MATHEMATICA APPLICANDA
Vol. 47(1) 2019, p. 127–139
doi: 10.14708/ma.v47i1.6497

Agata Lonc * (Warsaw)
Monika Joanna Piotrowska * (Warsaw)
Konrad Sakowski * (Fukuoka)

Analysis of the AOK Plus data and derived hospital network

Abstract We present an analysis of admission/discharge data from insurance provider for Saxony and Thuringia (Germany) for years 2010–2016. A study of such data is necessary to derive a structure of a healthcare system transfer network, as no patients' transfer data are available. Hospital network is a basis for simulation of multidrug-resistant bacteriae spread allowing to study the effectiveness of disease-control strategies. In this paper, the properties of the dataset under consideration are presented and discussed. Moreover, the resulting inter-hospital network structure is analyzed.

2010 Mathematics Subject Classification: Primary: 62-07; Secondary: 92C42.

Key words and phrases: healthcare data analysis, overlapping data, healthcare network, multidrug-resistant bacteria..

1. Introduction. Multidrug-resistant bacteria (MDR) recently became more serious health threat [2, 8]. According to the European Centre for Disease Prevention and Control and the European Medicines Agency about 25 000 patients die in the EU from an infection with MDR each year, [5]. In addition, it is estimated that these infections result in extra healthcare costs of at least EUR 1.5 billion each year. These pathogens can spread within the healthcare system population, e.g. by contact between undiagnosed infectious patients and susceptible patients. Pathogen control strategies exist (see e.g. [1, 2]), but they mainly focus on the individual healthcare facility level, while the cooperative approach may be more beneficial [7]. Unique properties of MDR make them immune to typical prevention strategies, mainly due to their antibiotic-resistant nature. In order to understand how to inhibit spreading of such bacteria, new models describing their transmissions are introduced. Such attempts to model the spread of some particular bacteria within the hospital network have already been undertaken, see e.g. [3, 4, 6].

* 0000-0002-7045-111X
* 0000-0003-0156-8290
* 0000-0003-1511-074X

Clearly, to model such a phenomenon one needs to have access to a good database providing a lot of information about patients but also transfer data of patients between the hospitals. Towards the end of the second decade of the 21st century, when the computerization is present in virtually every domain of human activity, it seems natural to expect that the patient transfer data are available as a database, which can be used for an efficient data analysis. Unfortunately, this is not the case even in Europe. Hospitals, or healthcare facilities in general, collect information about patient admissions, discharges, age, gender, diagnosis, procedures in they own patient registers, however do not have information about hospitalisation history of patients. Nevertheless, an excellent source of data could be healthcare providers collecting patient hospitalization records for years. They store information that is the most relevant for the modellers, unfortunately they do not keep track of patients' transfers. So one of the difficulties for the modellers would be to extract the transfer data form the hospitalisation history. Such a task seems not to be too hard but, depending on the database, can be less or more challenging. In addition, modellers need to better know the dataset in a way allowing them e.g. to stratificate patients according to certain risk factors. To provide so much important information for the future work concerning creating models describing the dynamics of the pathogen spread within the inter-hospital network we perform an analysis of the dataset provided by AOK Plus – one of the largest health insurance company in Germany.

This paper is organized as follows. In Section 2 and 3 we briefly describe the database provided by AOK Plus and analyse it. Next, in Section 4, we describe basic properties of the generated inter-hospital network. Finally, in Section 5, we comment on the results and present future plans concerning modelling the spread of MDR bacterial infection.

2. Description of dataset. We consider the anonymized patients dataset provided by AOK Plus – a healthcare provider in Saxony and Thuringia. The dataset consists of 4 826 823 hospitalisation records of 1 623 567 patients covering the period of 7 years (2010 – 2016). In particular, the database stores the following information: patient anonymized ID, anonymized healthcare facility ID, the federal state of healthcare facility, the day of admission, the day of discharge, the diagnosis (international ICD-10-GM code), the patient's sex and year of birth.

