Nation-Wide ePrescription Data Reveals Landscape of Physicians and Their Drug Prescribing Patterns in Slovenia

Pavlin G. Poličar[1(✉)], Dalibor Stanimirović[2], and Blaž Zupan[1]

[1] Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia
pavlin.policar@fri.uni-lj.si

[2] Faculty of Public Administration, University of Ljubljana, 1000 Ljubljana, Slovenia

Abstract. Throughout biomedicine, researchers aim to characterize entities of interest, infer landscapes of cells, tissues, diseases, treatments, and drugs, and reason on their relations. We here report on a data-driven approach to construct the landscape of all the physicians in Slovenia and uncover patterns of their drug prescriptions. To characterize physicians, we use the data on their ePrescriptions as provided by the Slovenian National Institute of Public Health. The data from the entire year of 2018 includes 10,766 physicians and 23,380 drugs. We describe physicians with vectors of drug prescription frequency and use the t-SNE dimensionality reduction technique to create a visual map of practicing physicians. We develop an embedding annotation technique that describes each visually-discernible cluster in the visualization with enriched top-level Anatomical Therapeutic Chemical classification terms. Our analysis shows that distinct groups of physicians correspond to different specializations, including dermatology, gynecology, and psychiatry. The visualization also reveals potential overlaps of drug prescribing patterns, indicating possible trends of physicians practicing multiple disciplines. Our approach provides a helpful visual representation of the landscape of the country's physicians, reveals their prescription domains, and provides an instrument to inform and support healthcare managers and policymakers in reviewing the country's public health status and resource allocation.

Keywords: drug prescription · drug prescribing patterns · physicians · two-dimensional embedding · explainable point-based visualisations · ATC classification

1 Introduction

The practice of medicine is a broad field comprising various medical specialties, each focusing on specific areas of the human body and its functions. One way to characterize a physician's area of interest and operation is by analyzing the drugs they prescribe to their patients. Different medical specialists will typically

J. M. Juarez et al. (Eds.): AIME 2023, LNAI 13897, pp. 283–292, 2023.
https://doi.org/10.1007/978-3-031-34344-5_34

prescribe different medications based on their expertise and the types of patients they are treating. For example, a cardiologist will typically prescribe medications for hypertension and heart disease, while a neurologist will prescribe medicines for neurological disorders such as seizures and Parkinson's.

The literature includes only a limited number of reports in that aim to characterize physicians based on their drug-prescribing patterns. Akhlaghi *et al.* [1] inspect data from Iran's second most populous province and identify important pairs of co-prescribed drugs. They use the occurrence of these pairs as features in a random-forest model trained to predict physician specialty. Shirazi *et al.* [10] use an unsupervised community-detection approach to identify nine major groups of physician specialties from a bi-partite *physician-drug* graph. In a different targeted study, Garg *et al.* [4] focus on identifying family physicians using machine learning methods. However, their research does not include prescription data but uses physician-specific attributes, such as sex, age, and various certifications.

In contrast to previous work, we do not have access to information about pyhsicians' actual specialties. We instead leverage an unsupervised machine learning approach to uncover and characterize groups of physicians with similar drug-prescribing patterns. Our study focuses on all physicians from Slovenia. To the best of our knowledge, this is the first study to identify physician specialties based on drug prescribing patterns encompassing the entire country's data. To this end, we develop a straightforward statistical approach based on p-value hypothesis testing that identifies characteristic top-level Anatomical Therapeutic Chemical (ATC) classification terms for visually-discernible clusters in a given two-dimensional embedding.

We start our report with a description of the data, a two-dimensional embedding approach, and a method to explain visually discernable clusters. We provide the results in annotated two-dimensional maps of physicians. In the discussion, we show that the map visually and interpretably reveals the landscape of Slovenian physicians and their drug-prescribing patterns.

2 Data and Methods

We here introduce the ePrescription data used throughout the study and provide an overview of the data construction and filtering approaches. We then describe our proposed method to identify overrepresented top-level ATC classification terms in visually discernible clusters in a t-SNE two-dimensional embedding.

