City Research Online

City, University of London Institutional Repository

Citation: Christensen, J. H., Pontoppidan, N. H., Bossing, R., Anisetti, M., Bamiou, D-E., Spanoudakis, G., Murdin, L., Bibas, T., Kikidiks, D., Dimakopoulos, N., et al (2019). Fully Synthetic Longitudinal Real-World Data From Hearing Aid Wearers for Public Health Policy Modeling. Frontiers in Neuroscience, 13, 850. doi: 10.3389/fnins.2019.00850

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/22981/

Link to published version: https://doi.org/10.3389/fnins.2019.00850

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

frontiers in Neuroscience

DATA REPORT
published: 13 August 2019
doi: 10.3389/fnins.2019.00850

Fully Synthetic Longitudinal Real-World Data From Hearing Aid Wearers for Public Health Policy Modeling

Jeppe H. Christensen[1], Niels H. Pontoppidan[1], Rikke Rossing[1], Marco Anisetti[2], Doris-Eva Bamiou[3], George Spanoudakis[4], Louisa Murdin[5], Thanos Bibas[6], Dimitris Kikidiks[6], Nikos Dimakopoulos[7], Giorgos Giotis[7] and Apostolos Ecomomou[8]*

[1] Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark, [2] Department of Computer Science, University of Milan, Milan, Italy, [3] The Ear Institute, Brain Institute, UCL, London, United Kingdom, [4] Department of Computer Science, City University of London, London, United Kingdom, [5] Guy's and St. Thomas' NHS Foundation Trust, London, United Kingdom, [6] Department of Otolaryngology, National & Kapodistrian University of Athens, Athens, Greece, [7] ATC Innovation Lab, Athens, Greece, [8] Athens Medical Group, Athens, Greece

OPEN ACCESS

Edited by:
Mary Rudner, Linköping University, Sweden

Reviewed by:
Michael A. Stone, University of Manchester, United Kingdom
Kathryn Arehart, University of Colorado Boulder, United States

***Correspondence:**
Jeppe H. Christensen
jych@eriksholm.com

Specialty section:
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *29 May 2019*
Accepted: *30 July 2019*
Published: *13 August 2019*

Citation:
Christensen JH, Pontoppidan NH, Rossing R, Anisetti M, Bamiou D-E, Spanoudakis G, Murdin L, Bibas T, Kikidiks D, Dimakopoulos N, Giotis G and Ecomomou A (2019) Fully Synthetic Longitudinal Real-World Data From Hearing Aid Wearers for Public Health Policy Modeling. Front. Neurosci. 13:850. doi: 10.3389/fnins.2019.00850

Keywords: real-world data, longitudinal data, hearing aids, public health policy, evidence-based, hearing loss, sound exposure

INTRODUCTION

Approximately one-third of people over 65 years of age, and 5% of the world's population, is affected by hearing loss (HL) (World Health Organization, 2017). Disabling HL is associated with early cognitive decline in adults (Olusanya et al., 2014), and when unaddressed, HL restricts social integration and reduces employment and educational opportunities, hampers emotional well-being and, thus, poses an economic challenge at both the individual and national level (Wilson et al., 2017). Moreover, more and more individuals suffer from HL, which is primarily due to increases in everyday noise exposure and an increase of the aging population (World Health Organization, 2017). Despite the fact that age-related HL is the third leading cause of years lived with disability (Vos et al., 2017), the population of individuals with hearing loss is underserved because few public health policies focus on prevention, intervention, and rehabilitation for age-related HL (Reavis et al., 2016). This inadequate focus has been attributed to a lack of evidence supporting policies that actively promote hearing healthcare (Moyer, 2012; Barker et al., 2016). However, this specific issue is targeted in the EU-funded H2020 project EVOTION (www.h2020evotion.eu), which collects a large volume of heterogenous data from almost 1,000 hearing aid (HA) users with varying degrees of hearing loss to support the development of evidence-based policy making within the hearing healthcare field (Spanoudakis et al., 2017; Gutenberg et al., 2018). Data are being collected from five sources: (i) hearing aids, (ii) a smartphone app, (iii) a biosensor, (iv) audiology clinics, and (v) electronic health records. The hearing aids log data about the user's sound environment (Pontoppidan et al., 2018), hearing aid use (i.e., on/off) and hearing aid settings on a minute-by-minute basis; a phone app developed for the study collects information about the user's physical location via GPS (Dritsakis et al., 2018). Thus, EVOTION will provide an evidence base for formulating and evaluating the impacts of public health policy pertaining to prevention, early diagnosis, and treatment/rehabilitation for adults with hearing impairment.

