Subject-based association networks for clustering analysis and phenotyping of pediatric sleep apnea

J. Gómez Pilar[1,2], D. Ferreira-Santos[3,4], P. Pereira-Rodrigues[3,4], D. Gozal[5], R. Hornero Sánchez[1,2], G.C. Gutiérrez Tobal[1,2]

[1] Grupo de Ingeniería Biomédica, Universidad de Valladolid, Valladolid, España

[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), España

[3] MEDCIDS-FMUP – Community Medicine, Faculty of Medicine of the University of Porto, Porto, Portugal

[4] CINTESIS – Center for Health Technology and Services Research, Porto, Portugal

[5] Department of Child Health, Child Health Research Institute, The University of Missouri School of Medicine, Columbia, MO, United States

Abstract

Pediatric patients suffering from obstructive sleep apnea (OSA) show different responses to treatment, being the reasons for this heterogeneity not completely understood. The phenotypic characterization of this disease could shed new light to personalized treatments by identifying predictive factors of the patient's evolution. In this context, the objective of this study is to identify subgroups of patients, as well as the most relevant variables to differentiate them. This was conducted using a modularity analysis applied to a subject-based association network. The Childhood Adenotonsillectomy Trial (CHAT)-baseline dataset participated in the study. It comprises 464 pediatric patients ranging 5-10 years of age that includes a variety of sociodemographic and clinical data. A novel variation of the association network generation methodology allowed to obtain a network where each node represents a subject, and the links represent the associations between its characteristics. After applying a bootstrap-based approach and network modularity analysis following Blondel's algorithm, 3 subgroups of patients with similar intragroup phenotypic characteristics, but significantly different intergroup features, were identified. The main defining characteristics of each group were age, sex, presence of hypertension, and weight-related characteristics. Together, these findings highlight the need for acquiring clinical and sociodemographic data beyond those purely derived from polysomnography to provide a tailored diagnosis and treatment, while showing the ability of the new proposed methodology to find modules of subjects with well-differentiated phenotypical characteristics.

1. Introduction

Pediatric obstructive sleep apnea (OSA) is a breathing disorder with high prevalence, affecting between 1 to 5% of the children worldwide [1]. It is characterized by the recurrent repetition of episodes of complete cessation (apneas) and decreases (hypopneas) of breathing during sleep that lead to hypoxemia, changes in intrathoracic pressure, surges in sympathetic activity, and changes in the heart rate regulation. Together, this get rise to the specific disruption of normal oxygenation and sleep patterns that depends on each patient [1].

OSA is diagnosed using standardized protocols based on overnight polysomnography (PSG) [2]. Then, adenotonsillectomy is the first line of treatment for the affected children. Nonetheless, the effectiveness of adenotonsillectomy is estimated to be no greater than 79% of cases [3]. Those children who are not candidates for adenotonsillectomy typically require continuous positive airway pressure (CPAP) therapy. Adherence and tolerance to treatment is highly heterogeneous though, which suggests the need for tailored treatments. There is therefore a need to develop personalized clinical protocols based on precision medicine for an objective characterization of the disease leading to more accurate diagnosis and treatments.

In adult OSA, factors predicting treatment adherence have been scarcely studied, obtaining disparate results among the few works oriented to this problem. These findings varied among studies, with the most consistent ones being the severity of OSA based on the initial PSG parameters and the daytime sleepiness [4]. On the contrary, illness severity does not appear to be a predictive factor for adherence in children [5]. Although it has been suggested that OSA treatment maximize the reversibility of the adverse effects in the affected children [6], potential adherence predictors have remained even more unexplored to date. Therefore, the search of subgroups of pediatric OSA population with differentiating characteristics could help provide patient-oriented diagnosis and treatment.

In this context, the objective of this study is to identify subgroups of pediatric OSA patients, as well as the most relevant variables of each subgroup. Some previous studies in adults successfully used an automatic analysis based on a *k*-means algorithm extension to determine 3 adult clusters/phenotypes [7]. Alternatively in this work, we propose a novel technique via modularity analysis applied to associations networks. Particularly, our proposal is the use of a variation of this technique to identify subgroups of patients with characteristics that could have been hidden by other conventional clustering techniques. This family of techniques, widely known in fields such as genetics [8], [9], has not been applied in the context of pediatric OSA, except for a single previous study in which it showed enormous potential to differentiate electroencephalographic patterns [6].

Based on the previous evidence shown in studies involving adults [4], [7], our starting hypothesis states that it is possible to find groups of pediatric OSA patients with well-differentiated phenotypic characteristics.