2.1 Data

We use anonymized data containing all ePrescription records from Slovenia prescribed in 2018. We obtained the data from the Slovenian National Institute of Public Health. We have to note that the data, albeit anonymized, is not public and, at this stage, cannot be openly shared, subject to restrictive Slovenian law. Slovenia has a centralized healthcare system and national eHealth solution,

making ePrescription records representative of the whole population of Slovenia. We chose 2018 to inspect physicians' landscape in the year before the pandemic era. In the early stages of the COVID-19 pandemic, many non-essential medical procedures were temporarily halted in Slovenia. These interruptions may have made interpreting the final embedding more difficult.

To construct the data matrix, we considered each physician and counted the number of times they prescribed each drug in 2018. To avoid drug brand preferences, we map each of the 23,380 drugs to its corresponding ATC classification term. We characterize each physician through a vector of prescription counts. The proposed procedure created a data matrix comprising 10,766 unique physicians and 920 unique level-5 ATC classification terms. To make our analysis more robust, we removed physicians that prescribed fewer than 25 drugs throughout the year, leaving us with 7,290 physicians. Metadata, such as patient age and sex, were excluded from the main analysis but also proves helpful, as discussed in Sect. 3.

2.2 Methods

Our approach aims to identify and annotate clusters in a given two-dimensional embedding. We here use t-SNE embeddings [7] because as it prioritizes cluster identification, though we could conduct a similar analysis on any other dimensionality reduction approach that yields two-dimensional maps of physicians. As is standard in dimensionality reduction of high-dimensional data, we first extract the top fifty principal components from the data [5]. We perform t-SNE dimensionality reduction with the openTSNE library v0.6.2 [8] using cosine distances, a perplexity value of 50, an exaggeration factor of 1.5, and degrees of freedom set to 0.8. We found that these parameter settings produce clear, well-separable clusters.

The first step of our pipeline is to identify visually discernible clusters in the two-dimensional embedding. As such, the embedding method of choice should produce discrete, well-separated clusters. On the resulting map, we then use the DBSCAN clustering algorithm [3] and manually tune the hyperparameters to achieve a visually sensible clustering. We found that setting DBSCAN's parameter values epsilon to 1.4 and setting the minimum number of samples to 25 identified clusters that best coincided with visually discernable groups of points in the embedding. It is worth noting that this aspect of the procedure could be automated, for instance, by maximizing the silhouette score [9] or by using DBSCANs internal parameter selection procedure [3].

To identify drugs prescribed more frequently in the identified clusters, we use a two-sample t-test and perform multiple hypothesis correction using the false discovery rate (FDR) with a threshold of $\alpha < 0.01$. To avoid selecting common drugs widely prescribed by all physicians or drugs where only minor differences in prescription frequency occur, we remove any drugs with a log-fold-change smaller than 0.25. The log-fold-change is the ratio of change between two values on a logarithmic scale and is used to compare changes over multiple orders of magnitude. As we are interested primarily in common prescription patterns in various

physician specializations, we also require drugs to be prescribed by at least 25% of all physicians within the cluster. This produces a list of overrepresented level-5 ATC classification terms in each cluster. We then use these terms in the next step for identifying overrepresented, top-level ATC classification terms.

Finally, to perform enrichment of top-level ATC terms, we use the hypergeometric test and perform multiple hypothesis corrections using FDR with a threshold of $\alpha < 0.01$.

3 Results and Discussion

Figure 1 shows the resulting annotated map of physicians from Slovenia. With the proposed methods, we could determine the specific drug-prescribing patterns for most of the clusters in the physician map and abstract these patterns and their related specializations. We also find that some common physician specialties are missing from our annotations, including rheumatologists, oncologists, and radiologists. This result does not imply that there are no such specialists in Slovenia, but rather that their drug prescribing patterns likely coincide with other specializations and are probably contained within other clusters. In a related study, Shirazi *et al.* [10] apply community detection to ePrescription data from Iran and identify clusters of different specializations. However, they find that many identified clusters contain multiple specializations. Among these are groups of neurologists and cardiologists, internists, general practitioners, and dermatologists.

Our procedure only identifies overrepresented ATC classification terms. It is still up to us to manually determine which combination of these terms corresponds to each specialization. In many cases, the specialization of each cluster is fairly obvious, as the drugs prescribed by physicians in each cluster predominantly come from a single, top-level ATC classification group. We observed such results with gynecologists, neurologists, dermatologists, ophthalmologists, and endocrinologists.

