48% left, 46% right), followed by HFGS (23% for both ears) and flat (14% left, 15% right). A chi-squared test of proportion revealed that the distribution of audiogram configurations was significantly different between the left and the right ear (p-value < 0.01).

DISCUSSION

We presented a recently established database that assembles hearing examination data for 143,794 adults (i.e., 271,575 records) who have undergone audiometric testing in the public system of Southern Denmark. The raw data gathered by the AuditBase system required an intensive preprocessing procedure, which also included the development of variables able to classify hearing, more specifically in terms of severity, asymmetry, audiogram configuration and type of lesion.

A variety of definitions for audiometric categorization can be found in literature; however, there is a lack of consistency and standardization among them (Margolis and Saly, 2008). In our study, some of the derived rules were based on the classification system developed by Margolis and Saly (2007), which has been previously validated and defined in order to maximize the agreement among the expert judges involved. This was the case for the definition of asymmetric HL in our study. Our results have shown that the majority of the audiograms were correspondent to asymmetric HL. This high rate of asymmetry may be explained by the fact that these definitions were considerably broad as they, for example, do not require interaural AC thresholds differences at consecutive frequencies.

Given the high number of missing BC thresholds, we were unfortunately unable to categorize the type of lesion for 43% of the cases. Nevertheless, our results showed sensorineural HL as the most predominant type of lesion, followed by mixed HL. Similar results were also found in a previous study from Margolis and Saly (2008), after analyzing audiometric records for a large sample of 16,818 patients.

In terms of the audiogram configuration, we found steep sloping AC curves (i.e., HFSS) to be the most prevalent shape for audiograms, followed by gently sloping

(i.e., HFGS). This result is in agreement with Hannula *et al*. (2011), who have assessed the prevalence of audiogram configurations among 850 adults between 54 and 66 years old, using similar configuration categories as in our study. On the other hand, Demeester *et al*. (2009), who have also used similar methodology, have found flat audiograms as the most prevalent configuration. However, the prevalence for HFGS and HFSS configuration was also shown to be high for their study sample. It is important to point out the limitations of the comparison between results found in the HESD and these studies, as there are fundamental sampling differences (i.e., the HESD is based on clinical data of adults of all ages, whereas the other studies are based on population data of adults between 54-66 years old).

Age at exam		
< 60 years	87,039 (32)	
≥ 60 years	184,536 (68)	
Sex *		
Male	145,565 (54)	
Female	124,211 (46)	
Asymmetry		
Asymmetric HL	144,953 (53)	
Symmetric HL	112,505 (42)	
Not classified	14,117 (5)	
Severity	**Left ear**	**Right ear**
Low or no hearing loss	31,017 (11)	33,247 (12)
Mild	75,181 (28)	76,555 (28)
Moderate	128,232 (47)	124,520 (46)
Severe	22,432 (8)	22,553 (8)
Profound	7,035 (3)	7,030 (3)
Not classified	7,678 (3)	7,670 (3)
Audiogram configuration	**Left ear**	**Right ear**
Flat	38,935 (14)	42,413 (15)
High freq. gently sloping (HFGS)	62,595 (23)	63,579 (23)
High freq. steeply sloping (HFSS)	131,145 (48)	125,250 (46)
Low frequency ascending (LFA)	4,639 (2)	4,964 (2)
Mild frequency U-shape (MFU)	1,436 (1)	1,520 (1)
Mild freq. reverse U-shape (MFRU)	4,362 (2)	4,133 (2)
Unspecified	28,463 (10)	29,716 (11)
Type of lesion	**Left ear**	**Right ear**
No hearing loss	15,816 (6)	17,625 (6)
Conductive	3,209 (1)	3,346 (1)
Sensorineural	102,122 (38)	101,076 (37)
Mixed	33,458 (12)	34,158 (13)
Unspecified	116,970 (43)	115,370 (43)

* Sex data was not available for 1799 audiograms.

Table 2: Demographics and prevalence of HL characteristics in terms of asymmetry, severity, type of lesion and configuration. Results are given in number (%).

Even though the HESD database demanded intensive preprocessing steps, it is remarkable the amount of hearing data that has been merged into a single database. The insights obtained in the study highlight the potential of the HESD database as a promising source of audiology-related epidemiological data, not just to evaluate hearing profiling among adults, but to further explore the effects of hearing impairment on a range of health outcomes.

REFERENCES

Agrawal, Y., Platz, E.A., and Niparko, J.K. (**2009**). "Risk Factors for Hearing Loss in US Adults: Data from the National Health and Nutrition Examination Survey, 1999 to 2002," Otol. Neurotol., **30**, 139-145.

Cunningham, L.L., and Tucci, D.L. (**2017**). "Hearing Loss in Adults," New Engl. J. Med., **377**(25), 2465-2473. doi:10.1056/NEJMra1616601.

Demeester, K., Wieringen, A., Hendrickx, J., Topsakal, V., Fransen, E., Laer, L., Camp, G.V., and Heyning, P.V. (**2009**). "Audiometric shape and presbycusis," Int. J. Audiol., **48**, 222-232. doi: 10.1080/14992020802441799

Elberling, C., Ludvigsen, C.W., and Lyregaard, P.E. (**1989**). "DANTALE: a new Danish speech material," Scand. Audiol., **18**(3), 169-175.

Hannula, S., Bloigu, R., Majamaa, K., Sorri, M., and Mäki-Torkko, E. (**2011**). "Audiogram configurations among older adults: Prevalence and relation to self-reported hearing problems," Int. J. Audiol., **50**, 793-801. doi: 10.3109/14992027.2011.593562

ISO. (**2011**) ISO 14155:2011 "Clinical investigation of medical devices for human subjects – Good clinical practice." Geneva.

ISO. (**2010**) ISO 8253-1:2010 "Acoustics — Audiometric test methods — Part 1: Pure-tone air and bone conduction audiometry." Geneva.

Margolis, R.H., and Saly, G.L. (**2007**). "Toward a standard description of hearing loss," Int. J. Audiol., **46**, 746-758. DOI: 10.1080/14992020701572652.

Margolis, R.H., and Saly, G.L. (**2008**). "Distribution of Hearing Loss Characteristics in a Clinical Population," Ear Hearing, **29**(4), 524-532. doi: 10.1097/AUD.0b013e3181731e2e

Margolis, R.H., Eikelboom, R.H., Johnson, C., Ginter, S.M., Swanepoel, D.W., and Moore, B.C.J. (**2013**). "False air-bone gaps at 4 kHz in listeners with normal hearing and sensorineural hearing loss," Int. J. Audiol., **52**(8), 526–532. doi:10.3109/14992027.2013.792437.

Stevens, G., Flaxman, S., Brunskill, E., Mascarenhas, M., Mathers, C. D., and Finucane, M. (**2011**). "Global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries," Eur. J. Public Health, **23**(1), 146-152. doi:10.1093/eurpub/ckr176

Thomson, R.S., Auduong, P., Miller, A.T., and Gurgel, R.K. (**2017**). "Hearing loss as a risk factor for dementia: A systematic review," Laryngoscope Invest. Otolaryngol., **2**, 69-79. doi: 10.1002/lio2.65