2. Materials and methods

2.1. Participants

In this study, we used the multicenter Childhood Adenotonsillectomy Trial (CHAT)-baseline dataset [3], after obtaining proper approval (www.sleepdata.org). This dataset is comprised of 464 pediatric patients ranging 5-10 years of age. All these patients met the criteria for being considered for adenotonsillectomy. PSG was conducted and apnea/hypopnea events were scored according to the American Academy of Sleep Medicine 2007 guidelines [2]. The OAHI was defined as the number of all obstructive apneas and hypopneas. In addition to PSG data, clinical, and sociodemographic variables were also acquired. Among all the information, a total of 26 variables were used to try to identify the phenotyping subgroups. These were selected to get the set of variables with the highest coincidence degree with previous studies in adults among all the variables included in the CHAT database [7].

2.2. Association networks

The association network analysis is a technique used in different fields to assess the relationship between categorical variables [6]. In this approach, the nodes in the network represent the variables under study, while the values of the connection between the nodes are an index of the association between each pair of variables.

We here propose a modification of this approach in which the nodes represent the subjects under study (unlike the usual technique where each node represents a variable). Each subject is then characterized by a vector of categorized variables (26 in this case). Therefore, the values of each link between nodes are computed as the *X*-square value (appropriate for categorical data) between the vector of a given pair of subjects. Thereby, subjects with higher degree of association between them will obtain a high value in their connection.

In order to remove the links with a residual association between the respective nodes, we only maintain those links with a statistical degree of association between them ($p < 0.05$). This results in a semi-weighted network, i.e., the values of non-significant links are set to zero, while the other links are set to the *X*-square value of a given pair of nodes.

2.3. Bootstrap procedure

Aiming at increasing the robustness of the results, a previously validated bootstrap procedure was applied [6], [9]. We first randomly selected a subgroup of 26 variables with possible repetition. Then, the association network, i.e., association matrix, was estimated following the previously described computations between nodes (subjects). This procedure was repeated 1000 times, selecting the mean value matrix for subsequent analysis among all those obtained. This methodology provides more robust networks in the sense of higher reproducible results, while allowing the evaluation of the stability of the network. Thus, if the parameters obtained in the generated networks have a high variance, it implies that the network is not very stable and, therefore, reaches less generalizable results [9], [10].

2.4. Network visualization

The resulting network were represented using *Gephy* (version 0.9.6) software. It allows to represent the network using ForceAtlas2 approach [11]. ForceAtlas2 is based on visualizing the associations between nodes (subjects) by mirroring physical attraction and repulsion forces. Considering both the distance and the node degree (sum of the *X*-square values reaching a node) of the connected nodes, this method turns structural proximities into visual proximities. In this way, subjects with a higher level of association between them are more likely to appear closer in the network.

2.5. Cluster analysis

Cluster analysis aimed at finding similar groups of subjects, where "similarity" means a global measure over the full set of characteristics. In this study, we used an unsupervised learning algorithm for clustering the data in the association network, meaning in this context that the algorithm neither has prior information about the number of clusters that exist before running the model nor does it make assumptions about relationships within the data. Particularly, we used Blondel's modularity [12]. This is a heuristic method based on modularity optimization. It is shown to outperform all other known community detection methods in terms of computation time, being important in large networks. The algorithm provides a modularity value that indexes the modular characteristic of the network and, more importantly, it gives a label for each node indicating the cluster to which said node belongs.

3. Results

The generated network model is shown in Figure 1. The color of the nodes (patients) indicates each of the 3 groups automatically identified by Blondel's algorithm. On the other hand, the size of each node represents the average level of association with the rest of the nodes in the network. In addition, after applying the ForceAtlas2 algorithm, the nodes are placed so that those with the highest level of association between their parameters will appear spatially closer. As both the size of the nodes and the position of them depends on the level of association, a relationship between them can be observed. Thus, the largest nodes tend to appear centered in the network, while the smallest nodes are located on the periphery of the network.

Among the 26 variables used, Table 1 shows the 8 with the highest differences among the 3 modules automatically defined. As observed, these modules showed well differentiated values among these variables. First, module A (green) consists of 136 subjects, all of them males. On the contrary, module B (orange) is composed of 123 female patients. Module C (purple) include both male and female patients with similar percentages (41.1% and 58.9%, respectively). Age also proved to be one of the most differentiating characteristics among modules. Interestingly, module C had a significantly higher mean