Within the provided dataset we have found 2 991 597 hospital/healthcare facility stay records for the facilities located in Saxony with the numeric diagnosis code, 1 566 451 for Thuringia, 268 182 for other German federal states and 593 records without any location given. There are 1 925 unique hospital facilities among the whole database and 134 of them are situated in Saxony and Thuringia. Since the overwhelming number of records concerns Saxony and Thuringia in the further study we focus on these two federal states.

3. Data analysis. To investigate the provided dataset (admissions, discharges, duration of stays, sizes of the facilities etc.), to analyse their structure and statistics, we used the previously developed Python code, freely available (with documentation) on the web page www.mimuw.edu.pl/~monika/emergenet. Clearly, similarly as in our previous study of the AOK Lower Saxony dataset [9], grouping the records with the same patient identification number we faced the problem of the so-called *overlaps* – the non-empty intersection of stay periods for distinct sets of two or more records for a given patient, either within the same facility or in other facilities, for details see Subsection 3.5.

3.1. Population structure. After limiting the records to the facilities located in Saxony and Thuringia, we end up with data for 706 827 men with 1 up to 155 hospitalizations and 845 666 women with 1 up to 147 hospitalizations. The average number of admissions per person is 3.24 for men and 3.0 for women, the median of admissions per person is 2 for both sexes. The average length of hospitalisation is 9.856 days, and the average period spent outside the facilities, between two hospitalisations is 286.305 days. In Figure 1 we present the structure of patients' population. Clearly, we do not consider patients' age, as the database covers seven years. Furthermore, we do not know exactly which patients died during that period – such information is only available in the database if a patient died while being insured by AOK Plus. Thus, for our statistical purposes we consider the birth year to investigate the age structure of patients.

Among 4 558 048 hospitalisation records for healthcare facilities located in Saxony and Thuringia there are 843 512 (18.5%) cases of diseases of the circulatory system. Other significant groups are diseases of the digestive system (466 739, 10.2%), neoplasms (464 666, 10.2%), injury, poisoning and certain other consequences of external causes (460 118, 10.1%) and mental and behavioural disorders (327 221, 7.2%).

3.2. Admissions. From Figure 2 it is clear that the majority of facilities had between 10 000 and 99 999 admissions and the other groups were less common. This distribution is suitable for studying the inter-hospital transmissions of infections, because the facilities with very few patients (less than 10) in seven years do not have much impact on migrations of patients. In fact, out of 134 considered facilities, there is only one with less then 10 patients and five with 10-99 patients during this whole period. Thus they might be omitted as their contribution to the patient transfer is insignificant. For separately-treated years most Saxony and Thuringia healthcare facilities had between 1 000 and 9 999 reported admissions and the number of facilities with given intervals of reported admissions did not change much with time.

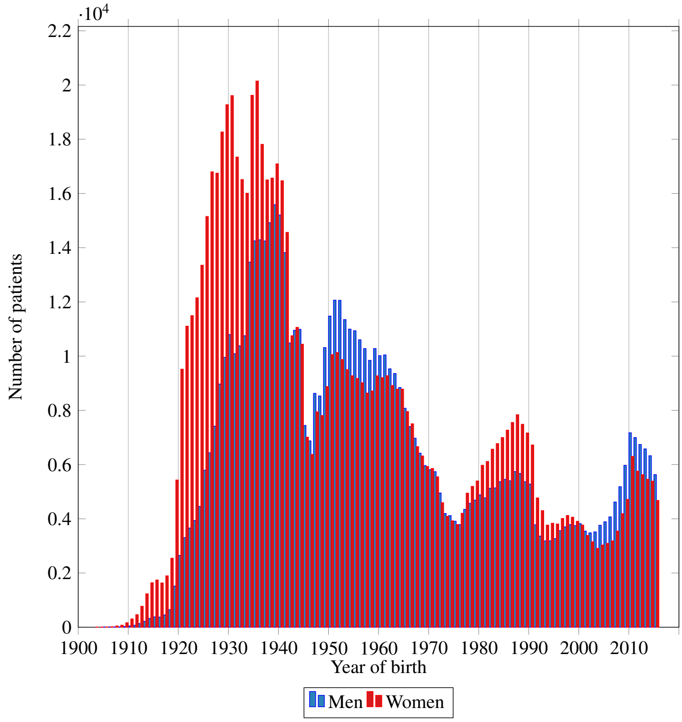

Figure 1: Structure of the patients' population.

