Assessing the Heterogeneity of Complaints Related to Tinnitus and Hyperacusis from an Unsupervised Machine Learning Approach: An Exploratory Study

Guillaume Palacios\textsuperscript{a}  Arnaud Noreña\textsuperscript{b}  Alain Londero\textsuperscript{c}

\textsuperscript{a}PainKILLAR, TELECOM ParisTech, Paris, France; \textsuperscript{b}Laboratoire de Neurosciences Sensorielles et Cognitives, UMR CNRS 7260, Aix-Marseille Université, Marseille, France; \textsuperscript{c}Hôpital Européen Georges-Pompidou, Service d’Oto-Rhino-Laryngologie et Chirurgie Cervico-Faciale, Paris, France

Keywords
Tinnitus · Hyperacusis · Text mining · Machine learning · Computational predictive medicine

Abstract

\textbf{Introduction}: Subjective tinnitus (ST) and hyperacusis (HA) are common auditory symptoms that may become incapacitating in a subgroup of patients who thereby seek medical advice. Both conditions can result from many different mechanisms, and as a consequence, patients may report a vast repertoire of associated symptoms and comorbidities that can reduce dramatically the quality of life and even lead to suicide attempts in the most severe cases. The present exploratory study is aimed at investigating patients’ symptoms and complaints using an in-depth statistical analysis of patients’ natural narratives in a real-life environment in which, thanks to the anonymization of contributions and the peer-to-peer interaction, it is supposed that the wording used is totally free of any self-limitation and self-censorship.

\textbf{Methods}: We applied a purely statistical, non-supervised machine learning approach to the analysis of patients’ verbatim exchanged on an Internet forum. After automated data extraction, the dataset has been preprocessed in order to make it suitable for statistical analysis. We used a variant of the Latent Dirichlet Allocation (LDA) algorithm to reveal clusters of symptoms and complaints of HA patients (topics). The probability of distribution of words within a topic uniquely characterizes it. The convergence of the log-likelihood of the LDA-model has been reached after 2,000 iterations. Several statistical parameters have been tested for topic modeling and word relevance factor within each topic.

\textbf{Results}: Despite a rather small dataset, this exploratory study demonstrates that patients’ free speeches available on the Internet constitute a valuable material for machine learning and statistical analysis aimed at categorizing ST/HA complaints. The LDA model with $K = 15$ topics seems to be the most relevant in terms of relative weights and correlations with the capability to individualizing subgroups of patients displaying specific characteristics. The study of the relevance factor may be useful to unveil weak but important signals that are present in patients’ narratives.

\textbf{Discussion/Conclusion}: We claim that the LDA non-supervised approach would permit to gain knowledge on the patterns of ST- and HA-related complaints and on patients’ centered domains of interest. The merits and limitations of the LDA algorithms are compared with other natural language processing methods and with more conventional methods of qualitative analysis of patients’ output. Future directions and research topics emerging from this innovative algorithmic analysis are proposed.

© 2020 The Author(s)
Published by S. Karger AG, Basel
Introduction

Subjective tinnitus (ST) is the most prevalent form of tinnitus (>95%) [McCormack et al., 2016]. It is defined as a non-hallucinatory auditory perception that is not produced by any actual acoustic source [Erlandsson and Dauman, 2013]. ST is often associated with another debilitating condition namely hyperacusis (HA), that is, intolerance to certain environmental sounds [Dauman and Bouscau-Faure, 2005; Tyler et al., 2014]. Indeed both ST and HA display a huge heterogeneity in their clinical presentation and may result from a variety of causes involving either peripheral auditory damage or central maladaptive plasticity [Nicolas-Puel et al., 2002; Pienkowski et al., 2014]. Moreover their pathophysiology is not fully understood, even if they most probably share common mechanisms, [Eggermont and Roberts, 2012] and thus far, no objective biomarkers of ST and HA are available that may be used to identify and categorize their countless clinically relevant subtypes [Haider et al., 2018].

