Big Data Analytics in Healthcare: Design and Implementation for a Hearing Aid Case Study

Jeppe H Christensen
Eriksholm Research Centre
Oticon A/S
Snekkersten, Denmark
jych@eriksholm.com

Michael K Petersen
Eriksholm Research Centre
Oticon A/S
Snekkersten, Denmark
mkpe@eriksholm.com

Niels H Pontoppidan
Eriksholm Research Centre
Oticon A/S
Snekkersten, Denmark
npon@eriksholm.com

Marco Cremonini
Department of Computer Science
University of Milan
Milan, Italy
marco.cremonini@unimi.it

***Abstract*—Modern hearing aids (HAs) are not simple passive sound enhancers, but rather complex devices that can log (via smart-phones) multivariate real-time data from the acoustic environment of a user. In the EVOTION project (www.h2020evotion.eu) such hearing aids are integrated with a Big Data analytics (BDA) platform to bring about ecologically valid evidence for policy-making within the hearing healthcare sector. Here, we present the background of the BDA platform and a concrete case study of how longitudinally sampled data from HAs can 1) support hypotheses about HA usage prognosis, and 2) bring new knowledge of how HAs are used across a typical day. In five participants, we found that the hourly HA usage was negatively associated with both the mean and the variance of the signal-to-noise ratio, and that increases in the daily total HA usage were associated with higher and more diverse sound levels.**

***Index Terms*—hearing aids, Big Data analytics, mixed models, multilevel clustered data, evidence-based public-health policies**

I. Introduction

Hearing loss (HL) affects approximately one-third of people over the age of 65 and 5% of the world's population [1]. In addition, disabling HL is associated with early cognitive decline in older adults [2], and if unaddressed, HL restricts social integration and educational and employment opportunities, hampers emotional well-being and poses an economic challenge at both the individual and national levels [3]. Moreover, HL prevalence is on the rise worldwide, primarily due to increased noise exposure and increase in the aging population [1]. Thus, HL poses a rising threat to the overall health and well-being of a large part of the general population and to the economic stability of healthcare sectors worldwide.

On the other hand, advances in generating health-related data by devices and sensors together with the development of technology and methodologies for processing large datasets (i.e. so-called Big Data) now permits the realization of data-driven solutions better adapted to aid individuals affected by HL [4]. Research and developments aiming to support traditional healthcare solutions with management strategies informed by Big Data now abounds [5], [6]. In this paper, we consider how modern hearing aids (HAs) combined with a platform for processing Big Data could improve the treatment of HL both on the level of the individual and of the general population.

Specifically, today, the leading management strategy for the majority of patients with HL is the provision of HAs. The use of HAs improves general health-related quality of life and hearing-specific quality of life associated with participation in daily activities and listening abilities [7]. However, HA users still face significant challenges, such as listening in noisy environments, poor sound quality, and difficulty to select among predefined programs and settings [8], which to a large part can be ascribed to ineffective or poorly fitted HAs. Ideally, HA fitting should adapt to the challenging and changing situations for individual HA users on a continuous basis rather than apply a "one-size fits all" strategy [9].

Given today's technologically advanced HAs, it is now possible to integrate them in a data-driven framework. This enables hearing care professionals to: 1) collect a rich set of information from patients with HL and their environment, 2) extract knowledge from data with preset analytics (e.g., HAs daily usage patterns), and 3) discover factors for low HAs usage and, for instance, correlations with the acoustic environment. As a result, HL patients using HAs could benefit from HAs adapting to the variable conditions of the environment and offering a more natural usage experience, which in turn ultimately increase the average usage of HAs and, thus, the benefit they produce to the quality of life. On top of that, novel data-insights regarding determining factors for successful HA uptake could help public-health policy-makers and healthcare administrators with making informed decisions.

The work presented in this paper is part of the ongoing research project 'EVidence-based management of hearing impairments: public health pOlicy making based on fusing Big Data analytics and simulatTON' (EVOTION; www.h2020evotion.eu), which aims to build the evidence base for the formulation of public health policies related to the prevention, early diagnosis, long-term treatment and rehabilitation of HL [10].

We first discuss some relevant works in the context of Big Data platforms for healthcare and of HAs. Next, we present a short summary of the Big Data analytics (BDA) platform developed in the EVOTION project. Lastly, we introduce a case study concerning five patients with HL and present

This work was partly supported by EU-funded project EVOTION (No. H2020-727521)

novel insights of HA usage behavior derived from modeling longitudinally sampled HA data.

II. Related Works

Evidence-based and data-driven public health policies have attracted remarkable attention in the last years both in the medical and in the data science communities, witnessed by related projects and publications [11]–[16]. What all these publication have in common is that they consider Big Data techniques as useful tools in translating personalized medicine initiatives into clinical practice. This is done by offering the opportunity to use analytic capabilities over highly heterogeneous data of different origin. For example, medical analyses could be improved by linking health-related data (e.g., medication list and family history) to lifestyle data (e.g., income, education, neighborhood, military service, diet habits, sport activity, entertainment) and to environmental data (e.g. polluted or noisy workplaces).