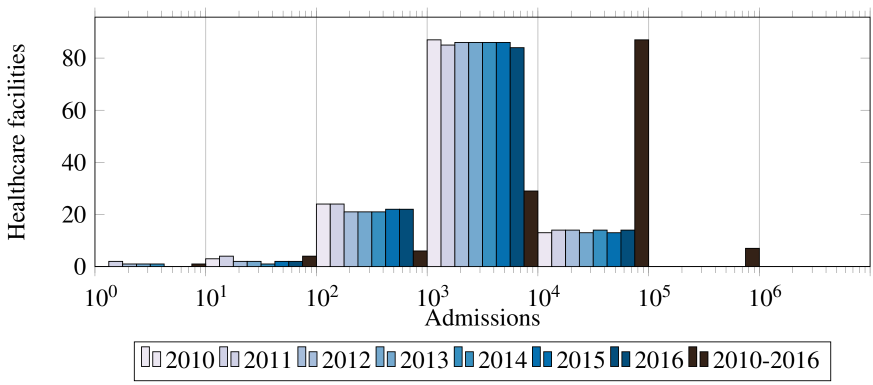

Figure 2: Number of healthcare facilities having given number of admissions for Saxony and Thuringia within years 2010-2016 and for separate years.

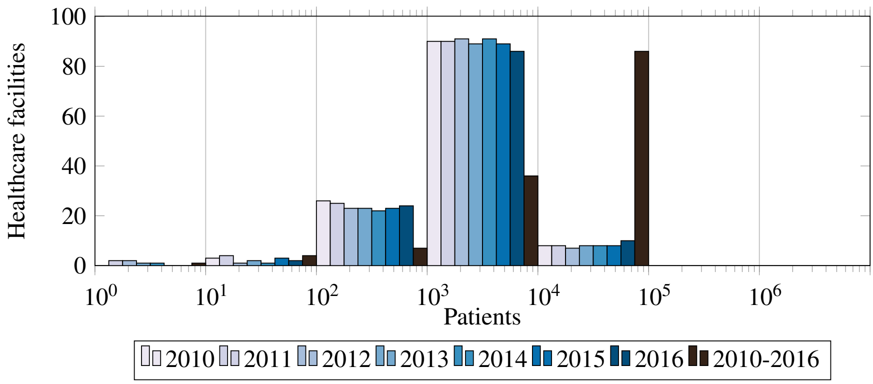

Figure 3: Number of healthcare facilities having given number of patients for Saxony and Thuringia within years 2010-2016 and for separate years.

3.3. Numbers of patients. In order to estimate the probabilities of patient transfers, one often needs information about the size of the healthcare facilities. Thus, we also categorise the healthcare facilities by the number of patients admitted to them. From Figure 3 it is clear that during years 2010–2016 the facilities with 10 000 to 99 999 patients were the largest group. However, looking at the years separately, we see that for each year facilities having between 1 000 and 9 999 patients dominated by far. We analyse the changes of the number of patients of each healthcare facility in time (from 2010 up to 2016). In general, for the biggest hospitals, we can distinguish two kinds of processes. Clearly, there are periodic variations of the number of patients, which occur simultaneously among the healthcare system. On the other hand, there are long-term increase/decrease of the facility population's size, which are specific to healthcare facility.

3.4. Duration of stays. In Figures 4a and 4b we report the duration of stays (until 31.12.2016) of patients in particular healthcare facilities and lengths of stays at home between hospitalizations, respectively. Clearly, the majority of hospitalizations do not exceed 10 days and most of them are 3 days long. The number of hospitalizations quickly decreases for the duration longer than 3 days. We can see that hospitalizations that last at least a month constitute only a marginal part of all records (less then 5% of all hospitalizations). When it comes to stays at home between hospitalizations, the number of stays first increases as the duration growths, reaches maximum for six-day-long stays and then decreases with some fluctuations. However, the decline is considerably slower than for the hospitalizations, and the stays lasting one hundred days or longer are still significant, as they constitute slightly more than a half of all home stays.