Here, we share the first outcome of EVOTION in the form of a data-set to inspire, encourage, and motivate a data-driven analytical approach to evidence-based healthcare policy modeling using real-world longitudinal data. The data-set includes information relating to patterns of real-world hearing aid usage and sound environment exposure. Undoubtedly, many such data-sources will be

available for researchers and policy-makers in the future, and the data-set presented here can act as a first step of building and testing potential statistical models (Christensen et al., 2018, 2019).

Specifically, the data-set represents a sub-sample of the data being collected in EVOTION. It contains longitudinally sampled observations from 53 individuals and includes the following measures: the sound environment, the hearing aid setting, logging time (timestamps), ID, and the degree of hearing loss on the best hearing ear of the individuals. Note that the ID (an integer between 1 and 53, randomly assigned to each individual) does not link to the real identity of the participants.

Data are considered sensitive as they contain personal and health related information, and EVOTION adhere to strict data ethics by applying privacy-aware big data analytics (Anisetti et al., 2018). Here, we overcome the problem of sharing such personal data by working with a fully synthetic data-set that preserves structural and statistical properties of the original data (see section Technical Validation), without allowing the extraction of personal information (see section Data Synthesization). Thus, the synthetic data-set can readily be shared among professionals.

METHODS

Protocol

Data collection in EVOTION follow a published protocol (Spanoudakis et al., 2017; Dritsakis et al., 2018), is ongoing, and spans 12 months from the day of recruitment to the end of study participation. The data-set presented here represents a synthesized data-sample from EVOTION. The source data span a mean of 17 days of hearing aid usage (minimum 2 and maximum 54 days), 53 participants, and a total of ∼5,000 h of hearing aid usage.

Data Acquisition

Each participant in EVOTION is supplied with a pair of EVOTION hearing aids and a Samsung A3 smartphone. The hearing aids are connected to the smartphone via low-energy Bluetooth, and a custom developed EVOTION app (developed by ATC, Athens) on the smartphone logs a real-time data vector every minute consisting of data parameters from both the hearing aid's processing of sound from the microphone, hearing aid settings, and the smartphone's GPS. When connected to a wireless network, the smartphone app transmits the logged data vector (see **Table 1**) to the EVOTION data repository, which is located on secure distributed servers.

Clinical (e.g., audiometric tests) and demographical data are collected from the hearing clinics that have been involved in recruiting participants for EVOTION.

Acquisition of Acoustic Variables

The EVOTION hearing aids implements proprietary algorithms for continuous estimates of the acoustic environment sensed by the calibrated hearing aid microphones. The continuous estimates are derived by level estimators that implements very short time constants in four frequency channels: full bandwidth, low, mid, and high frequencies. The dynamic range of the estimators covers the dynamic range of the microphones. Noise floor is estimated from the lowest values and the modulation envelope from the largest values within a longer time-window. The SNR is obtained by subtracting the noise floor from the SPL, and the modulation index by subtracting the noise floor from the modulation envelope. Also, from the full bandwidth signal, a proprietary algorithm estimates if the current sound environment is quiet, noise, speech, or speech-in-noise dominated (i.e., the "SoundClass" variable in **Table 1**). In total, the acoustic environment is described by 21 variables, which the hearing aid transmits to the smartphone over Bluetooth every minute.

Data Synthetization

To enable data-sharing, and uphold differential privacy, we synthesized the data-set from a subset of the EVOTION repository data (the source) using DataSynthesizer (for details, see Ping et al., 2017). First, DataSynthesizer generates empirical conditional probability density functions (PDFs) for each variable of the data-source by computing a Bayesian network using the GreedyBayes algorithm with up to 4 parents—that is, the values in one variable can be conditioned on the values of up to four other data variables. Next, the synthesized data-set is generated by randomly drawing from the empirical PDFs while injecting each drawn sample with Laplacian noise with location 0 and scale $4(d - k)/n\varepsilon$ to preserve privacy. Here, n is the size of the source input (rows), $\varepsilon = 0.1$, d is the number of variables, and $k = 4$. Thus, the covariance between source parameters are preserved by allowing the empirical PDFs to be conditioned in the Bayesian network. In addition, to mask absolute position from GPS measures, each latitude and longitude coordinate were centered for each individual (i.e., subtracted by the mean latitude and longitude) prior to synthesization.