Moreover, a novel stream of data is increasingly available from sensors and medical devices, providing even more opportunities to study correlations between multiple factors related to healthcare [17], [18]. However, as also indicated in a report by the Institute of Medicine (IoM), several open problems still remains [19].

With respect to hearing healthcare and the problem of hearing loss, Big Data approaches are still underexplored. The most relevant work, up to now, was presented by Mellor, Stone, & Keane [4], [20] as "proof-of-concept" examples of how BDA methods could be used to gain data-insights from healthcare data, and they went on showing how clustering methods applied to a large data-set containing a number of variables concerning hearing aid users could be used to mine for interesting patterns. For example, the authors found that specific hearing aid settings (i.e. gain-reduction profiles) could be associated with specific distributions of sound pressure levels of the acoustic environment.

However, no other attempts at integrating a BDA framework with real-time data-feeds from hearing aids for holistic care of hearing loss patients have been made to date.

III. BDA Technology

Figure 1 shows the main components of the Big Data platform developed in EVOTION, which is based on the Apache Foundation ecosystem [21]. Here, we briefly describe the crucial components:

Hadoop - YARN: Hadoop is a tool for data-intensive distributed applications, based on the YARN programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS) [22]. Hadoop allows writing applications that rapidly process large amounts of data on large clusters of compute nodes. YARN permits a division of the input data-set into independent subsets that are processed in parallel (i.e. batch processing) [23].

HBase: A database engine built on Hadoop and modeled after Google's Big Table [24]. HBase is optimized for real-time data access to large tables with up to billions of rows. Among other features, it offers support for interfacing Hive.

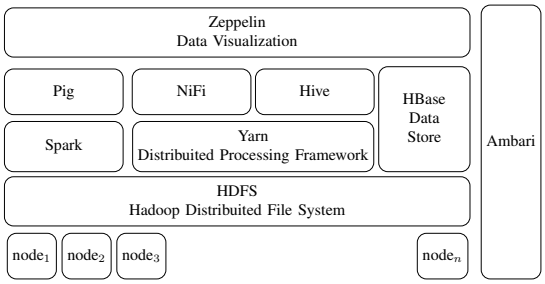

Fig. 1. Big Data platform technologies.

Hive: A data warehousing infrastructure, which runs on top of Hadoop. Hive provides a language called Hive QL to organize, aggregate, and run queries on data-sets [25]. Hive QL is similar to SQL, it uses a declarative programming model and results are described in one big query. HQL queries can be broken down by Hive to communicate to MapReduce jobs executed across a Hadoop cluster.

Spark: A general purpose cluster computing engine providing APIs to various programming languages such as Java, Python, or Scala [26]. Spark is specialized at making data analysis faster, it supports in-memory computing that enables it to query data much faster compared to disk-based engines such as Hadoop, and it also offers a general execution model that can optimize arbitrary operator graphs. Spark also offer several tools, such as machine learning tool Mllib [27], structured data processing, Spark SQL, graph processing tool Graph X, stream processing engine called Spark Streaming, and Shark for fast interactive question device.

Zeppelin: A web based and multipurpose notebook that enables interactive data analytics (Apache Zeppelin: https://zeppelin.apache.org). The notebook is the place for data ingestion, discovery, analytics, visualization and collaboration. Zeppelin supports many interpreters such as Apache Spark, Python, JDBC, Markdown and Shell.

IV. BDA Design

The BDA components within the EVOTION project is mainly focused on the execution of preset analytics over the data collected from the patients recruited for clinical trials (see section V). However, the overall goal of EVOTION is to support policy makers in the definition of public-health policies (PHPs) specified through the definition of data analytic workflows (DAWs) expressed in a proprietary language. A comprehensive presentation of policy specification, execution, and links to data analytics, as well as the whole EVOTION architecture is out of the scope of the present paper. Although, in order to illustrate how the BDA supports the execution of analytics based on data-feeds from HAs, we here shortly describe a DAW and the concepts of BDA interfacing.

A. Data analytic workflows

A DAW is an ordered sequence of *Data Analytic Tasks* (just *Tasks* in the following) of the following types:

- Data Processing Tasks: Including data preparation like data source selection for feature reduction, data cleaning, or data type transformation.
- Statistical Analysis Tasks: Performing statistical analysis on a data-set like ANOVA, Breusch-Pagan Test, etc.
- Data Mining Tasks: Exploiting supervised or unsupervised algorithms (e.g., Random Forest, K-means) for clustering and machine learning.