Then, ST/HA assessment essentially relies upon an in-depth analysis of patients’ reported output [Landgrebe et al., 2012; Hall et al., 2016]. Indeed ST/HA intrusiveness or induced distress is commonly assessed by the means of Visual Analog Scales or validated multidimensional questionnaires [Meikle et al., 2007] such as the tinnitus handicap inventory [Newman et al., 1996], the tinnitus functional index [Meikle et al., 2012] or the HA questionnaire [Khalfa et al., 2002]. These closed questionnaires are currently considered the standard evaluation of ST/HA impairment even if the validity and the consistency of such multi-item assessment tools have been questioned [Fackrell et al., 2016]. However, beyond the questionnaires, another main source of information aimed at assessing ST/HA subtypes, ST/HA severity and associated symptoms is the patient’s clinical interview during which he/she has to answer open questions. Some data is available on ST/HA related complaints freely reported by patients in open-ended questionnaires. A recent comprehensive review has analyzed the answers to open questions (8 studies, 885 patients) compared to closed ones (78 studies, 15,496 patients) with regard to ST related complaints. It showed that 42 discrete domains could be defined spanning physical health, psychological health, quality of life and negative attributes of ST. Interestingly the authors of this meta-analysis stated that “this diversity was not captured by any individual study alone,” and that “there was a good convergence between complaints collected using open- and closed-format questions” with “general mood and perceptual attributes of tinnitus reported only using closed questions” [Hall et al., 2018]. Meaning that open questions, asked by a healthcare professional in a small sample of patients (hundreds at most), may miss some key clinical patterns of ST/HA. And that these questionnaires could not reflect the real-life heterogeneity in ST/HA clinical presentation and may not be able to individualize small subgroups of patients displaying very specific characteristics. Another intrinsic limitation and potential weakness of all these questionnaire-based assessments is that they somehow reflect the medical insight and not the patient-focused one [Haider et al., 2016]. Either because the patient has to answer closed questions without any possibility to add information that is relevant for him/her, or because his/her answers to open questions are analyzed, hierarchized and clustered from a biased medical perspective. Indeed, if subjects are instructed in qualitative studies to tell their own story in their own words, their wording is eventually interpreted during the analytic process aimed at understanding and categorizing patients’ verbatim [Tyler and Baker, 1983; Dauman et al., 2017; Colagrosso et al., 2019]. And so even when using elaborated models such as the Grounded Theory aimed at theorizing adaptation to a chronic condition from the data themselves [Pryce and Chilvers, 2018; Hallberg et al., 2005], comprehensive classifications such as the International Classification of Functioning, Disability and Health [Manchaiah et al., 2018], or psychologically driven tools such as the interpretative phenomenological analysis [Talewar et al., 2019; Smith and Osborn, 2008]. Eventually, despite all these efforts to better capture patients concerns, do we really know which ST/HA related complaints are the most relevant to patients since only the patient is in position to describe his/her condition and its impact on his/her life. Do the patients thoroughly describe all their symptoms during a clinical interview, in particular those they could consider “useless” or “too exotic” to be told, potentially because of the fear of being ridiculed? Which words do they use to characterize their symptoms in an ecological or everyday environment (assuming that the clinical environment can be intimidating and source of anxiety [Dauman et al., 2015])? Which problems do they complain about the most frequently when they speak freely and not when they answer closed questionnaires or orientated questions. These questions are far from being trivial.

Therefore, in order to be productive, it appears mandatory for healthcare providers to gain knowledge on consensual and patient-centered evaluations [Williamson et al., 2012; Black, 2013]. Obtaining an exhaustive and accurate insight on patients’ symptoms may pave the way
for a better understanding into the “structural properties” of ST/HA subgroups of patients, appropriate assessment and outcome instruments also fitting the patients’ needs [Hall et al., 2019], and ultimately the development of efficient consensual therapeutic strategies [Londero and Hall, 2017; Pryce et al., 2018]. This is why a systematic quantitative study of tinnitus patients’ verbatim would be meaningful and informative.