Sometimes the top-level ATC classification terms alone are insufficient, so we examine overrepresented ATC classification terms further down the ontology by repeating the same enrichement scoring procedure. For instance, both cardiologists and hematologists predominantly prescribe drugs from the *B (blood and blood-forming organs)* ATC classification group making them difficult to distinguish from one another. However, upon closer inspection, we find that the top cluster tends to prescribe ACE inhibitors and diuretics more often than the lower cluster – drugs that are often used to treat cardiovascular conditions, indicating that the top cluster corresponds to cardiologists and the lower to hematologists.

We are able to confidently assign a specialization to most clusters, with the exception of the lower left cluster, which we guess corresponds to general practitioners. There are also two central clusters, colored green and light green, with no enriched top-level ATC classification terms. Inspecting lower-level ATC classification terms yielded no satisfactory conclusion. Therefore, we opt to leave these clusters unlabelled.

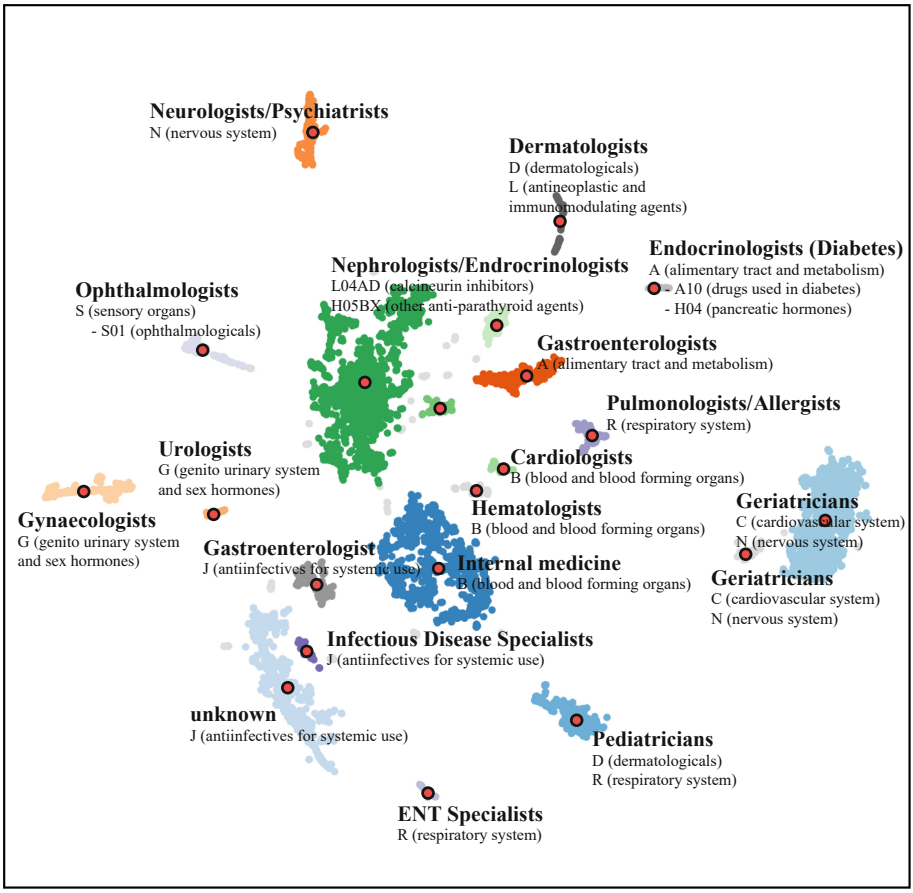

Fig. 1. Annotated t-SNE map of Slovenia phyicians characterized by their drug prescriptions. Due to the positional invariance of nonlinear embedding methods, we omit axis labels and scales. Each point on the plot represents a physician. Points are color-coded according to their cluster assignments. Physicians identified as outliers by the DBSCAN clustering algorithm are colored light gray. We indicate the center of each cluster by a red point, accompanied by the determined cluster specialization. We list the automatically-inferred, top-level ATC codes characterizing each cluster below the specialization. It is worth noting that two of the center-most clusters do not contain an annotation. Upon further investigation, we found that there were no overrepresented top-level ATC terms for these clusters, and thus, we could not assign them a distinct specialization. We hypothesize that these physicians are general practitioners. (Color figure online)

3.1 Non-drug Prescription Data Aids in the Identification of Physician Specializations

In certain instances, drug-prescribing patterns alone may not be sufficient to unambiguously determine a physician's specialization. In such cases, other meta-data available in ePrescription records, such as patient age and sex, can also provide valuable information in determining a physician's area of expertise. For example, a pediatrician typically treats patients under 18, while a geriatrician generally treats patients over 65. Similarly, a gynecologist predominantly treats female patients, while a urologist typically treats male patients.