DAWs in EVOTION could be serialized (i.e., the output of one DAW inputs a consecutive one, for instance, one may select features and the following may compute a cluster) and they could mix automatic and human actions. That way, workflows could be defined as consecutive DAWs and human interventions, for example when policy makers or clinicians evaluate intermediate results and request further analysis. The logical of a DAW is implemented as a procedural workflow and translated into an executable form called *Executable DAW* (EDAW in the following). Thus, the core mechanism implemented by the BDA platform enables the execution of analytics based on two main subsystem catalogues (see details next section):

- Task Catalogue: A lists of Tasks for which an executable implementation is available. For instance, an entry in the Task Catalogue could be `Spark_ANOVA`, a specific implementation of an ANOVA.
- Workflow Catalogue: A lists of available EDAWs, each representing logical workflows of Tasks to be executed. An EDAW can be scheduled (e.g. run every month) and also trigger the execution of other EDAWs (for example based on new incoming data).

B. Task Catalogue

The Task Catalogue handles the list of available implemented analytic tasks that can be composed into EDAWs to produce analytic algorithms. An implemented analytic task includes the following attributes: Unique ID, Task name, name and version of required software libraries, programming language used for the development, path of source code/executable, details on dependencies (e.g., another task to be extended) if any, and a textual description of the algorithm.

In addition, the Task Catalogue offers APIs for managing the stored implementations (add, delete, modify).

C. Workflow Catalogue

The Workflow Catalogue handles the EDAWs and is composed of three sub-components: 1) the Catalogue repository of EDAWs, 2) the Workflow *Scheduler* scheduling the execution of EDAWs according to the specification given (i.e., periodic, upon request, or driven by data changes), and 3) the Workflow *Manager*, which is responsible for keeping track of the running EDAWs.

An EDAW includes the following attributes: Unique ID; EDAW name; the list of implemented analytic tasks; a list of parameters for each task and global parameters for the Workflow; the language used for the development (e.g., Scala, Java) and the Orchestrator adopted for the orchestration (e.g., Oozie, Scala); the path to the source/executable code or the path of the Orchestration metadescription; the Workflow Execution Type (on request, scheduled, driven by data change) and corresponding parameters; and a textual description of the Workflow purpose and characteristics.

D. BDA interfacing

The BDA platform is a core component of an extended framework whose main goal is to support stakeholders (i.e., policy makers, clinicians) during the definition of PHPs. Therefore, an interactive process between users and the BDA is taking place. For illustration purposes, we sketches the steps required to execute a certain Big Data analytics workflow (EDAW).

The actor is a policy maker or a clinician interacting through a graphical dashboard for the specification of the PHP instance to be processed (see Figure 2). This includes the selection of a given PHP model from a catalogue and the specification of actual parameters, data-types, and other conditions for processing the model. Thus, a policy model contains a DAW that specifies what operations are to be performed on the data. After the actor has specified all required information and defined the scheduling, the BDA framework transform at run-time the DAW expressed in declarative language into an EDAW, as a set of executable directives. To perform this transformation, executable analytics are retrieved from the Task Catalogue (see Figure 2). From the Workflow Catalogue, directives for the execution of the EDAW are retrieved and used by the scheduler subsystem. During execution of the EDAW, notifications are sent to the dashboard to let the user receive information about the ongoing processing.

In the next section, we present a case study demonstrating how consecutive operational steps, as those defined in a workflow, underpin a PHP model. In this case, the PHP model attempts to discover factors causing low HA usage.

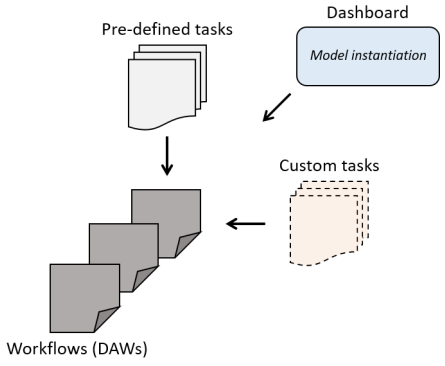

Fig. 2. Conceptualization of BDA interfaces. A stakeholder interacts with the BDA dashboard to define a policy model. The selected model (e.g. association of education and gender on HA use satisfaction) can consist of several predefined and custom tasks that make up a workflow. The workflows (DAWs) are executed according to a time-manager (e.g. once; every week; triggered by new data).

V. Hearing Aid Case Study

Several studies have found that demographic and self-reported factors predict the success of treatment with a HA. Typically, success of HA treatment equates HA uptake as usage in hours per day. [28] [29]. However, these studies are confounded by the fact that HA users tend to overestimate their HA usage as they report using their HAs more than what is measured by automatic HA logs [10]. It is thus important to use objective data logging of HA use for accurate information. In addition, Laplante-Levesque et al. [30] showed that patterns of HA use (i.e. "how" is the HA used) are at least as important in predicting HA outcomes as the duration of HA use (i.e. "how much" is the HA used). A fact that advocates the use of real-time longitudinal data-logging that will enable a more detailed tracking of usage patterns.

In EVOTION, participants with varying degrees of HL (N >1000) are supplied with "smart" HAs that measure and log the acoustic environment of a user each minute of use. This longitudinal real-time data feeds into a data-repository and enables a fusion with clinical and demographic data concerning each participant. Thus, analysis of data from EVOTION will 1) shed light on individualized HA usage and preferences; 2) enable a better understanding of factors influencing HA outcomes of a population (e.g. dynamics of the acoustic environment) and 3) support public-health policy-makers with ecologically valid factual evidence of HA user profiles and usage characteristics.