3.5. Overlaps. Among 4 558 048 detected in the Saxony and Thuringia

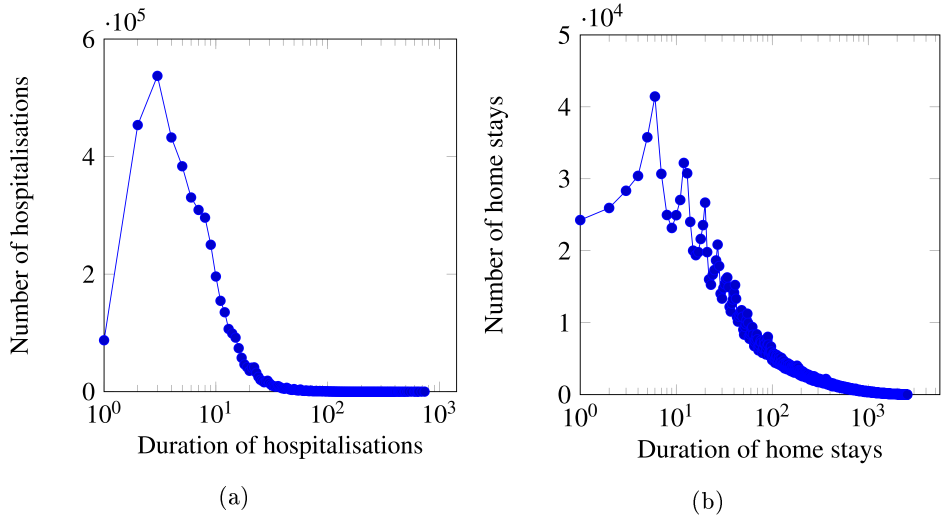

Figure 4: Durations of patients stays: (a) in healthcare facilities (b) at home between hospitalizations.

dataset hospitalizations we find 198 723 cases of overlapping records. Following [9], we distinguish several types of overlaps: *standard transfer* – one day overlap of two stay periods, where both periods are longer than one day and each record corresponds to a different facility; *first day transfer/last day transfer* – similar to above, but the duration of the stay in one facility is exactly one day long and it coincides with admission to/discharge from the latter facility; *simultaneous two admissions in a single institution* – two reported stays in the same place for the same period; *temporary transfer* – two records, period of one of them is contained in the other, and admission and discharge dates are not the same; *simultaneous two admissions in two different institutions* – periods are exactly the same, but the facilities are different; *unknown two admissions in two different institutions* – any two records for hospitalisations in different institutions, which are not covered by the cases already introduced; *two admissions in a single institution* – two reported stays in the same institution but for different (overlapping) periods and *unknown multiple admissions (n)* – more than two records of overlapping hospitalisation periods, with maximal number of records in a given day is n.

In Table 1 we present overlapping records within the years 2010–2016. The majority (over 77%) of them are typical transfers, meaning that both stay periods are covered only by one day and the stays are reported for different institutions – standard, the first day and the last day transfers. The other significant types are overlapping stays within one facility (over 13%) and temporary transfers (over 7%). Clearly, the most problematic cases such as simultaneous admissions in three or more facilities are marginal (around 0.5%) and therefore can be ignored. In general, the longer the overlaps are, the less

Table 1: Identified types of overlaps (units located in Saxony and Thuringia).

Overlap description	number of records
standard transfer	137 331 (69.1%)
two admissions in a single institution	26 402 (13.3%)
first day transfer	14 837 (7.5%)
temporary transfer	14 534 (7.3%)
unknown two admissions in two institutions	3 391 (1.7%)
unknown multiple admissions (3)	1 023 (0.5%)
last day transfer	900 (0.5%)
simultaneous two admissions in two institutions	271 (0.1%)
simultaneous two admissions in a single institution	32 (0.0%)
unknown multiple admissions (4+)	2 (0.0%)

often they appear in the database.