We plot the percentage of female patients of each physician in the physician map on Fig. 2.a. While we were able to determine clusters corresponding to urologists and gynecologists based on prescription patterns alone, we use Fig. 2.a to validate our cluster assignments.

We plot the median patient age of each physician in Fig. 2.b. The red dots correspond to physicians treating younger patients. While physicians with younger patients are scattered across several clusters, they are highly concentrated in the lower-right cluster. We conclude that this cluster corresponds to pediatricians. Interestingly, we also observe a reddish hue in points corresponding to

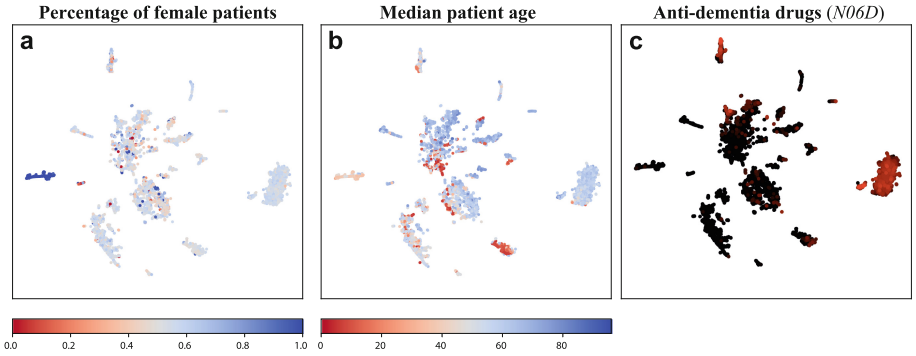

Fig. 2. Non-drug prescription data helps to gain additional insight into physician specializations or further validates the cluster annotations. For instance, panel (**a**) shows the proportion of female patients treated by each physician. This additional information confirms the cluster assignments of gynecologists and urologists, both very gender-specific specializations. Panel (**b**) depicts the median patient age treated by each physician. The cluster with the lowest median age, indicated in red, corresponds to pediatricians, who primarily treat patients under 18. We would also be interested in identifying geriatricians who treat older patients, but this plot reveals several candidate clusters. To identify geriatricians, we plot the frequency of the prescription of anti-dementia drugs in (**c**), as dementia is a prevalent disease in the elderly population. The plot reveals two potential clusters for anti-dementia drugs. From our annotation procedure, we have already identified the top cluster to correspond to neurologists/psychiatrists. Using extra information from meta-data gives us a high level of confidence that the right cluster corresponds to geriatricians.

gynecologists, indicating that the median age of gynecologist patients appears to be around 40 years old. Krause *et al.* [6] report similar findings in the German healthcare system, where they note that visits to gynecologists drop off with age.

Figure 2.b reveals several clusters containing physicians with older patients. We also investigate anti-dementia drug prescriptions to identify which corresponds to geriatricians. The incidence of dementia increases exponentially with age [2]. Therefore we would expect geriatricians to prescribe more anti-dementia drugs than their peers. Figure 2.c shows the frequency of the prescription of all drugs corresponding to anti-dementia drugs, with the ATC designation *N06D*. Figure 2.c reveals that physicians most frequently prescribe anti-dementia drugs from two distinct clusters, one at the top and one to the right of the embedding. Using our ATC-term enrichment approach, we determined that the top cluster corresponds to neurologists/psychiatrists. Based on this information, we infer that the cluster to the right corresponds to geriatricians. This finding is already hinted at in Fig. 2.b, as the top cluster contains a mix of red and blue dots. In contrast, the cluster on the right contains mostly physicians colored with a blueish hue, corresponding to physicians treating older patients.

3.2 Investigating Previously-Identified, Specialization-Specific Drugs

In a related study, Akhlaghi *et al.* [1] develop a predictive model for predicting physician specialization in Iran based on their drug-prescribing patterns. They identify several frequently-prescribed, specialization-specific, co-prescribed drug pairs that were found to be essential for distinguishing between different physician specializations. Unlike in our study, the authors had access to physicians' true specializations. We can then investigate whether the identified drug pairs also appear in Slovenia and evaluate their discriminatory power between different physician specializations.