In the following sections, we show that modeling of data from EVOTION with linear mixed-models (LMMs) can help identifying external factors influencing HA usage.

A. Data

We use a data-set from a study at Eriksholm Research Centre (part of Oticon A/S, Denmark), which includes five participants with hearing loss that wear identical HAs as the ones provisioned in EVOTION. Thus, this data-set is representative of the type of data collected in EVOTION. In the study, participants were given a pair of Oticon EVOTION hearing aids (based on the Oticon Opn^{TM}) and a Bluetooth connected smart-phone [31]. During normal HA use, the smart-phone logged a data-vector of 20 acoustic parameters (recorded by the HA) and a time-stamp every minute. The data-vector includes five sound characteristics consisting of the momentary sound pressure level (*SPL*), signal-to-noise ratio (*SNR*), noise floor (*Nf*), Modulation Index (*MI*), and the Modulation Envelope (*ME*). Each characteristic were measured in four frequency bands: 0-1.3kHz; 1.3-4.1kHz; 4.1-10kHz; 0-10kHz, however, for simplicity, we only use the full-bandwidth (0-10kHz) data.

Given the longitudinal data, we can investigate both *how* HA users use their HAs and *how much* they use them. These two scenarios lead to the following aggregation of data:

- HA usage in minutes per hour.
- HA usage in hours per day.

HA usage in min/hour were derived by accumulating the time-stamps within each two-hour period from 0AM-11PM. E.g., the HA usage for 1PM is derived of time-stamps from 12:00PM to 1:59PM. In turn, HA usage for 2PM consists of time-stamps from 1PM to 2:59PM, ensuring one hour overlap for each data-point. The hourly HA usage were then averaged across all recorded days (the average amount of days was 40.2 with 10.5 days SD). Usage in hours/day were derived by accumulating the time-stamps across all hours within each day. Similarly, the acoustic parameters are represented by averages either within hours or days. In addition, we computed the variance of the *SPL* (SPL_{var}) and the *SNR* (SNR_{var}) across both hours and days. Thus, the data-set has multilevel grouping (e.g. hours, days, participant ID, hearing loss, gender) and each grouping contain a different number of samples (e.g. some participants were enrolled in the study for longer than others).

B. Hypothesis

One of the sub-goals of EVOTION is to support a PHP model that describes the impact of environmental factors to HA use and outcomes [10]. Such description will enable more individualized hearing healthcare and improve the general knowledge about the environmental difficulties that a HA user faces.

In the current study, we hypothesize that HA usage is predicted by a combination of parameters of the acoustic environment. Specifically, we hypothesize that HA users proactively use their HAs more if they are faced with loud acoustic environments (*SPL*) but poor signals (*SNR*). In addition, as previous studies have shown [32], we also hypothesize that the diversity of the sound environment (i.e. SPL_{var} and SNR_{var}) impacts usage times in a way that more diverse listening situations predicts higher usage.

C. Model approach

To test the significance of predicting variables we take a linear mixed models (LMMs) approach. LMMs have proven highly effective when dealing with complex, longitudinal data with multilevel clusters in which observations might be correlated. In LMMs, inter-observational correlations are dealt with by including random effects besides the independent predictive variables (here, acoustic characteristics), which are treated as fixed effects with regression coefficients [33].

Briefly, LMMs are an extension of simple linear models to allow for both fixed and random effects. Fixed effects are contributed to variables that exhibit constant slopes and intercepts with the response variable regardless of any hierarchical grouping - that is, fixed effects are considered within-participants effects. On the other hand, random effects allow the slope and/or intercept to vary between grouping factors (such as age or gender). For example, we might expect that HA usage within a day vary between participants due to unmeasurable factors, which would suggest including participant IDs as a random factor for the intercept (i.e. a between-participant effect). Moreover, we might believe that the strength of the association between *SPL* and HA usage vary between the participants, which in turn would suggest

