FOLIA MEDICA CRACOVIENSIA
Vol. LIX, 1, 2019: 89–100
PL ISSN 0015-5616
DOI: 10.24425/fmc.2019.128028

Comparison of the efficacy of available statistical methods for prediction of the hospitalizations number: proof of concept and validation based on the analysis of Polish National Health Fund data in the years 2009–2017

Norbert Tuśnio[1], Jakub Fichna[2], Przemysław Nowakowski[3]

[1]The Main School of Fire Service, Warsaw, Poland
[2]Department of Biochemistry, Medical University of Lodz, Lodz, Poland
[3]Innovation & Technology Transfer Center, Medical University of Lodz, Lodz, Poland

Corresponding author: Norbert Tuśnio, PhD, Eng.
Faculty of Fire Safety Engineering, The Main School of Fire Service
ul. J. Słowackiego 52/54, 01-629 Warszawa, Poland
Phone: +48 602 896 982; E-mail: ntusnio@sgsp.edu.pl

Abstract: The aim of the study was to choose and validate the tool(s) to predict the number of hospitalized patients by testing three predictive algorithms: a linear regression model, Auto-Regressive Moving Average (ARMA) model, and Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH) model. The study used data from the collection of data on inflammatory bowel diseases (IBD) from the public database of the National Health Fund for the years 2009–2017, data recalculation taking into account the population of provinces and the country in particular years, and prediction making for the number of patients who would require hospitalization in 2017. The anticipated numbers were compared with real data and percentage prediction errors were calculated. Results of prediction for 2017 indicated the number of hospitalizations for Crohn's disease (CD) and ulcerative colitis (UC) at 17 and 16 respectively per 100,000 persons and 72 per 100,000 persons for all IBD cases. The actual outcomes were 21 for both CD and UC (81% and 75% accuracy of prediction, respectively), and 99 for all IBD cases (73% accuracy). The prediction results do not differ significantly from the actual outcome, this means that the prediction tool (in the form of a linear regression) actually gives good results. Our study showed that the newly developed tool may be used to predict with good enough accuracy the number of patients hospitalized due to IBD in order to organize appropriate therapeutic resources.

Key words: inflammatory bowel disease, Crohn's disease, ulcerative colitis, small intestine, large intestine, epidemiology, medical statistics.

Introduction

Inflammatory bowel diseases (IBD) is a group of inflammatory conditions of the colon and the small intestine, which include Crohn's disease (CD) and ulcerative colitis (UC). A characteristic feature of inflammatory lesions in CD is their focal or segmental nature which means that the segments affected by disease are usually separated by completely healthy tissue. The second feature that differentiates CD from UC is that the inflammatory process covers not only the mucous membrane, but the entire thickness of the intestinal wall, from the mucosa to the serous membrane. Ulcerative colitis is a chronic non-specific inflammation of the large intestine, occurring most often in the rectum, but which can spread continuously to its further sections. In UC, the inflammation is usually limited to the mucous membrane, while the remaining layers of the intestinal wall (muscular and serous membrane) remain unchanged.

The main goals in anti-IBD therapy include management of inflammation and abdominal pain as well as other clinical symptoms such as diarrhoea and rectal bleeding. These can be attained through the use of non-steroid anti-inflammatory drugs, corticosteroids, or biologics (e.g. anti-TNFα or anti-α4β7 integrin antibodies), depending on the stage and the severity of the disease. Alleviation of symptoms and maintenance of remission require a long-term treatment which in most cases is associated with the development of adverse side effects needing additional therapy; moreover, for some, like biologic drugs (e.g. infliximab), there is an unsatisfactory number of responders [1, 2].