We select three reliably-identified specialization clusters from our results and inspect the occurrence and discriminatory power of the drug pairs identified in Iran. We show three such combinations in Fig. 3. The plots show that while some of the identified drug pairs in Iran are, in fact, highly prescribed in their target specializations, they are not altogether discriminatory. For instance, the Clotrimazole-Fluconazole drug pair is characteristic of gynecologists in Iran. Inspecting these individually in the top two panels of Fig. 3, we can see that they are, in fact, overrepresented in the gynecology cluster. However, both of these drugs, especially Fluconazole, also appear in several other physician clusters throughout the embedding. To inspect the pairing of these two drugs, we calculate the product of the prescription counts of each of the two drugs. We show the results in the bottom panel of Fig. 3. While many gynecologists prescribe the drug pair Clotrimazole and Fluconazole, it is also prescribed by other specialists, including geriatricians, pediatricians, dermatologists, and ENT specialists. Neurologists/psychiatrists behave similarly. The Acetaminophen-Ketotifen drug pair identified for pediatricians appears to have less discriminatory power.

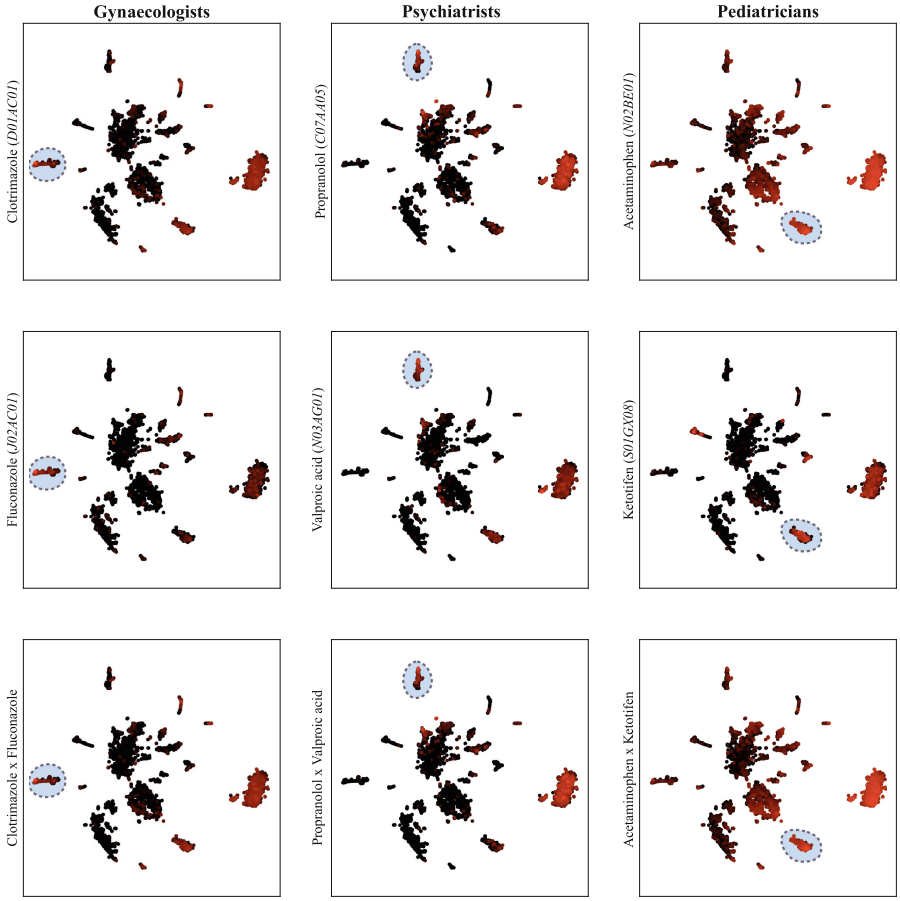

Fig. 3. We plot the prescription frequency of specilization-specific drug pairs identified by Akhlaghi *et al.* [1] for each physician. The first and second rows show the prescription frequencies of the individual drugs, while the third row plots the product of their respective frequency values. The color intensity indicates the log-transformed frequency of each drug/drug pair. The light-blue areas indicate the regions where each previously-identified physician specialty actually occurs.