4. Properties of the inter-hospital network. Although we do not know the geographic localization of the reported in the database healthcare facilities we are still able to build the inter-hospital network on the basis of the hospitalization records. In the hospital network healthcare facilities are represented by nodes, while weighted edges represent the probabilities of transfers between healthcare facilities. Derivation of the patient transfers and then the inter-hospital network along with per-patient transfer probabilities is not straightforward and requires solving the overlaps issues first and next the precise counting of patient transfers for each day and the size of the population in each node. Thus, due to the complexity of that process the details will not be described here and we focus on the properties of the obtained network. Nevertheless, we expect that the obtained stochastic probability matrix is regular. It can be checked numerically by empirical verification, but this result is always dependent on numerical errors. Thus, we prefer to find an analytic argument, based on the following lemma.

Lemma 4.1 *Assume that* $A = [A_{ij}]_{i,j=1}^{k}$ *is a* $k \times k$ *dimensional real matrix such that*

1. $\forall i, j \in \{1, \ldots, k\}$, $i \neq j$, $A_{ij} \geq 0$,
2. $\forall j \in \{1, \ldots, k\}$ $A_{jj} > 0$,
3. $\forall i \in \{1, \ldots, k\}$ $\sum_{j=1}^{k} A_{ij} = 1$,
4. $\forall i, j \in \{1, \ldots, k\}$ $\exists\, i_0 = i, i_1, \ldots, i_{n-1}, i_n = j$ *such that* $\forall m \in \{1, \ldots, n\} \quad A_{i_{m-1} i_m} > 0$.

Then A *is a stochastic regular matrix.*

Proof A is a right stochastic matrix by Assumption 3 of this lemma. Thus, it is enough to prove that there exists sufficient large N such that all elements of $B = A^N$ are positive. Let us define $n(i,j)$ to be the smallest n so that Assumption 4 is true for indices i, j. Then, due to non-negativity of A_{ij} (Assumption 1), Assumption 4, and standard matrix multiplication rules we have

$$e_i^T A^{n(i,j)} e_j \geq \prod_{m=1}^{n} A_{i_{m-1} i_m} > 0, \tag{1}$$

where T denotes the transposition of the vector. Then, for any natural $l > 0$, due to Assumption 2, we have

$$e_i^T A^{n(i,j)+l} e_j \geq A_{jj}^l \prod_{m=1}^{n} A_{i_{m-1} i_m} > 0 \tag{2}$$

If we therefore define $N := \max\big\{n(i,j) : i, j \in \{1, \dots, k\}\big\}$, then by inequality (2) we get

$$B_{ij} = e_i^T B e_j = \left(e_i^T A^N\right)_j = \left(e_i^T A^{n(i,j)+l(i,j)}\right)_j > 0, \tag{3}$$

where $l(i,j) := N - n(i,j) \geq 0$ is an integer. Thus, matrix $B = A^N$ has only positive elements and A is regular by the definition. ∎

Clearly, Assumption 1 is satisfied as all the elements are non-negative, they are probabilities. Assumption 2 is more subtle, as in general A_{jj} can be equal to 0. It is however rather unlikely, as it would imply that no patients would stay in the healthcare facility overnight ever. While theoretically possible, this is unrealistic and such facility would be most likely removed from the simulation. Assumption 3 ensures that the probability matrix is stochastic. Since A_{ij} is a probability of a jump from ith node to jth node, clearly these probabilities must sum up to one, otherwise the Markov Process is not defined correctly. Assumption 4 means that for every two facilities, there is a (potential) transfer path between them. Actually, it is not necessary that any patient follows this whole path, but there must be a patient transfer for every component. Thus, the transfer path must exist between every two facilities, in both directions.

In Figure 5 we present visualization of the inter-hospital network built on hospitalization records for Saxony and Thuringia. In the analysis we examined 133 out of 134 healthcare facilities located in Saxony and Thuringia, as one of them has too few admissions in the database and may cause problems in the future models (e.g. creating absorbing state, in which all patients eventually end up). We examine every day of every hospitalization and check whether the patient changed the healthcare facility (transfer) or not (auto-transfer). There are 42 255 828 such direct transfers, among which there are 27 420 531 transfers between two facilities located in Saxony (including both transfers to another facility and auto-transfers), 14 828 627 transfers between two facilities located in Thuringia (both transfers and auto-transfers), 3 430 transfers from Saxony to Thuringia and 3 240 transfers from Thuringia to Saxony. The overwhelming majority (42 059 912, 99.5%) of these transfers are auto-transfers, which results in very high rate of re-admission to the same hospital.