For such a purpose, the huge amount of information that is freely available on the Internet and social media platforms (Facebook, Twitter, Blogs…), could be of crucial interest. Indeed, the Internet allows people to learn about ST/HA, advocate for tinnitus awareness and research funding [Deshpande et al., 2018]. And mobile apps have been developed allowing an ecological momentary assessment aimed at collecting subjective measurements under real-life conditions [Schlee et al., 2016]. But for the patients, the Internet is of utmost importance when they are seeking support and connections with medical professionals and other ST/HA patients. During those interactions, patients share quite a great deal of information about their condition (how ST/HA is portrayed individually) and their feelings about it (what is the current perception of a specific ST/HA issue). This huge amount of ecological data is worth a deeper insight and analysis.

But it is beyond the capabilities of any human being (healthcare professional or scientist) to read through this huge amount of data and to organize it. This is why one should apply modern natural language processing (NLP) techniques powered by machine learning algorithms. These algorithms are particularly suited to the problem of classification and categorization of large quantity of textual data aimed at extracting a text pattern analysis from a set of documents written in a natural language. An exhaustive survey of machine learning and NLP techniques is beyond the scope of this article. For a general overview of NLP techniques, see Mannings [Manning and Schütze, 1999]. In short, there are two categories of machine learning algorithms: supervised and unsupervised. In the supervised approach, the algorithm learns from a set of pre-labelled data. Conversely, an unsupervised algorithm will unravel, by itself, the underlying patterns present in the data [Blei and McAuliffe, 2010].

The main goal of the present paper is to present a methodology of extraction of topics (or dimensions) that form the matrix of the speeches of ST/HA patients. Since the topics we seek to extract are not known a priori, neither the number of topics nor their relative importance, we have favoured a so-called unsupervised text categorization approach using a Latent Dirichlet Allocation (LDA) algorithm [Bhattacharya and Getoor, 2006]. One should note that other techniques such as latent semantic analysis (LSA) and recurrent neural networks (RNN) may be used for the same purpose [Landauer, 2006]. LSA is the oldest technique of text categorization [Dumais, 2005]. It is based on a Matrix Factorization technique called singular value decomposition used in many areas of science from image compressing to quantum mechanics. The idea underpinning LSA is to work in a reduced space of latent topics that should describe a large group of text in a simpler and more condensed way. Because of its maturity and its simplicity, LSA algorithms are widely used in NLP. Recently LSA was used in the field of psychiatry [Bedi et al., 2015]. In the latter, the authors used LSA and dimensionality reduction techniques to study the onset of psychosis in the verbatim of a population of patients presenting a high risk of developing schizophrenia. RNN have arisen as a promising tool for text analysis and speech recognition. In a recent paper, RNN were used to detect emotions in human speech opening the way to a huge field of research in psychology, psychiatry and clinician-patient interactions [Lee and Tashev, 2015]. For a review of RNN techniques applied to NLP, see [Goldberg, 2015]. These data mining and information retrieval techniques have been applied in pharmacology, medicine, and in social media data: Feldman et al., [2015] the authors use text mining methodology to extract adverse effects of drugs from medical forums. Another interesting study focused on performing sentiment analysis on several medical forums dedicated to hearing loss, where the authors categorized messages posted on the forums as positive, negative, and neutral [Ali et al., 2013]. Finally ontology-based text mining methods were also applied in an example of a study of Chronic Kidney Disease forums [Burckhardt and Padman, 2015].

The goal of this proof-of-concept study is to test the feasibility and interest of the LDA method. This method is aimed at extracting topics related to HA from freely posted comments on an Internet forum (TinnitusTalk.com). Future directions and research topics of this algorithmic analysis in the field of ST/HA are also discussed.