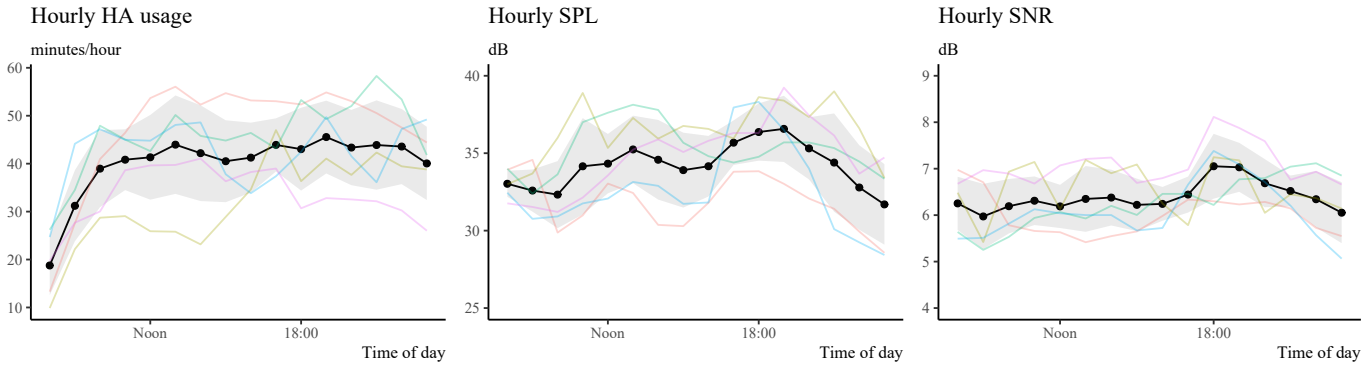

Fig. 3. Mean HA usage (left), sound pressure levels (middle), and signal-to-noise ratio (right). Each colored line represents data from each participant, and the black line represents the grand-average across all participants and days. Shaded area represents the 95% CI. Since only few logs were made at times between midnight and 7am these data-points were omitted.

modeling the fixed effect of *SPL* on HA usage with a random slope.

Using LMMs, we can model the dependent variable (in our case HA usage), y, on the following form:

$$y = X\beta + Z\gamma + \epsilon, \tag{1}$$

were y is a $N \times 1$ column vector of N observations; X is a $N \times p$ matrix of the p predictor variables (see first column in Tables II and III); β is a $p \times 1$ column vector of the fixed effects regression coefficients (see second column in Tables II and III); Z is the $N \times q$ design matrix for the q random effects; γ is a $q \times 1$ vector of random effects; and ϵ is the $N \times 1$ column vector of the residuals, i.e. the part of y not explained by our model. In our case study, we model the within-participant HA usage in minutes/hour with *SPL*, *SNR*, SPL_{var}, and SNR_{var} as fixed effects. In addition, we include random intercepts for each participant and hour, and random slopes for each of the fixed effects. We model HA usage in hours/day with the same fixed effects (now aggregated across all hours of a day) and allow random intercepts for each participant and day, and random slopes for each of the fixed effects. Modeling were done in the statistical programming language, *R*, using the *lme4* package [34].

D. Workflow implementation

The BDA Workflow (i.e. DAW) for this case study can be summarized with the following sub-flows, each consisting of several tasks:

Data selection: Relevant data are selected from the BDA platform data-repository and aggregated according to the needs of the hypotheses being tested (i.e. Data Processing Tasks). Here, two levels of aggregation are needed (as previously stated) and the selected data includes *SPL*, *SNR*, the time-stamps of each data-log, and any relevant grouping factor (i.e. participant ID, age).

Model selection: The predictor variables of the selected data are included as fixed effects and model selection identifies the most parsimonious parameterization of the random effects using LMMs (i.e. Statistical Analysis Tasks). Each combination of random effects (i.e. random intercepts and slopes) are compared wrt. Akaike's Information Criteria (*AIC*, see Table I). The model with the lowest *AIC* is chosen for further evaluation.

Model evaluation: The quality of the model fit is evaluated by inspecting the residuals and the predictive power of the final model is evaluated by the amount of explained variance. In addition, main effects of the regression coefficients for each predicting variable (here, acoustic parameters) are tested for significance by a MANOVA using Wald's chi-square tests (see Table II and III).

E. Results

Model selection (see section V-D) identified the optimal (most parsimonious) models by comparing each parameterization of the random effects with a *NULL* hypothesis that does not allow random variation in neither slopes nor intercepts.

Table I shows the AIC for each combination of random effects. Columns 1-3 display the results for models predicting HA usage between *hours* and column 4-6 show results for

TABLE I
COMPUTED AIC FOR EACH MODEL. THE MODELS aM_0 AND bM_0 CORRESPONDS TO THE *NULL* MODELS - THAT IS, THE LEAST COMPLEX MODELS THAT ONLY INCLUDE FIXED EFFECTS.

minutes/hour			hours/day		
Model	*df*	AIC	Model	*df*	AIC
aM_0	6	601.44	bM_0	6	1067.22
aM_1	7	584.01	bM_1	7	1019.15
aM_2	7	581.19	bM_2	7	1069.22
aM_3	8	554.88	bM_3	8	1021.15
aM_4	10	557.63	bM_4	10	1024.74
aM_5	10	558.85	bM_5	10	1021.72
aM_6	10	558.27	bM_6	10	1024.28
aM_7	10	558.63	bM_7	10	1020.07
aM_8	12	561.76	bM_8	12	1023.24
aM_9	12	562.40	bM_9	12	1022.91
aM_{10}	12	559.99	bM_{10}	12	1024.50
aM_{11}	12	561.00	bM_{11}	12	1025.41