	Module A (green)	Module B (orange)	Module C (purple)	X-square test (X | p)
Subjects (#)	136	126	202	-
Sex (males% | females%)	100 | 0	0 | 100	41.1 | 58.9	267.8 | <0.001
Age (% between 5-6 years | 7-8 years | 9-10 years)	77.2 | 18.4 | 4.4	68.3 | 26.2 | 5.6	34.2 | 43.6 | 22.2	76.8 | <0.001
OAHI (% < 1 event/hour | > 1 event/hour)	1.5 | 98.5	11.9 | 88.1	3.0 | 97.0	12.1 | 0.002
BMI (underweight% | normal% | overweight% | obese%)	6.6 | 87.5 | 4.4 | 1.5	4.8 | 89.7 | 4.8 | 0.8	0 | 1.5 | 25.7 | 72.8	398.3 | <0.001
Waist circumference (normal% | increased%)	73.0 |27.0	60.3 | 39.7	2.5 | 97.5	202.8 | <0.001
Neck circumference (normal% | increased%)	100 | 0	99.2 | 0.8	70.3 | 29.7	51.4 | <0.001
Arterial hypertension (No% | Yes%)	94.8 | 5.1	95.2 | 4.8	81.7 | 18.3	21.2 | <0.001
Epworth sleepiness scale (No% | Yes%)	83.8 | 16.2	80.9 | 19.0	74.75| 25.25	27.1 | <0.001

Table 1. *Summary of the main differences between modules of the clinical, socio-demographic, and PSG data.*

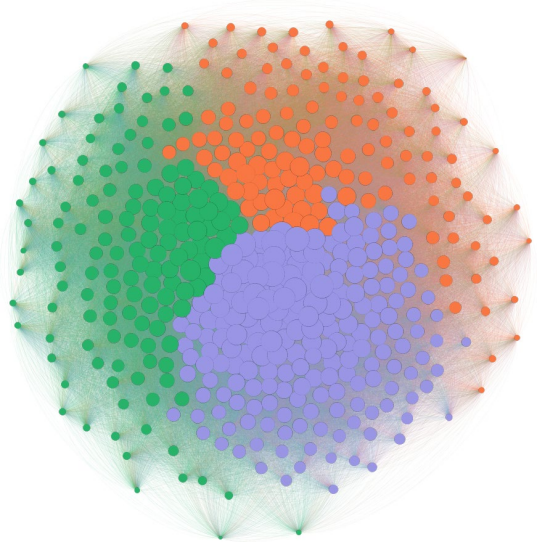

Figure 1. *Subject-based association network of the pediatric OSA patients. The network is derived from clinical, socio-demographic, and PSG data after a bootstrap procedure. The size of each node (patient) is determined by the average level of association with the rest of the nodes in the network. The color represents the module to which it belongs according to Blondel's algorithm* [12]. *Its spatial distribution follows a heuristic of attractive and repulsive force to reach a stable position of all the nodes, following ForceAtlas2 approach* [11].

age, with 65.8% of patients in this module older than 7 years of age.

Module C showed other noteworthy characteristics frequently related to weight. A noticeable example is that the 98.52% of patients in this module are overweight or obese. Also, waist and neck circumference showed significant higher values compared to modules A and B. In fact, among all overweight/obese, increased waist, and increased neck subjects, 93.75%, 98.36%, and 69.37% out of them belong to module C.

Finally, in line with the previously mentioned results, the subjects belonging to module C also showed differentiating clinical characteristics, as in the Epworth sleepiness scale (ESS). About a quarter of the subjects in this module showed excessive daytime sleepiness. Interestingly, OAHI was the characteristic with lower significant differences between the three modules. Together, these results highlight the noticeable association between different PSG, sociodemographic, and clinical parameters, as well as the ability of the proposed methodology to find modules of subjects with well-differentiated characteristics.

4. Discussion

In this study, we applied a novel association network-based approach to identify phenotypic profiles in children with OSA. After the use of a robust bootstrap-based statistical assessment, the conducted unsupervised modularity analysis revealed 3 distinctive groups of pediatric OSA patients.

One of the identified subgroups (module C) showed a distinctive profile in obesity-related variables, such as BMI, and waist and neck circumferences. Interestingly, previous studies related obesity to the ability to adhere to treatment in children suffering from OSA [13]. Moreover, and again in line with previous studies [5], this group did not show a higher OSA-severity profile in terms of OAHI. Indeed, OAHI was the characteristic with the lowest differences among the 8 variables of Table 1.

In addition to the obesity-related variables, our subject-based association networks let identify new characteristics that, although potentially associated with weight, have not been related to phenotypic subgroups of pediatric patients. This is the case of the age, sex, sleepiness, and arterial hypertension. All of them showed great significant differences between the subgroups of patients found. Based on these findings, it seems feasible to establish the hypothesis that patients belonging to module C may have either lower response to adherence or treatment [5], thus benefiting the most of personalized interventions. This, however, remains unexplored, so future studies with a longitudinal design could be conducted to evaluate it.

Curiously, a recent work focused in adult OSA phenotyping, and using 40 variables, showed 3 different modules too [7]. Despite some of the variables defining the subgroups are not suitable for children, such as alcohol consumption, the authors also reported the high importance of obesity, neck circumference, sex, and age to define the groups [7].