Limitations and Updates

While EVOTION collects a large amount of heterogeneous data, the dataset described here represents a sub-collection of the parameters from a sub-population of all the individuals enrolled in EVOTION, which limits the data-set's usability for hypothesis testing. We expect to update the data-set with more observations and data-types once these become available in the EVOTION project for synthesization. In addition, we do not have access to low-level details of the signal-processing taking place in the hearing aids. Thus, we do not include real-time data on how the hearing aids autonomously reacts to the sound environment (e.g., adjusting noise reduction or compression characteristics).

Data Format

The data-set is stored as a comma separated values (csv) file with each row representing one vector of observations associated to a timestamp. The included 399.500 observations represent 28 variables (columns) and they are described in **Table 1**.

Data Access

The newest version of the dataset is named "EVOreal_time_synth.csv" and is uploaded to zenodo.org

TABLE 1 | Data-set variables logged every minute.

Variable name	Description	Type	Units/levels
ID	Identifier	Integer	1:53
SoundClass	A value describing the sound environment. The value is derived by the hearing aids internal processing of the acoustic variables	Categorical	QUIET, SPEECH, SPEECH-IN-NOISE, NOISE
hProg	A value describing the active hearing aid program (Dritsakis et al., 2018)	Categorical	MEDIUM, LOW, HIGH, HIGH+
hVol	A value describing the active hearing aid volume state	Integer	Steps (−9:4) represent 2.5 dB up or down from default (0)
LonRel	Relative longitude (centered for each individual)	Continuous	GPS
LatRel	Relative latitude (centered for each individual)	Continuous	GPS
LowSPL	The sound pressure level (SPL) measured in low frequency bands	Continuous	dB
MidSPL	SPL measured in middle frequency bands	Continuous	dB
HighSPL	SPL measured in high frequency bands	Continuous	dB
fbSPL	SPL measured in full bandwidth	Continuous	dB
LowNf	The noise floor (Nf) measured in low frequency bands	Continuous	dB
MidNf	Nf measured in middle frequency bands	Continuous	dB
HighNf	Nf measured in high frequency bands	Continuous	dB
fbNf	Nf measured in full bandwidth	Continuous	dB
LowME	The modulation envelope (ME) measured in low frequency bands	Continuous	dB
MidME	ME measured in middle frequency bands	Continuous	dB
HighME	ME measured in high frequency bands	Continuous	dB
fbME	ME measured in full bandwidth	Continuous	dB
Timestamp	Local time of record	ISO 8601	YYYY:MM:DD HH:MM:SS
LowSNR	The signal-to-noise ratio (SNR) from low frequency bands as SNR = SPL – Nf	Continuous	dB
MidSNR	SNR from middle frequency bands	Continuous	dB
HighSNR	SNR from high frequency bands	Continuous	dB
fbSNR	SNR from full bandwidth	Continuous	dB
LowMI	The modulation index (MI) from low frequency bands as MI = ME – Nf	Continuous	dB
MidMI	MI from middle frequency bands	Continuous	dB
HighMI	MI from high frequency bands	Continuous	dB
fbMI	MI from full bandwidth	Continuous	dB
PTA4	Pure tone average (PTA) across 4 testing frequencies (0.5, 1, 2, and 4 kHz) on the best hearing ear	Continuous	dB hearing threshold

Frequency bands corresponds to 0–1.33; 1.33–2.14; 4.14–10; and 0–10 kHz.

and accessible via the following DOI: https://doi.org/10.5281/zenodo.2668210.

TECHNICAL VALIDATION

The Bayesian network generating the fully synthetic data ensures that covariance between different variables are preserved. To validate that, indeed, dependencies are still present in the fully synthetic data-set we computed statistics from both the acoustic variables (only the full-bandwidth variables were selected) and the "Timestamp" variable (see **Figure 1**).

Acoustic Variables

According to the hearing aids' estimation of the acoustic variables (see section Acquisition of Acoustic Variables), we would expect certain dependencies in the synthesized data. For example, the estimated noise floor (fbNf) should ideally always be lower than the estimated modulation envelope (fbME). The correlation matrix of the full bandwidth acoustic variables and their classification into the four discrete environments by color ("SoundClass") are shown in **Figure 1A**. As expected, the noise floor is almost always lower than the modulation envelope (expect for a few outliers, see row 3, column 2 in **Figure 1A**). The outliers that do not follow the expected pattern are not generated by the synthesization process but instead reflects noise in the hearing aids' estimation method (outliers are also present in the source data). The color-coding indicates that clustering of the sound environment depends on more than one acoustic parameter. For example, in **Figure 1A** (panel in row 3, column 2), the environment is dominantly classified as "Quiet" for low levels of noise floor and modulation envelope. But as the modulation envelope passes ∼60 dB the environment changes to either "Speech" or "Speech in Noise" despite no changes in the noise floor. This 3rd order dependency further validates that the data synthesization process preserves structural dependencies in the source data.