models predicting HA usage between *days*. The lowest AIC is obtained by the models aM_3 and bM_1. In both cases, adding random effects vastly improved the model prediction (AIC approx. 50 lower than for *NULL* models aM_0 and bM_0). For predicting HA usage between hours, the best combination of random effects were to allow for random intercepts due to participant and hour of day. With this partitioning, 13.6% of the total variance between observations were explained purely by the fixed effects, and including the random effects increased the proportion of explained variance to 76.4%. For predicting HA usage between days, the proportion of explained variance were 6.5% (fixed effects) and 36.8% (full model), respectively, and the optimal model only included a random intercept due to participant. Figure 3 shows the mean HA usage, *SPL*, and *SNR* over time (from 7am to midnight). Each colored line corresponds to data from each participant, and the black line represents the grand-average across days and participants with the 95% CI indicated by the gray shaded area. Usage is fairly stable at around 40 minutes/hour from noon to midnight. However, both the *SPL* and *SNR* exhibit fluctuations over time: *SPL* peaks at noon and both *SPL* and *SNR* peaks in the early evening (6pm to 7pm). These grand average fluctuations are likely to indicate routine activities taking place at roughly the same time every day (such as lunch and dinner activities).

Table II presents the estimated regression coefficients of the optimal model for predicting HA usage within a day. The significant predictors were *SNR* and SNR_{var}. Thus, hours of low and highly varying *SNR* exhibit less HA usage (increases in *SNR* and SNR_{var} with 1 SD yields 2.04 and 2.8 minutes of less use per hour, respectively).

In Figure 4 we plot the distribution of total HA usage per day (left) together with scatter-plots of two possible predictive variables; *SPL* and SPL_{var} (middle and rightmost panel). In only a small proportion of days were the HA barely used, and the overall mean were 10.05 hours/day. The regression slopes of HA usage with *SPL* seems fairly equal among the participants (color-coded solid lines in Figure 4 middle) although the intercepts seem to vary by visual inspection (which were also confirmed by the model selection). In contrast, the association between HA usage and SPL_{var} (see Figure 4 right) indicates both varying slopes and intercepts. However, adding a random slope for SPL_{var} to model bM_1 did not lower the *AIC*. The regression coefficients (Table III) indicate significant and positive effects of *SPL* and SPL_{var} to daily HA usage. Thus, days with overall higher *SPL* and a more diverse sound environment are associated with increased HA usage. We did not find that including days as a random effect improved the *AIC* (bM_3 in Table III), which indicates that the participants did not generally increase their daily HA usage during the days of the study.

VI. Discussion

We have introduced a Big Data analytics (BDA) architecture for processing large amount of static and dynamic data concerning patients with hearing loss and their use of hearing aids. The BDA platform enables rapid analytics of data by implementing executable workflows that specifies the relevant mathematical and algorithmic operations to be performed.

Using a linear mixed-models approach, we have identified distinct patterns of hearing aid (HA) usage in five patients with hearing loss. On a daily basis, hours with less-than-normal HA usage is associated with a higher-than-normal signal-to-noise ratio and variance in the signal-to-noise ratio (SNR_{var}) of the sound environment. One speculative explanation is that moments of high SNR_{var}, which could be brought on by

TABLE II
Regression coefficients and significance for predicting HA usage within a day for each (scaled and centered) predictor.

	Dependent variable:
	HA usage (minutes/hour)
SPL	-1.373
	(0.999)
SNR	-2.037^{*}
	(1.194)
SPL_{var}	0.158
	(0.789)
SNR_{var}	-2.389^{**}
	(1.027)
Constant	38.031^{***}
	(3.086)
Observations	86
Log Likelihood	-296.265
Note:	*p<0.1; **p<0.05; ***p<0.01

TABLE III
Regression coefficients and significance for predicting HA usage between days for each (scaled and centered) predictor.

	Dependent variable:
	HA usage (hours/day)
SPL	0.624^{***}
	(0.240)
SNR	0.086
	(0.359)
SPL_{var}	0.825^{***}
	(0.256)
SNR_{var}	-0.325
	(0.365)
Constant	9.976^{***}
	(0.955)
Observations	198
Log Likelihood	-497.703
Note:	*p<0.1; **p<0.05; ***p<0.01