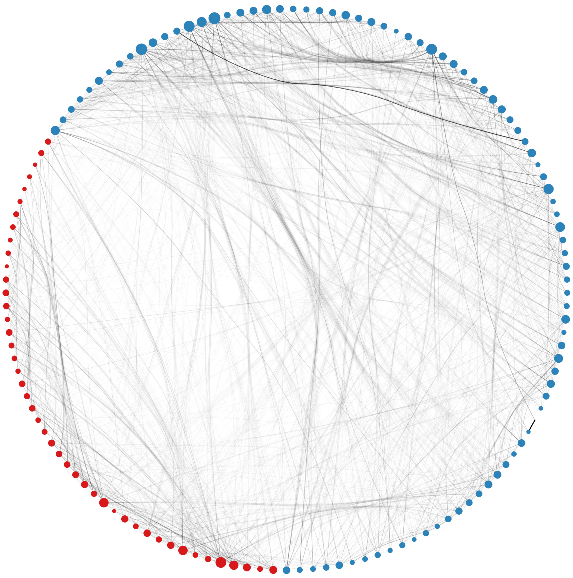

Figure 5: Visualization of the inter-hospital network considering only transfers between different facilities. Darker edges correspond to higher probability of transfer. Red nodes represent facilities located in Thuringia and blue nodes represent facilities located in Saxony. The size of the nodes represents the size of the set of both direct predecessors and successors.

The examined directed graph is strongly connected, i.e. it is possible to reach every node from any starting point. Since the size of the graph is not big, this fact can be proved by computer using a finite number of edges' combinations and checking whether they form a path between the two chosen nodes. The diameter of the graph, which is the longest path between any two nodes, is 3. Out of 17 556 ordered pairs of different nodes, 4 729 pairs are directly connected, 12 051 pairs have the shortest connecting path of length 2 and 776 pairs – of length 3. The radius of the graph is 2, meaning there exists a node, whose shortest path to every other node is less then or equal to 2 and no node is directly connected with all of the other nodes. Nodes have on average indegree and outdegree 36.6, which is the number of edges going respectively into the node and out of the node, excluding the edges connecting the nodes to themselves. For nodes representing facilities in Saxony the numbers are slightly higher – indegree 39.3 and outdegree 39.2 and for nodes representing facilities in Thuringia they are lower – 31.2 and 31.3, respectively. In Figure 6a

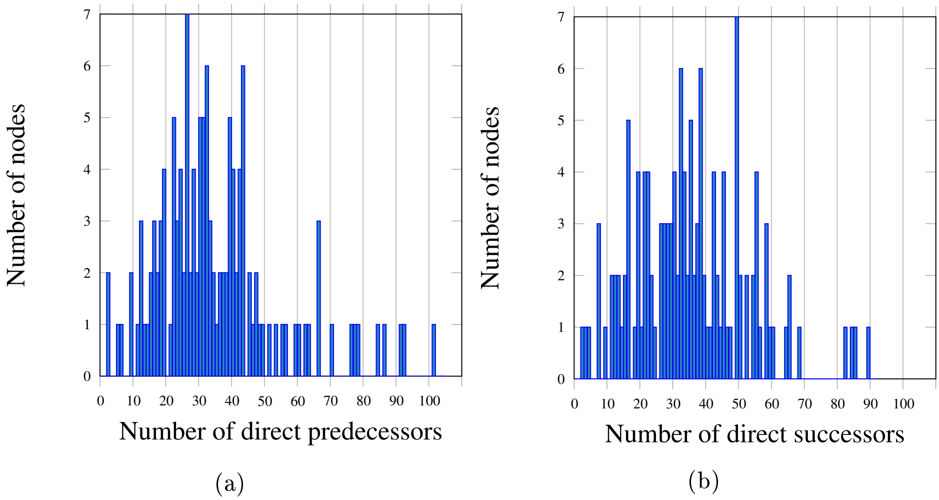

Figure 6: Number of nodes having the given number of: (a) direct predecessors, (b) direct successors.

we present the number of edges as a function of the indegree (number of direct predecessors) and in Figure 6b – as a function of the outdegree (number of direct successors).