Material and Methods

Design of the Study

This paper presents a preliminary statistical analysis of freely posted comments on an internet forum (TinnitusTalk.com). After having contacted the founder of the Tinnitus Hub (administrator of TinnitusTalk.com) and gained verbal permission of doing so, we extracted all data from a publicly available source and anony-
After analyzing the textual data, it is mandatory to clean and to pre-process it. After having checked the database for consistency, we removed duplicate forum posts. In order to perform a statistical analysis of the terms appearing in the data we had to break all text messages into sentences. Then, we split each sentence into a sequence of words (or tokens). This step is called "tokenization." In order to avoid as much noise as possible in the data signal, we removed all English common stop-words such as conjunctions (if, then, because…), determinants (the, a), pronouns (he, she…), and common verbs (is, be, have, had, will…). These words appear frequently in any English text but add no useful information to the content or significance of the document. Eventually, we removed all punctuation marks and normalized the text to lowercase. The output of this pre-processing pipeline is a so-called Bag of Words (BOW) representation of the text corpus. In our case, the corpus consists of all the post messages recovered from the "Hyperacusis" forum. The BOW representation of the textual data is an ensemble of all the tokens found in the text corpus (after preprocessing) where the order of the words in the original sentences in the text has been forgotten.

### Statistical Analysis

**Term Frequency**

As a first step towards a basic descriptive statistical analysis of the data, each item of the BOW was ranked according to its frequency of appearance in the forum. We wrote a simple Python script that counts the number of appearances of each token in the text corpus. The results are shown in Figure 1 and commented in the Results section.

### Topic Modeling

Next, we examined the data from the HA forum on TinnitusTalk.com under the lens of a topic model. Topic modeling is a type of statistical learning aimed at unveiling the abstract topics that occur in a collection of documents. The LDA algorithm is the most well-known example of topic modeling. It is a generative probabilistic model that unveils the underlying topics that structure a corpus of texts [Blei et al., 2003] and measure their relative importance. Within the LDA framework, each document (forum post) is modeled as a statistical mixture of different topics. And these topics can be interpreted as a set of words each with a certain probability of occurrence. The words with the highest probability are the representatives of the topics computed by the LDA algorithm. Recently, LDA has been used in different fields such as the classification and generation of political speeches [Kassarnig, 2016].

We took the Java-implementation of the LDA algorithm from the open-source machine learning library Mallet (http://mallet.cs.umass.edu/). The topic modeling of the TinnitusTalk.com data was conveyed using a set of high level R functions that performed...
the following steps: (1) read the data from the Mongo Database (JSON format), (2) put the data into a data frame and (3) run a parallel LDA topic modeling on the whole BOW corpus of “Hyperacusis” forum posts calling the Java LDA model from Mallet. To make the interface between the R program and the MALLET models (in Java), we used the rJava and Mallet-R packages.

The output of the whole process is a probabilistic model giving K topics, each topic described as a distribution of words. The words and association of words are characteristic of a topic. The set of topics discovered by the algorithm are to be considered a real-life picture of HA patients’ concerns on an Internet forum (i.e., symptoms and distress) as we shall see in the Results section. We reached convergence of the log-likelihood of the LDA-model after 2,000 iterations. The number of topics (K) of the model is a free parameter of the model that needs to be fine-tuned. We have tested a set of 3 candidates for the number (K) of topics representing our corpus. K = 10, 15, and 20. The LDA model with K = 15 topics produced the most relevant topics in terms of relative weights and correlations from a patients’ perspective. A visualization of the distributions of topics for the TinnitusTalk.com data is shown in Figure 2 and commented in length in the Results section.

Note that since the redaction of the present paper, a new study of the impact of tinnitus in social media was recently published. In this article, the authors conduct a cluster analysis of a textual corpus (News, Facebook pages etc.) using the Reinert hierarchical classification method [Manchaiah et al., 2018]. Interestingly, the authors claim that topics derived from the Reinert method (i.e., a divisive hierarchical clustering) are more accurate than the ones computed by LDA. Further research is needed to test this claim on our data and compare the results obtained by both algorithms.
Results

The word count of the 50 most frequent words occurring in the text corpus is displayed in Figure 1.