Among the limitations of the study, it is worth highlighting the lack of cognitive variables for the identification of subgroups of subjects. In this regard, recent works show the importance of exhaustively characterize the phenotypic profile of patients and recommend decreasing the threshold for conducting adenotonsillectomy from 5 event/hour of OAHI to 1 events/hour if the child presents associated neurocognitive symptoms [14]. Therefore, future studies should include these types of variables into account when establishing subgroups of patients with the aim of finding those who respond appropriately. In addition, a longitudinal study design could provide important clues regarding the adherence capability to treatment of each subgroup of pediatric patients. Finally, the use of variables tailored to children could derive in higher modularity degrees.

5. Conclusions

In this study, a new method grounded on subject-based association networks, and its modularity analysis, was proposed to identify subgroups of OSA patients with well-differentiated phenotypic characteristics. Our novel proposal reached results coherent with other studies conducted in both adults and children. One of the three subgroups automatically identified shows a profile of variables consistent with lower adhere/response to treatment. Additionally, the usefulness of previously unexplored variables relevant for OSA subgrouping were determined. This opens the door to the search for new predictive variables of adherence and response to pediatric OSA treatment, while highlighting the need for simultaneously acquiring clinical and sociodemographic variables beyond those purely derived from PSG to provide tailored diagnosis and interventions.

Acknowledgements

This research has been developed under the grants PID2020-115468RB-I00 and PDC2021-120775-I00 funded by 'Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033/' and ERDF, A way of making Europe; and by 'CIBER en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)' through 'Instituto de Salud Carlos III' co-funded with ERDF funds as well as under the project SleepyHeart from 2020 CIBER-BBN valorization call.

References

[1] C. L. Marcus *et al.*, "Diagnosis and Management of Childhood Obstructive Sleep Apnea Syndrome," *Pediatrics*, vol. 130, no. 3, pp. e714–e755, Sep. 2012, doi: 10.1542/peds.2012-1672.

[2] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules," *Terminology and Technical Specification*, 2007, Accessed: Oct. 13, 2021. [Online]. Available: http://ci.nii.ac.jp/naid/10024500923/en/

[3] C. L. Marcus *et al.*, "A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea," *New England Journal of Medicine*, vol. 368, no. 25, pp. 2366–2376, Jun. 2013, doi: 10.1056/NEJMoa1215881.

[4] M. Kohler, D. Smith, V. Tippett, and J. R. Stradling, "Predictors of long-term compliance with continuous positive airway pressure," *Thorax*, vol. 65, no. 9, pp. 829–832, Sep. 2010, doi: 10.1136/thx.2010.135848.

[5] N. DiFeo *et al.*, "Predictors of Positive Airway Pressure Therapy Adherence in Children: A Prospective Study," *Journal of Clinical Sleep Medicine*, vol. 08, no. 03, pp. 279–286, Jun. 2012, doi: 10.5664/jcsm.1914.

[6] G. C. Gutiérrez-Tobal *et al.*, "Pediatric Sleep Apnea: The Overnight Electroencephalogram as a Phenotypic Biomarker," *Front Neurosci*, vol. 15, Nov. 2021, doi: 10.3389/fnins.2021.644697.

[7] D. Ferreira-Santos and P. P. Rodrigues, "Obstructive sleep apnea: A categorical cluster analysis and visualization," *Pulmonology*, Nov. 2021, doi: 10.1016/j.pulmoe.2021.10.003.

[8] P. J. Gutiérrez-Díez, J. Gomez-Pilar, R. Hornero, J. Martínez-Rodríguez, M. A. López-Marcos, and J. Russo, "The role of gene-to-gene interaction in the breast's genomic signature of pregnancy," *Sci Rep*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-81704-8.

[9] N. Jimeno *et al.*, "Main Symptomatic Treatment Targets in Suspected and Early Psychosis: New Insights From Network Analysis," *Schizophr Bull*, pp. 1–12, Jan. 2020, doi: 10.1093/schbul/sbz140.

[10] S. Epskamp, D. Borsboom, and E. I. Fried, "Estimating psychological networks and their accuracy: A tutorial paper," *Behav Res Methods*, vol. 50, no. 1, pp. 195–212, Feb. 2018, doi: 10.3758/s13428-017-0862-1.

[11] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software," *PLoS One*, vol. 9, no. 6, pp. 1–12, 2014, doi: 10.1371/journal.pone.0098679.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[13] S. L. Katz *et al.*, "Factors related to positive airway pressure therapy adherence in children with obesity and sleep-disordered breathing," *Journal of Clinical Sleep Medicine*, vol. 16, no. 5, pp. 733–741, May 2020, doi: 10.5664/jcsm.8336.

[14] H.-L. Tan, M. L. Alonso Alvarez, M. Tsaoussoglou, S. Weber, and A. G. Kaditis, "When and why to treat the child who snores?," *Pediatr Pulmonol*, vol. 52, no. 3, pp. 399–412, Mar. 2017, doi: 10.1002/ppul.23658.