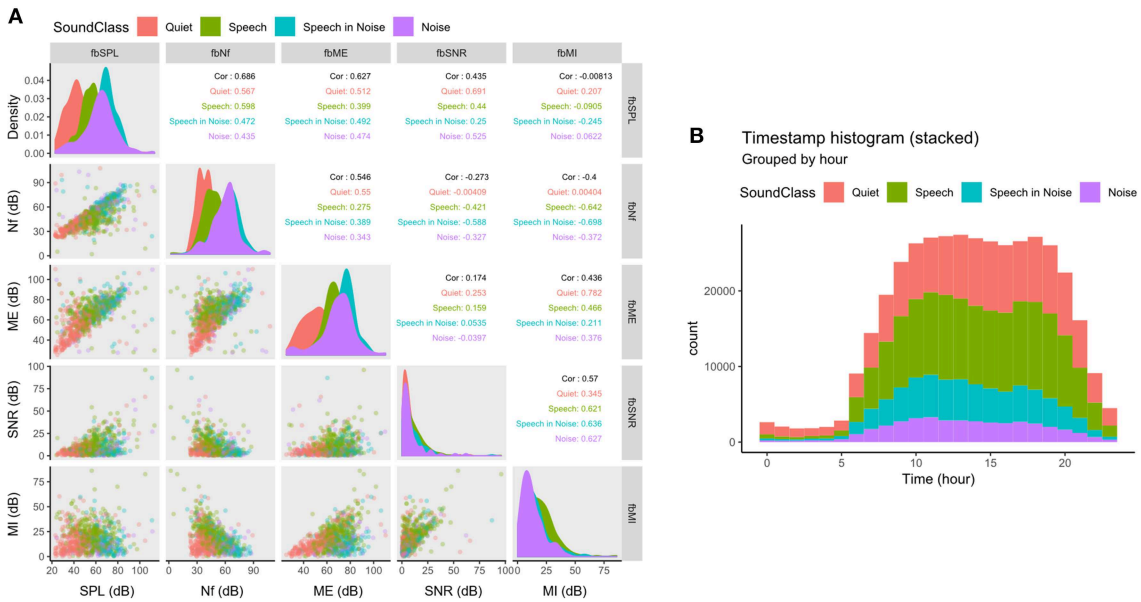

FIGURE 1 | Scatterplot matrix (panels below the diagonal), density plots (panels in the diagonal), and correlation matrix (panels above the diagonal) of the full-bandwidth acoustic variables **(A)** and a stacked histogram of the "Timestamp" variable **(B)**. Data in both **(A,B)** are grouped and colored by the classified sound environment by the variable "SoundClass." SPL, Sound pressure level; Nf, Noise floor; ME, Modulation envelope; SNR, Signal-to-noise ratio; MI, Modulation index.

Timestamps

Each observation in the data-set is associated with a timestamp. Thus, we can aggregate the timestamps to test the hypothesis that hearing aid usage is not uniformly distributed throughout the day and, from this, validate that the data synthesization process preserves the distributional statistics of the "Timestamp" variable. **Figure 1B** shows the histogram of timestamps binned by hour from 0 to 23. Most timestamps fall between 5 a.m. and 9 p.m. (usual awake hours) with peaks around noon and evening (6 p.m.). In addition, most data are logged in "Quiet" or "Speech" environments, which reflects what is reported in the literature (Humes et al., 2018). Thus, the distribution of timestamps and the dependency between timestamps and the sound environment both exhibit characteristics expected from real life use of hearing aids.

CONCLUSION

We present a synthesized data-set containing longitudinal observations of hearing aid use and associated sound environments. The data represent real life behavior of individuals with hearing loss wearing hearing aids. The underlying reason for sharing these data is to motivate the use of such data for public health policy modeling—that is, identifying which models derive useful "high-level" information and insights from such "low-level" data observations in the field of hearing healthcare.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript/supplementary files.

AUTHOR CONTRIBUTIONS

JC wrote the text, analyzed and synthesized the data-set, and prepared the figures. NP co-authored the text. RR, D-EB, LM, TB, DK, and AE recruited the participants. MA, GS, ND, and GG fascilitated the datalogging by technical developments.