In this thread dedicated to HA the 5 words most frequently used are: ear ($n = 1,504$), HA ($n = 1,358$), sound ($n = 1,172$), tinnitus ($n = 1,087$), and pain ($n = 965$). Figure 1 also shows some less frequent but highly relevant words with a psychological valence such as anxiety ($n = 350$), sensitivity ($n = 191$), feeling ($n = 190$), sleep ($n = 179$). The main results of running the LDA model on the BOW dataset on the “Hyperacusis” thread have been summarized in Figure 2 which shows a graphical representation of the 15 topics unravelled by the LDA algorithm [Sievert and Shirley, 2014]. Each topic is represented by a circle (Fig. 2a) whose area is proportional to the importance of the topic in the corpus of patients’ posts and whose distance with the other circles is proportional to their inter-correlation. For the sake of clarity, we have reordered the topics by importance, with Topic 1 being the most important one while Topic 15 is the least important.

Each topic is described by a set of words, and each word is assigned a probability of occurrence in the topic (e.g., frequency of words used in Topic 3: Fig. 2a). The probability of distribution of words within a topic uniquely characterizes it. The most important topics (Topic 1 and 2) are pretty general and have keywords such as tinnitus and hearing. This is often the case in BOW analysis and LDA in particular, where a trade-off between the scarcity of data and precision of the categorization has to be made. As a consequence, the main topics generally produce low valuable information. Conversely less frequent topics are more precise and thus may be able to unveil some very interesting features in the patients’ discussions. Table 2 displays 15 of such topics dealing with specific patients’ concerns when discussing on the forum.

Because of the small number of posts analyzed in this exploratory study (3,785 posts from 200 contributors),
the results presented here are, by nature, very preliminary and speculative. The study of a larger amount of data would be necessary to draw really informative and firm conclusions. Therefore, we did not conduct an in-depth interpretation of the output of the LDA algorithm. Nonetheless some interesting examples of topics extracted by the LDA algorithm can be given. Topic 3 is a general topic about HA. This topic counts for roughly 10% over all forum posts analyzed in our experiment. Indeed, the most frequent words for this topic are ear, pain, tinnitus and HA. But we can find in this topic other interesting words such as burning/cold, fullness/pressure or neck/head that could represent more subtle variants of patients’ symptoms (Fig. 2). Topic 5 with words like ear(s), loud, (ear)plugs, noise, dB, wear(ing), protection is related to hearing protection, which is an important everyday concern in the HA population and chief issue during HA patients’ counselling. Topic 6, which is closely linked to Topic 5 and that which accounts for 5.2% of the analyzed tokens, seems to be related to the Tensor Tympani Syndrome with top words such as ear, ttts, tensor, tympani, sound or muscle. This suggests that the symptoms and concerns of this subgroup of subjects do not strictly overlap those of the overall HA population. Topic 7 deals with hearing tests, with keywords such as ENT, hearing loss or frequencies meaning that hearing damage and hearing loss are key concerns that should be considered during patient counselling, at least for this subgroup of HA subjects. Topic 8 with words such as sleep, room, quiet, bed is a less frequent signal about sleeping problems highlighting the relevance of this symptom in a subset of patients. Topic 11 deals with inner ear and vestibular disorders and their diagnosis. We find here words such as MRI, vertigo, neurologist, Meniere and dizziness. This topic is rather small in comparison to others and accounts for only 3.8% of the conversations but defines quite probably another specific subgroup of HA patients displaying both auditory and vestibular impairments. Topic 13 with words such as stress, system, anxiety, life, work, limbic, stressed, brain, depression, people emphasizes the relevance of psychological co-factors associated with HA intrusiveness as described by the patients themselves. Finally, Topic 14 is a broad topic centered around quality-of-life issues such as music, nature or love suggesting that psycho-social negative consequences of HA go far beyond the auditory domain. Note that Topic 13 and 14 are very closely linked (short distance between topics) suggesting that psychological concerns may be crucial when considering the impairment of the quality of life among HA patients.