The number of patients with IBD, both in Poland and worldwide is growing rapidly in recent years, and therefore becoming an increasing medical, social and economic problem. For example, in 2015 there were about 3 million (1.3% of adults) new IBD cases in the US alone [3]. In line, in their systematic review Molodecky et al. estimated that the highest annual incidence of CD and UC in Europe was reaching up to 12.7 and 24.3 per 100,000 persons, respectively [4]. Moreover, in time-trend analyses, 75% of CD studies and 60% of UC studies had an increasing incidence of statistical significance ($P < 0.5$). Because of its chronic nature and our poor understanding of its pathology, what makes available therapies only partly effective, IBD is expensive in terms of treatment. The direct costs in Poland in 2013 amounted to over 98 million PLN (25 million EUR), and the annual value of medicines reimbursement was equally considerable [5]. Moreover, state budget has been economically burdened by hospitalization of patients due to IBD-related diseases and by the costs covering the sick leave. Unfortunately, there are currently no tools to facilitate epidemiologic data analysis and predictions for the future, and these could be very useful for the healthcare system.

In this study we attempted at validation of selected statistical models for determination of the trends of IBD incidence in Poland with the use of the public data from the National Health Fund. Three algorithms were selected as potentially useful for the prediction of the number of patients who will require hospitalization:

- Linear Regression method, which assumes that the time trend over time is linear. This assumption was adopted due to the growing IBD disease trend.
- Auto-Regressive Moving Average (ARMA) model, which provides a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the autoregression and the second for the moving average.
- Generalized Auto-Regressive Conditional Heteroskedasticity (GARCH) model, where ARCH is a statistical model for time series data that describes the variance of the current error term or innovation as a function of the actual sizes of the previous time periods' error terms.

The study used the collection of data on IBD from the public database of the National Health Fund from the years 2009–2017, data recalculation taking into account the population of provinces and the country in particular years, and making predictions for the number of patients who would require hospitalization in 2017. The expected indications were compared with real data and percentage prediction errors were calculated.

Material and Methods

A public database of the National Health Fund was used in this study [6]. Specifically, a trial version of the data browser on services provided in the Diagnosis-Related Groups system in 2009–2017 released by NHF was employed. The browser allows generation of cross-section analyses for a selected group based on, among other parameters, the number of occurrences, the median length of hospitalization and the average value of group implemented. The results are presented in both the tabular version and the graphical form.

Importantly, data for individual provinces and for the whole country, broken down into specific cases of IBD classified according to IDC-10 (K50 for CD and K51 for UC) were taken into consideration (Table 1).

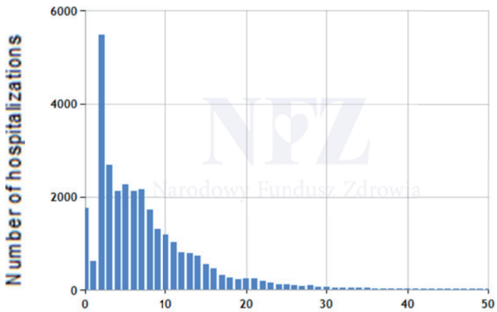
General information found in the NHF database consists of — among others — the number of patients and hospitalizations together with the rehospitalization rate, length of hospitalization — median and dominant (number of days) and medical costs incurred by the fund. It is possible to determine the number of hospitalizations and the median duration of stay in individual provinces, with the division into gender. The database also concerns the number of hospitalizations and length of stay based on the patient's admission and discharge mode. The analysis used data on the number of patients' hospitalizations for relevant disease entities. Information on the population (in a given year and province) was taken from data provided by the Statistics Poland government.

Table 1. Diagnosis-Related Groups, categories and types of IBD selected for analysis.

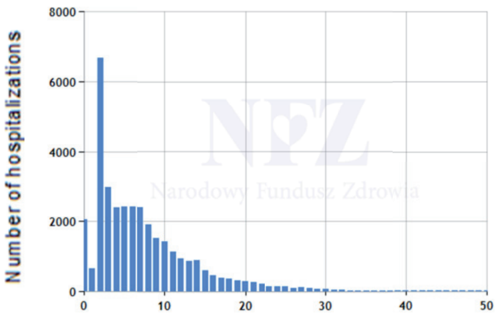
DRG	Type of data	Years
F56	Inflammatory bowel disease >17 years of age	2009–2013
F57	Inflammatory bowel disease <18 yrs.	2009–2013
F58	Inflammatory bowel disease	2014–2015
F58E	Inflammatory bowel disease >65 years of age	2016–2017
F58F	Inflammatory bowel disease <66 yrs.	2016–2017