These properties determine the problems to be faced in the context of the pathogen spread simulations in such a network. The incidence matrix has over 1/4 non-zero elements, which along with average predecessor/successor number makes it a dense matrix ($O(n^2)$ nonzero elements, rather than $O(n)$ for a sparse matrix). Moreover, no loosely-connected hospital groups can be found below the region-level. As we see in Figure 5, even between regions the ties may be significant. This poses a problem for the parallel simulations, as the inter-process communication will be significant. Moreover, from a practical viewpoint, it is very hard to prevent the pathogen transmission through the hospital transfer path, as there is no clear method to localize the disease to a subgroup of facilities, as it would basically mean that it breaks all the ties, which is hardly feasible.

5. Discussion The analysis of the AOK Plus database is extremely useful for modelling the spread of MDR bacterial infection. Information about the population structure enables us to divide patients based on the risk factors, such as age or gender. Furthermore, the classification of healthcare facilities based on the number of admissions or the number of patients allows us to identify which facilities may be more prone to infection outbreaks.

Out of the admissions database we extracted the information about patients' transfers between the hospitals. Based on the records we created the inter-hospital network that contains probabilities of direct transfers. Our analysis shows that the network is rather well connected in favour of the Saxony subnetwork. However, the transfers between Saxony and Thuringia are also well represented.

Since the examined network is built on the direct transfers only, it does not take into account cases when the patient stays at home between the hospitalizations. Hence, it does not reflect the actual spread of the bacteria as after leaving one hospital a patient may stay infected for a long period of time and still be infected while being admitted to another facility. A solution to this problem would be an introduction of additional node(s) representing stays outside of healthcare facilities leading to the analysis of a greater network.

6. Acknowledgements. This work was supported by grant no. 2016/22 /Z/ST1/00690 of National Science Centre, Poland within the transnational research programme JPI-EC-AMR (Joint Programming Initiative on Antimicrobial Resistance) entitled "Effectiveness of infection control strategies against intra- and inter-hospital transmission of MultidruG-resistant Enterobacteriaceae — insights from a multi-level mathematical NeTwork model" (EMerGe-Net). We thank the AOK Plus for providing anonymized record data.

References

[1] C. Camus, E. Bellissant, A. Legras, A. Renault, A. Gacouin, S. Lavoué, B. Branger, P. Y. Donnio, P. le Corre, Y. Le Tulzo, D. Perrotin, and R. Thomas. Randomized comparison of 2 protocols to prevent acquisition of methicillin-resistant Staphylococcus aureus: Results of a 2-Center study involving 500 patients. *Infection Control and Hospital Epidemiology*, 32 (11):1064–1072, 2011. doi: 10.1086/662180. Cited on p. 127.

[2] I. F. Chaberny, F. Schwab, S. Ziesing, S. Suerbaum, and P. Gastmeier. Impact of routine surgical ward and intensive care unit admission surveillance cultures on hospital-wide nosocomial methicillin-resistant Staphylococcus aureus infections in a university hospital: An interrupted time-series analysis. *Journal of Antimicrobial Chemotherapy*, 62(6):1422–1429, 2008. doi: 10.1093/jac/dkn373. Cited on p. 127.

[3] T. Donker, J. Wallinga, R. Slack, and H. Grundmann. Hospital networks and the dispersal of hospital-acquired pathogens by patient transfer. *PLoS ONE*, 7(4):e35002, apr 2012. doi: 10.1371/journal.pone.0035002. Cited on p. 127.

[4] T. Donker, T. Smieszek, K. L. Henderson, A. P. Johnson, A. S. Walker, and J. V. Robotham. Measuring distance through dense weighted networks: The case of hospital-associated pathogens. *PLOS Computational Biology*, 13(8):e1005622, aug 2017. doi: 10.1371/journal.pcbi.1005622. Cited on p. 127.

[5] ECDC/EMEA. The bacterial challenge: time to react. A call to narrow the gap between multidrug-resistant bacteria in the EU and the development of new antibacterial agents., 2009. Cited on p. 127.