This lack of specificity could result from several factors or a combination thereof. Firstly, the discrepancies between our and Iran's study could indicate potential cultural and regional differences in prescribing patterns between physicians in different countries or systemic differences in their healthcare medical and education systems. Direct comparisons between countries can be misleading, as the general population's demographic makeup and dietary habits play an essential role in the overall drug prescribing trends. Secondly, the drug pairs identified by Akhlaghi *et al.* [1] were used for prediction in a highly non-linear random-forest model and not individually, as was shown in our case. Perhaps

the top-rated drug pairs are only discriminatory when used in conjunction with other drug pairs. Lastly, it may be difficult to identify drugs that are truly specific to each specialization alone. Some specializations, e.g., gynecologists or psychiatrists, are easily identifiable since most prescribed drugs originate from the corresponding ATC classification group. However, as seen in Figs. 2.c and 3, even drugs that are relatively specific to these specialists often appear in at least one other cluster as well, e.g., anti-dementia drugs are prescribed by both neurologists and geriatricians. Using their approach, the authors were able to achieve 74% accuracy when predicting physician specialization, indicating that this may not be an altogether trivial task.

4 Conclusion and Future Directions

In this study, we developed an unsupervised computational approach to infer annotated maps of physicians based on their drug prescribing characteristics. The annotations for clusters on the map use an external database with ATC classification terms. We found that the inferred physician map includes highly discernable and interpretable clusters. Unlike previous studies, we did have access to physicians' true specializations. While we used our approach to the map of physicians, we could similarly apply our method to other data modalities or clusters, e.g., patients, diseases, drugs, or institutions.

The results of this study raise several interesting points. Firstly, it is interesting that our embedding and clustering procedure uncovered such distinct groups of physicians. Using our enrichment scheme, we were able to associate these with different physician specializations. Additionally, we found that we could not unambiguously classify certain physician clusters without additional metadata, emphasizing the need to incorporate different data modalities and ontologies into similar analyses. This also highlights the benefits of constructing data maps from one data source and providing explanations from other data sources or ontologies. Interestingly, certain physician specializations were missing from our final embedding. This indicates that similar drugs and drug-prescribing patterns likely appear between different specializations and are encompassed within a single cluster. We also compare the different drug-prescribing patterns between Iranian and Slovenian physicians. Our preliminary results indicate potential differences between the drug-prescribing patterns between the two countries. This could be a result of a variety of factors, including regional, cultural, or educational factors. The differences between drug-prescribing patterns in different countries are poorly understood, and we have here provided an example of this discordance. While our results are interesting, their benefits and real-world impact are in the hands of domain experts and policymakers. In our case, we are continuing collaboration with the Slovenian National Institute of Public Health to see how the knowledge gained from our study can lead to a better understanding of the needs of the Slovenian healthcare system and provide them with the tools for a more evidence-based decision-making approach.

Acknowledgements. This work was supported by the Slovenian Research Agency Program Grant P2-0209. We would also like to thank the Slovenian National Institute of Public Health for their constructive cooperation.

References

1. Akhlaghi, M., Tabesh, H., Mahaki, B., Malekpour, M.R., Ghasemi, E., Mansourian, M.: Predicting the physician's specialty using a medical prescription database. Computat. Math. Methods Med. (2022)
2. Birkenhäger, W.H., Forette, F., Seux, M.L., Wang, J.G., Staessen, J.A.: Blood pressure, cognitive functions, and prevention of dementias in older patients with hypertension. Arch. Intern. Med. **161**(2), 152–156 (2001)
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
4. Garg, A., Savage, D.W., Choudhury, S., Mago, V.: Predicting family physicians based on their practice using machine learning. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 4069–4077. IEEE (2021)
5. Kobak, D., Berens, P.: The art of using t-SNE for single-cell transcriptomics. Nat. Commun. **10**(1), 5416 (2019)
6. Krause, L., Dini, L., Prütz, F.: Gynaecology and general practitioner services utilisation by women in the age group 50 years and older. J. Health Monit. **5**(2), 15 (2020)
7. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
8. Poličar, P.G., Stražar, M., Zupan, B.: openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. BioRxiv p. 731877 (2019)
9. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
10. Shirazi, S., Albadvi, A., Akhondzadeh, E., Farzadfar, F., Teimourpour, B.: A new application of community detection for identifying the real specialty of physicians. Int. J. Med. Inform. **140**, 104161 (2020)