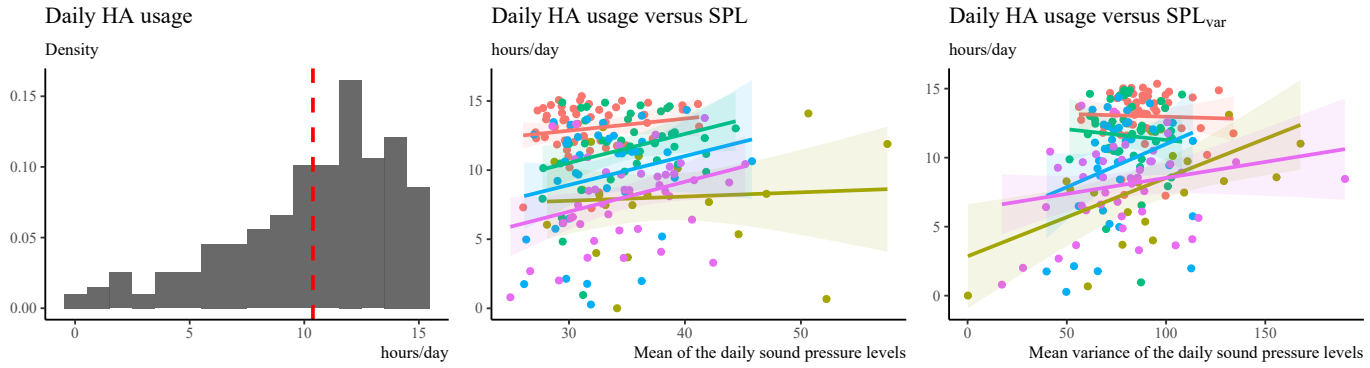

Fig. 4. Distribution of daily HA usage pooled from all participants (left), scatter-plot of daily HA usage and *SPL* (middle), and scatter-plot of daily HA usage and *SNR* (right). In the leftmost panel, grand average HA usage is represented by the red dashed line at around 10 hours/day. In the scatter-plots, regression lines with SDs are plotted for each (color-coded) participant.

changes in the contextual situation of a HA user (such as moving from a busy meeting to a quiet office), comprise fewer listening situations than moments of constant *SNR*. In addition, periods with an overall lower *SNR* might demand more HA use, which effectively increase the perceived *SNR*. On a monthly basis, days with more-than-normal HA usage is associated with higher-than-normal sound pressure levels (*SPL*) and more diverse sound environments. The diversity of the sound environment (here approximated by SPL_{var}) can be considered a proxy for life-style activity levels [35]. Thus, an overall louder and more active day seems to be associated with increased HA usage. This finding corroborates earlier studies, showing that when using the HAs more often, and reporting greater satisfaction, older adults indicated more diverse listening situations [32].

Our findings are based on recordings of HA usage and acoustic parameters in only five participants over the course of approximately 40 days. However, data in EVOTION will comprise recordings from more than one thousand participants with diverse demography and clinical backgrounds, and recordings of HA use will be acquired for up to one year. By projecting the presented modeling approach to EVOTION data via the BDA platform, we expect to generalize outcomes to the general population of people with hearing loss, and to identify predictors of HA usage that are specific to specific sub-populations according to either demographics or clinical conditions (such as educational level or severity of hearing loss). This will enable evidence-informed public-health policy making within the hearing healthcare sectors. For example, from the current results, a stakeholder could argue that the provision scheme of HAs should not only be guided by age and/or hearing loss but also by the acoustical environment a potential HA user faces.

Lastly, in a review study, Perez and Edmon [36] found that the level of reporting in studies investigating HA usage was inconsistent and of variable quality. Here, and in EVOTION, we overcome these problems by relying on actual data-logs instead of self-reported or accumulated HA usage statistics.

VII. Conclusions

In this work we have presented one of the first so far experimental application of analytics implemented over a Big Data platform and designed to process health data related to hearing impaired patients. The work is part of the ongoing EU project EVOTION aimed at supporting public-health policy makers with Big Data technology.

The work demonstrates the possibilities that a data-driven approach to healthcare could provide by considering a restricted set of user data ($N = 5$). Our findings are aligned with earlier studies and confirm the suitability of a linear mixed-models approach.

The same approach is currently being adopted over the much larger clinical trial ($N > 1000$) of the EVOTION project, where we expect to achieve results representative of the whole population of hearing impaired people.

Acknowledgment

We would like to thank Benjamin Johansen and Maciej Korzepa from Eriksholm Research Centre/Danish Technical University, Denmark, for sharing their data for our case study.

References

[1] World Health Organization, *Global costs of unaddressed hearing loss and cost-effectiveness of interventions.* 2017. OCLC: 975492198.

[2] B. O. Olusanya, K. J. Neumann, and J. E. Saunders, "The global burden of disabling hearing impairment: a call to action," *Bulletin of the World Health Organization*, vol. 92, pp. 367–373, May 2014.

[3] B. S. Wilson, D. L. Tucci, M. H. Merson, and G. M. O'Donoghue, "Global hearing health care: new findings and perspectives," *The Lancet*, vol. 390, pp. 2503–2515, Dec. 2017.

[4] J. Mellor, M. A. Stone, and J. Keane, "Application of Data Mining to a Large Hearing-Aid Manufacturer's Dataset to Identify Possible Benefits for Clinicians, Manufacturers, and Users," *Trends in Hearing*, vol. 22, p. 233121651877363, Jan. 2018.

[5] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[6] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, no. 1, p. 3, 2014.

[7] M. A. Ferguson, P. T. Kitterick, L. Y. Chong, M. Edmondson-Jones, F. Barker, and D. J. Hoare, "Hearing aids for mild to moderate hearing loss in adults," *Cochrane Database of Systematic Reviews*, Sept. 2017.

[8] A. McCormack and H. Fortnum, "Why do people fitted with hearing aids not wear them?," *International Journal of Audiology*, vol. 52, pp. 360–368, May 2013.