FUNDING

This work was supported by EU-funded project EVOTION (contract n. H2020-727521).

REFERENCES

Anisetti, M., Ardagna, C., Bellandi, V., Cremonini, M., Frati, F., and Damiani, E. (2018). Privacy-aware big data analytics as a service for public health policies in smart cities. *Sustain. Cities Soc.* 2018, 68–77. doi: 10.1016/j.scs.2017.12.019

Barker, F., Mackenzie, E., Elliott, L., Jones, S., and de Lusignan, S. (2016). Interventions to improve hearing aid use in adult auditory rehabilitation. *Cochrane Database Syst. Rev.* 2016:cd010342. doi: 10.1002/14651858.cd010342

Christensen, J. H., Petersen, M. K., Pontoppidan, N. H., and Cremonini, M. (2018). "Big data analytics in healthcare: design and implementation for a hearing aid case study," in *2018 14th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)* (Washington, DC).

Christensen, J. H., Pontoppidan, N. H., Anisetti, M., Bellandi, V., and Cremonini, M. (2019). "Improving hearing healthcare with Big Data analytics of real-time hearing aid data," in *2019 IEEE World Congress on Services (SERVICES)* (Washington, DC).

Dritsakis, G., Kikidis, D., Koloutsou, N., Murdin, L., Bibas, A., Ploumidou, K., et al. (2018). Clinical validation of a public health policy-making platform for hearing loss (EVOTION): protocol for a big data study. *BMJ Open* 8:e020978. doi: 10.1136/bmjopen-2017-020978

Gutenberg, J., Katrakazas, P., Trenkova, L., Murdin, L., Brdarić, D., Koloutsou, N., et al. (2018). *Big Data for Sound Policies: Toward Evidence-Informed Hearing Health Policies*.

Humes, L. E., Rogers, S. E., Main, A. K., and Kinney, D. L. (2018). The acoustic environments in which older adults wear their hearing aids: insights from datalogging sound environment classification. *Am. J. Audiol.* 27, 594–603. doi: 10.1044/2018_AJA-18-0061

Moyer, V. A. (2012). Screening for hearing loss in older adults: US Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* 157, 655–661. doi: 10.7326/0003-4819-157-9-201211060-00526

Olusanya, B. O., Neumann, K. J., and Saunders, J. E. (2014). The global burden of disabling hearing impairment: a call to action. *Bull. World Health Org.* 92, 367–373. doi: 10.2471/BLT.13.128728

Ping, H., Stoyanovich, J., and Howe, B. (2017). "DataSynthesizer: privacy-preserving synthetic datasets," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management - SSDBM '17* (Washington, DC).

Pontoppidan, N. H., Li, X., Bramsløw, L., Johansen, B., Nielsen, C., Hafez, A., et al. (2018). "Data-driven hearing care with time stamped data-logging," *Proceedings of the International Symposium on Auditory and Audiological Research, 6, Adaptive Processes in Hearing*. Retrieved from: https://proceedings.isaar.eu/index.php/isaarproc/article/view/2017-15/332

Reavis, K. M., Tremblay, K. L., and Saunders, G. (2016). How can public health approaches and perspectives advance hearing health care? *Ear Hearing* 37, 376–380. doi: 10.1097/AUD.0000000000000321

Spanoudakis, G., Kikidis, D., Bibas, A., Katrakazas, P., Koutsouris, D., and Pontoppidan, N. H. (2017). "Public health policy for management of hearing impairments based on big data analytics: EVOTION at genesis," in *17th IEEE International Bio-Informatics and Bio-Engineering Conference* (Washington, DC). Retrieved from: http://openaccess.city.ac.uk/18205/1/BIBE_2017_paper_85.pdf

Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., et al. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* 390, 1211–1259. doi: 10.1016/S0140-6736(17)32154-2

Wilson, B. S., Tucci, D. L., Merson, M. H., and O'Donoghue, G. M. (2017). Global hearing health care: new findings and perspectives. *Lancet* 390, 2503–2515. doi: 10.1016/S0140-6736(17)31073-5

World Health Organization (2017). *Global Costs of Unaddressed Hearing Loss and Cost-effectiveness of Interventions*. World Health Organization. Retrieved from: http://apps.who.int/iris/bitstream/10665/254659/1/9789241512046-eng.pdf?ua=1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Christensen, Pontoppidan, Rossing, Anisetti, Bamiou, Spanoudakis, Murdin, Bibas, Kikidiks, Dimakopoulos, Giotis and Ecomomou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.