So far, the results presented in Figure 2 show the terms “w” displayed in each topic t ordered by decreasing probability of a term “w” appearing in topic t, p (w|t). It is possible to look deeper into a Topic by introducing the rele-
vance factor [Sievert and Shirley, 2014] of a term $w$ in a topic $t$ defined as:

$$\text{Relevance (term } w \mid \text{ topic } t) = \lambda \cdot p(w \mid t) + (1 - \lambda) \cdot p(w)$$

where $\lambda$ is a parameter taking values in $(0, 1)$ and $p(w)$ is the probability of finding the term $w$ in the whole corpus. We see from Eq. (1) that the case $\lambda = 1$ corresponds to the regular ordering of the terms that we have discussed so far with a relevance (or probability) of each term in the topic being:

$$\text{Relevance (term } w \mid \text{ topic } t) = p(w \mid t)$$

whereas for $\lambda = 0$

$$\text{Relevance (term } w \mid \text{ topic } t) = p(w)$$

which puts more weight on terms that appear only in a given topic and less weight on frequent terms. Equipped with this new tool, we can now investigate what happens when tuning the parameter from 1 to 0, thus giving more weight to terms that appear very specifically in 1 topic with respect to the overall corpus [Chuang et al., 2012]. Figures 3–7 show the 30 most relevant terms for topic 11 ranked by decreasing order of relevance (from top to bottom) for various values of $\lambda$. Figure 3 corresponds to $\lambda = 1$, Figure 4 to $\lambda = 0.6$, Figure 5 to $\lambda = 0.4$, Figure 6 to $\lambda = 0.2$ and finally Figure 7 corresponds to $\lambda = 0$.

Interestingly, we see that decreasing $\lambda$ brings to the surface new terms that are very specific to the Topic and highlight technical and medical terms. For example, the terms *canal*, *semicircular*, *Tullio*, *perilymph*, *fistula(s)* emerge for Topic 11 for a value $\lambda = 0$ (Fig. 7). This shows clearly that the LDA algorithm is able to detect relatively rare medical subtypes such as superior canal dehiscence (Minor’s Syndrome) or perilymph fistulae by analyzing the verbatim of subjects who discuss of HA associated with a vestibular impairment. In other words, this study of the relevance allows unveiling weak signals in the distribution of words. Indeed those weak signals are potentially very interesting for the clinician, as they capture less common though actual features in the discussion among patients.

**Discussion**

Even if NLP can be applied to any kind of textual data (i.e., clinical reports, patients’ testimonies, speech to text output) this preliminary study demonstrates that patients’ free speeches available on the Internet constitute a valuable dataset for machine learning statistical analysis aimed at categorizing ST/HA complaints. Indeed, by collecting all the posts exchanged among patients of the HA thread on TinnitusTalk.com, we have been able to perform a frequency analysis of these textual data. We found that the top 10 keywords associated with the HA thread are ear, HA, sound, tinnitus, pain, time, loud, noise, hearing and people. Those terms are in accordance with the clinical knowledge about HA highlighting the links between HA and hearing impairment, ST, chronic pain and noise exposure. Nonetheless, less frequent words stress other important clinical issues such as anxiety, sleep problems or hearing protection. Even more subtle domains of interest emerge from this frequency-based ranking of words: for example, the time-course of the condition (time, years, months, weeks, long), or the emotional context linked to it (feel, feelings, hope, bad, good). Our results are in line with what has already been shown by the use of open-ended problem questionnaires. Tyler and Baker [1983] have shown that ST is associated with negative effects on lifestyle (93%), emotional difficulties (70%), negative consequences on general health (56%) and hearing impairment (53%). Insomnia, depression, annoyance, and insecurity being the most commonly cited associated symptoms and concerns. With a similar methodology, Sanchez and Stephens [1997] have highlighted the global consequences of ST perception inducing a large variety of symptoms and concerns that were categorized into psychological (30.1%), hearing (23.5%), health (20.7%), sleep (14.6%), and situational (11.1%) difficulties.