[9] M. A. Ferguson and H. Henshaw, "Auditory training can improve working memory, attention, and communication in adverse conditions for adults with hearing loss," *Frontiers in Psychology*, vol. 6, May 2015.

[10] G. Dritsakis, D. Kikidis, N. Koloutsou, L. Murdin, A. Bibas, K. Ploumidou, A. Laplante-Lévesque, N. H. Pontoppidan, and D.-E. Bamiou, "Clinical validation of a public health policy-making platform for hearing loss (EVOTION): protocol for a big data study," *BMJ Open*, vol. 8, p. e020978, Feb. 2018.

[11] R. C. Brownson, J. F. Chriqui, and K. A. Stamatakis, "Understanding evidence-based public health policy," *American Journal of Public Health*, vol. 99, no. 9, pp. 1576–1583, 2009.

[12] J. Hemerly, "Public policy considerations for data-driven innovation," *Computer*, vol. 46, pp. 25–31, June 2013.

[13] M. TB and D. AS, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.

[14] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, vol. 1, no. 1, pp. 51–59, 2013.

[15] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016.

[16] M. J. Khoury and J. P. Ioannidis, "Big data meets public health," *Science*, vol. 346, no. 6213, pp. 1054–1055, 2014.

[17] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[18] R. K. Pathinarupothi, P. Durga, and E. S. Rangan, "Iot based smart edge for global health: Remote monitoring with severity detection and alerts transmission," *IEEE Internet of Things Journal*, 2018.

[19] *The Global Use of Medicines: Outlook Through 2016*. MS Institute, 2016.

[20] J. C. Mellor, M. A. Stone, and J. Keane, "Application of Data Mining to "Big Data" Acquired in Audiology: Principles and Potential," *Trends in Hearing*, vol. 22, p. 233121651877681, Jan. 2018.

[21] M. M. Rathore, A. Paul, A. Ahmad, M. Anisetti, and G. Jeon, "Hadoop-based intelligent care system (hics): Analytical approach for big data in iot," *ACM Transactions on Internet Technology (TOIT)*, vol. 18, no. 1, p. 8, 2017.

[22] A. Holmes, *Hadoop in practice*. Manning Publications Co., 2012.

[23] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, *et al.*, "Apache hadoop yarn: Yet another resource negotiator," in *Proceedings of the 4th annual Symposium on Cloud Computing*, p. 5, ACM, 2013.

[24] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[25] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy, "Hive-a petabyte scale data warehouse using hadoop," in *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pp. 996–1005, IEEE, 2010.

[26] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, *et al.*, "Apache spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[27] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, *et al.*, "Mllib: Machine learning in apache spark," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.

[28] A. Laplante-Lévesque, L. Hickson, and L. Worrall, "Factors influencing rehabilitation decisions of adults with acquired hearing impairment," *International Journal of Audiology*, vol. 49, pp. 497–507, July 2010.

[29] B. Gopinath, J. Schneider, D. Hartley, E. Teber, C. M. McMahon, S. R. Leeder, and P. Mitchell, "Incidence and Predictors of Hearing Aid Use and Ownership Among Older Adults With Hearing Loss," *Annals of Epidemiology*, vol. 21, pp. 497–506, July 2011.

[30] A. Laplante-Lévesque, C. Nielsen, L. D. Jensen, and G. Naylor, "Patterns of Hearing Aid Usage Predict Hearing Aid Use Amount (Data Logged and Self-Reported) and Overreport," *Journal of the American Academy of Audiology*, vol. 25, pp. 187–198, Feb. 2014.

[31] B. Johansen, Y. P. R. Flet-Berliac, M. J. Korzepa, P. Sandholm, N. H. Pontoppidan, M. K. Petersen, and J. E. Larsen, "Hearables in Hearing Care: Discovering Usage Patterns Through IoT Devices," in *Universal Access in Human–Computer Interaction. Human and Technological Environments* (M. Antona and C. Stephanidis, eds.), vol. 10279, pp. 39–49, Cham: Springer International Publishing, 2017.

[32] B. Williger and F. R. Lang, "Hearing Aid Use in Everyday Life: Managing Contextual Variability," *Gerontology*, vol. 61, no. 2, pp. 158–165, 2015.

[33] J. Fox and J. Fox, *Applied regression analysis and generalized linear models*. Los Angeles: SAGE, third edition ed., 2016. OCLC: ocn894301740.

[34] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," *Journal of Statistical Software*, vol. 67, no. 1, 2015.

[35] Y.-H. Wu, E. Stangl, X. Zhang, and R. A. Bentler, "Construct Validity of the Ecological Momentary Assessment in Audiology Research," *Journal of the American Academy of Audiology*, vol. 26, pp. 872–884, Nov. 2015.

[36] E. Perez and B. A. Edmonds, "A Systematic Review of Studies Measuring and Reporting Hearing Aid Usage in Older Adults since 1999: A Descriptive Summary of Measurement Tools," *PLoS ONE*, vol. 7, p. e31831, Mar. 2012.