[6] B. Y. Lee, S. M. McGlone, Y. Song, T. R. Avery, S. Eubank, C. C. Chang, R. R. Bailey, D. K. Wagener, D. S. Burke, R. Platt, and S. S. Huang. Social network analysis of patient sharing among hospitals in Orange County, California. *American Journal of Public Health*, 101(4):707–713, 2011. doi: 10.2105/AJPH.2010.202754. Cited on p. 127.

[7] B. Y. Lee, S. M. Bartsch, K. F. Wong, S. L. Yilmaz, T. R. Avery, A. Singh, Y. Song, D. S. Kim, S. T. Brown, M. A. Potter, R. Platt, and S. S. Huang. Simulation shows hospitals that cooperate on infection control obtain better results than hospitals acting alone. *Health Affairs*, 31(10): 2295–2303, oct 2012. doi: 10.1377/hlthaff.2011.0992. Cited on p. 127.

[8] A. Y. Peleg and D. C. Hooper. Hospital-acquired infections due to gram-negative bacteria. *New England Journal of Medicine*, 362(19):1804–1813, 2010. doi: 10.1056/NEJMra0904124. Cited on p. 127.

[9] M. J. Piotrowska and K. Sakowski. Analysis of the AOK Lower Saxony hospitalisation records data (years 2008 – 2015). *arXiv*, 1903.04701v1, 2019. Cited on pp. 129 and 132.

Analiza danych AOK Plus i zbudowanej na ich podstawie sieci szpitalnej

A. Lonc, M.J. Piotrowska, K. Sakowski

Streszczenie W artykule analizujemy dane o hospitalizacjach pacjentów dostarczone przez jednego z ubezpieczycieli działajacego na terenie dwóch landów niemieckich: Saksonii i Turyngii. Dostarczona baza danych obejmuje lata 2010–2016. Ze wzgledu na brak informacji dotyczacych transferów pacjentów miedzy szpitalami, przeprowadzona została niezbedna analiza pozwalajaca na wyekstrahowanie takich informacji.Wefekcie stworzono siec złozona z placówek opieki zdrowotnej, konieczna do symulowania rozprzestrzeniania sie wielolekoopornych bakterii szpitalnych. Taka siec pozwoli na badanie skutecznosci działan majacych na celu kontrolowanie i zwalczanie tego typu zakazen. W niniejszym artykule zostały przeanalizowane dane z udostepnionej bazy danych. Ponadto poddano analizie strukture otrzymanej sieci transferów miedzyszpitalnych.

Klasyfikacja tematyczna AMS (2010): 62-07; 92C42.

Słowa kluczowe: analiza danych opieki zdrowotnej, nakładające się dane, sieć międzyszpitalna, wielolekooporne bakterie.

Agata Lonc is a master's Mathematics student at the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw. Interested in applications of mathematics, especially in biology-related problems.

Monika J. Piotrowska, got habilitation in mathematics at the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw. Interested in mathematical and computational modelling of various biological and medical phenomena including: tumour growth, angiogenesis (blood vessels formation), multicellular spheroids growth and necrotic core formation, tumour-immune system interactions, transmissions of multidrug-resistant bacteriae within hospital networks, investigation and optimization of (combined) anticancer therapies and infection control strategies.

Konrad Sakowski, got his PhD in mathematics. Interested in interdisciplinary research related to applied mathematics, numerical modelling of semiconductor devices, semiconductor physics, numerical methods for elliptic differential equations, numerical methods for nonlinear equations.

Agata Lonc
University of Warsaw
Institute of Applied Mathematics and Mechanics
Banacha 2, 02-097 Warsaw
E-mail: alonc96@gmail.com

Monika Joanna Piotrowska
University of Warsaw
Institute of Applied Mathematics and Mechanics
Banacha 2, 02-097 Warsaw
E-mail: monika@mimuw.edu.pl

Konrad Sakowski
University of Warsaw
Institute of Applied Mathematics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
Polish Academy of Sciences
Institute of High Pressure Physics
Sokołowska 29/37, 01-142 Warsaw, Poland
Kyushu University
Research Institute for Applied Mechanics
Kasuga, Fukuoka, 816-8580, Japan
E-mail: konrad@mimuw.edu.pl

Communicated by: Urszula Foryś

(Received: 13th of June 2019; revised: 8th of July 2019)