Beyond the simple frequentist point of view, we also analyzed the HA forum data using a topic modeling approach (LDA). Even if the more important Topics 1 and 2 are pretty general, less frequent ones seem to be related to definite subgroups of patients and/or conditions such as the tensor tympani syndrome (Topic 6), Menière’s disease (Topic 11) or comorbid psychological factors (Topic 13). Apparently the LDA algorithm has the ability to categorize the patients’ complaints into a discrete number of clusters that may be representative of homogeneous subgroups of subjects/patients. Even if this has not been performed in this study for ethical reasons (respect of patients’ anonymization) a further step, would be to analyze statistically the contribution of each individual subject to each topic. And then, using an appropriate measure of distance, we would be able to compute clusters of similar patients instead of discussion topics. This could potentially lead to a totally automated method enabling to allocate each individual subject to a specific homogeneous
subgroup of individuals displaying the same medical condition. This could be of crucial interest for clinical research [Landgrebe et al., 2012].

Despite these results, one should question the real advantages of clustering raw data (i.e., unstructured patients’ posts on a forum) through a purely statistical algorithm when compared to already existing tools aimed at clustering patient information with open-ended or closed questionnaires. A lot of efforts have been devoted to the analysis and the characterization of the dimen-

Fig. 3. Diagram of the top-30 most relevant terms for topic 11 identified to discussions on neurological issues for a relevance parameter of $\lambda = 1$. Here, according to Eq. (1), relevance (term $w$ | topic $t$) = $p(w | t)$, that is, the terms in the topic are simply ordered with respect to the probability distribution of words inside the topic. The most probable a word is to belong to the topic, according to the machine-learned model, the higher it is in the relevance diagram.

1. Saliency(term $w$) = frequency($w$) * $\sum \frac{p(t | w) \times \log(p(t | w)/p(t))}{t}$ for topics $t$; see Chuang et al (2012)
2. Relevance(term $w$ | topic $t$) = $\lambda \times p(w | t) + (1 - \lambda) \times p(w | t)/p(w)$; see Sievert and Shirley (2014)
sions underlying the complaints of patients with tinnitus and HA by the means of standardized questionnaires, survey or medical literature mining [Haider et al., 2016]. But the limit of studies based on questionnaires is two-fold. First, studies relying on validated questionnaires look at the data through a pre-established reading grid, and so the questions asked are based on the analysis of patients wording [Newman et al., 1996]. Second, surveys are limited to a relatively small number of patients (hundreds at most). By contrast, the algorithmic method pro-

---

**Fig. 4.** Diagram of the top-30 most relevant terms for topic 11 identified to discussions on neurological issues for a relevance parameter of $\lambda = 0.6$. By changing $\lambda$ from 1 to 0.6, we see a clear reordering of the most relevant words. In particular, the terms neurologist and Meniere now appear, respectively, in second and third rank. This is due to the fact that these 2 terms have a very high ratio of overall term frequency versus term frequency in the topic.
posed in our paper is based on unsupervised statistical learning. Unsupervised means the algorithm is agnostic of the data is fed and very little information is known a priori about the data. Statistical learning means the more data we feed to our algorithms, the better the classification and characterization of the dimensions in patients’ speeches. Our claim is that this statistical method applied to a very large dataset would be able to reveal topics matching specific subgroups of subjects or, at least, specific concerns or dimensions linked to ST/HA intrusiveness. However, both qualitative studies and statistical methods have their own merits. Indeed, qualitative studies based on questionnaire analysis rely on few data but require a lot of medical, psychological and scientific

Fig. 5. Diagram of the Top-30 most relevant terms for topic 11 identified to discussions on neurological issues for a relevance parameter of $\lambda = 0.4$. 

1. Saliency(term $w$) = frequency($w$) * $\sum_t p(t | w)$ * $\log(p(t | w)/p(t))$ for topics $t$; see Chuang et al (2012) 
2. Relevance(term $w$ l topic $t$) = $\lambda * p(w | t) + (1 - \lambda) * p(w)/p(w)$; see Sievert and Shirley (2014)
knowledge to thoroughly interpret the data. On the other hand, statistical methods such as the one presented here assume very little knowledge about the underlying medical background but extract information from a huge volume of data (thousands of patients interacting on forums and social media daily). Even if this is beyond the scope of this paper, further research is warranted to compare the relevance and usefulness of both strategies, that is, the standard one with “few data but lots of knowledge” versus the non-supervised algorithmic one with “no a-priori knowledge but lots of data.”

**Fig. 6.** Diagram of the Top-30 most relevant terms for topic 11 identified to discussions on neurological issues for a relevance parameter of \( \lambda = 0.2 \).
The statistical machine learning approach presented has four main strengths. First, data from patients’ forums such as TinnitusTalk.com constitute a large set of textual data from a wide audience. This allows for a large-scale statistical analysis of the patients’ speeches using NLP techniques and data-mining algorithms. Second, since those methods are based on machine learning techniques, they are not hampered by any medical bias. In other words, those algorithms consider every word and every patient’s opinion from a purely statistical point of view, without requiring any medical knowledge. The LDA algorithmic method has the key benefit of having...

**Fig. 7.** Diagram of the Top-30 most relevant terms for topic 11 identified to discussions on neurological issues for a relevance parameter of $\lambda = 0$. 

1. $\text{Saliency}(\text{term } w) = \text{frequency}(w) \times \sum_{t} p(t \mid l w) \times \log(p(t \mid l w)/p(t))$ for topics $t$; see Chuang et al (2012).
2. $\text{Relevance}(\text{term } w \mid \text{topic } t) = \lambda \times p(w \mid t) + (1 - \lambda) \times p(w \mid l t)/p(w)$; see Sievert and Shirley (2014).
even if forums are moderated, the dataset may not be representative of the overall population of ST/HA patients. On the other hand, the information shared among people inside of patients’ forums is not fully validated and might not be all medically accurate. It also contains a lot of disturbances in terms of noise. The best way to overcome this problem is to boost the signal/noise ratio by increasing the volume of the data to be treated. The fourth limitation is that we present a statistical study of the actors on the forums as a “frozen cohort.” In particular, we do not take into account the interactions among actors nor the intra-subject variability during the time course of the ST/HA, which have been argued to be important.

Conclusion and Perspectives

In this proof-of-concept study, we have tested the feasibility of a novel statistical method aimed at analyzing the speeches of ST/HA patients in real-life conditions through the collection of open web data. It is an ab-initio statistical approach that supposes no a priori knowledge about the complaints of ST/HA patients. Despite the small amount of data analyzed in this exploratory study, the LDA algorithm and related statistical techniques (term frequency, topic modeling, relevance factor) indeed display the capability to unveil both general and very specific ST/HA patients’ concerns, to cluster the speeches into discrete informative topics and to unravel uncommon medical conditions from natural language input. We acknowledge the fact that this purely statistical approach should be compared by further research to other already existing phenomenological methods of analyzing patients’ verbatim via open-ended questionnaires. Moreover, this paper opens up to various extensions of the method presented here. In particular, in future works, we plan on leveraging the power of open and social data and to include patients verbatim from various sources: blogs, forums, and social media data (Twitter) as well as patients’ expressions from various countries and languages. The latter would be particularly interesting to compare the influence of language and culture on the patients’ perception of ST/HA. This could be achieved by using neural networks methods to address the issues of synonymy and lexical field and translations of concepts between the texts in English and those written in other languages.

So far, our method proposes to take a snapshot of the wording of the topics discussed among tinnitus and HA patients in free speech blogs and forums. But it is a static
picture and we know that the variability of sensitivity to tinnitus or to a particular stimulus is high. Developing new algorithmic methods that would follow the evolution of patient’s concerns and topics of discussion over time is the next challenge [Blei and Lafferty, 2006; Amr and Xing, 2012; Brugggermann et al., 2016]. Finally, the ultimate goal of the presented method is to lay the ground for a thorough clinical study with open-ended interview of patients. We would record their testimony using speech recognition techniques and transform it into analyzable text data. This would allow a follow-up of a cohort before and after therapy, thus offering a novel, patient-centered and totally unbiased way of evaluating the outcome of therapeutic interventions.

References


Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013 Jan;346:f167.


Londero A, Hall DA. Call for an evidence-based consensus on outcome reporting in tinnitus intervention studies. Front Med (Lausanne). 2017 Apr;4:42.


Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality; 2013.