Group	Category	Type of disease
K50	K50.0	Crohn's disease of the small intestine
	K50.1	Crohn's disease of the large intestine
	K50.8	Other forms of Crohn's disease
	K50.9	Crohn's disease, undetermined
K51	K51.0	Ulcerative (chronic) inflammation of the small and large intestine
	K51.1	Ulcerative (chronic) ileitis
	K51.2	Ulcerative (chronic) inflammation of the handpiece
	K51.3	Ulcerative (chronic) inflammation of the handpiece and sigmoid
	K51.5	Mucosal inflammation of the handpiece and colon
	K51.9	Ulcerative colitis, unspecified

Of note, in the NHF database, the numbers of Diagnosis-Related Groups F58 occur entirely in the years 2014–2015, whereas in 2009–2013 data for IBD occur divided into two parts in relation to age (F56 — above 17 years old and F57 — under 18). To compare F58 with data from 2009–2013, it was necessary to combine data F56 and F57. Similarly, the length of stay histograms are available in F56 and F57, broken down by age, and in F58 without such a division (Fig. 1). Similar situation occurs for the years 2016–2017, where F58 data were divided into F58E (>65 years old) and F58F (<66 years old).



Length of stay in the hospital in 2014 [days]



Length of stay in the hospital in 2015 [days]

Fig. 1. The number of hospitalizations due to inflammatory bowel disease and the length of hospital stay in year 2014 (left) and 2015 (right). Source: https://prog.nfz.gov.pl/app-jgp/AnalizaPrzekrojowa.aspx.

Three categories of data have been prepared:

1. The number of IBD hospitalizations in 2009–2016, broken down by provinces
2. The number of CD hospitalizations in 2009–2016, for the whole country
3. The number of UC hospitalizations in 2009–2016, for the whole country

Three prediction methods in the time series analysis were selected. The analysis of time series was linked to the forecasting methods. The aim of the time series analysis was to build a model of a certain phenomenon / process (population incidence on IBD) based on the observed changes in time of certain measurable quantities describing this process (number of hospitalizations). Using the obtained model, the predictions (exploration) of the series of incidents (IBD hospitalization rate) and its components (CD and UD hospitalization rate) were made. The trend of the phenomenon was determined. The general direction of the phenomenon (systematic changes that the phenomenon undergoes) has been examined. It was checked whether this trend is linear or non-linear. The periodic fluctuations were not checked due to the lack of seasonal data on a monthly basis. The adopted period was one year.

Results

The changes in the number of hospitalizations per 100,000 persons for individual provinces in 2009–2016 are presented in Table 2.

Table 2. Number of hospitalizations per 100,000 population due to IBD in individual provinces in 2009–2017.

Province / Year	2009	2010	2011	2012	2013	2014	2015	2016	2017
Lower Silesia	78	68	66	66	68	88	103	51	118
Kujawy-Pomerania	122	122	118	115	118	149	167	81	152
Lublin province	56	49	58	59	48	68	88	47	108
Lubusz	29	27	19	15	18	30	36	24	59
Lodz province	54	52	53	52	47	73	88	43	102
Lesser Poland	72	58	64	64	67	90	104	48	101
Masovia	55	60	74	83	81	112	126	65	139
Opole province	26	27	31	32	29	64	69	31	63
Podkarpackie	43	40	45	49	53	76	85	48	109
Podlaskie	73	61	72	74	65	82	84	41	86
Pomerania	48	46	49	36	37	64	80	36	72
Silesia	49	48	54	64	47	68	86	41	87
Swietokrzyskie	36	34	36	45	46	79	95	37	84
Warmia-Masuria	28	29	28	29	27	53	47	21	51
Greater Poland	36	37	35	39	38	49	51	31	67
West Pomerania	35	41	41	48	41	53	68	38	76

Figure 2 shows the provinces with the largest hospitalization rate due to IBD.

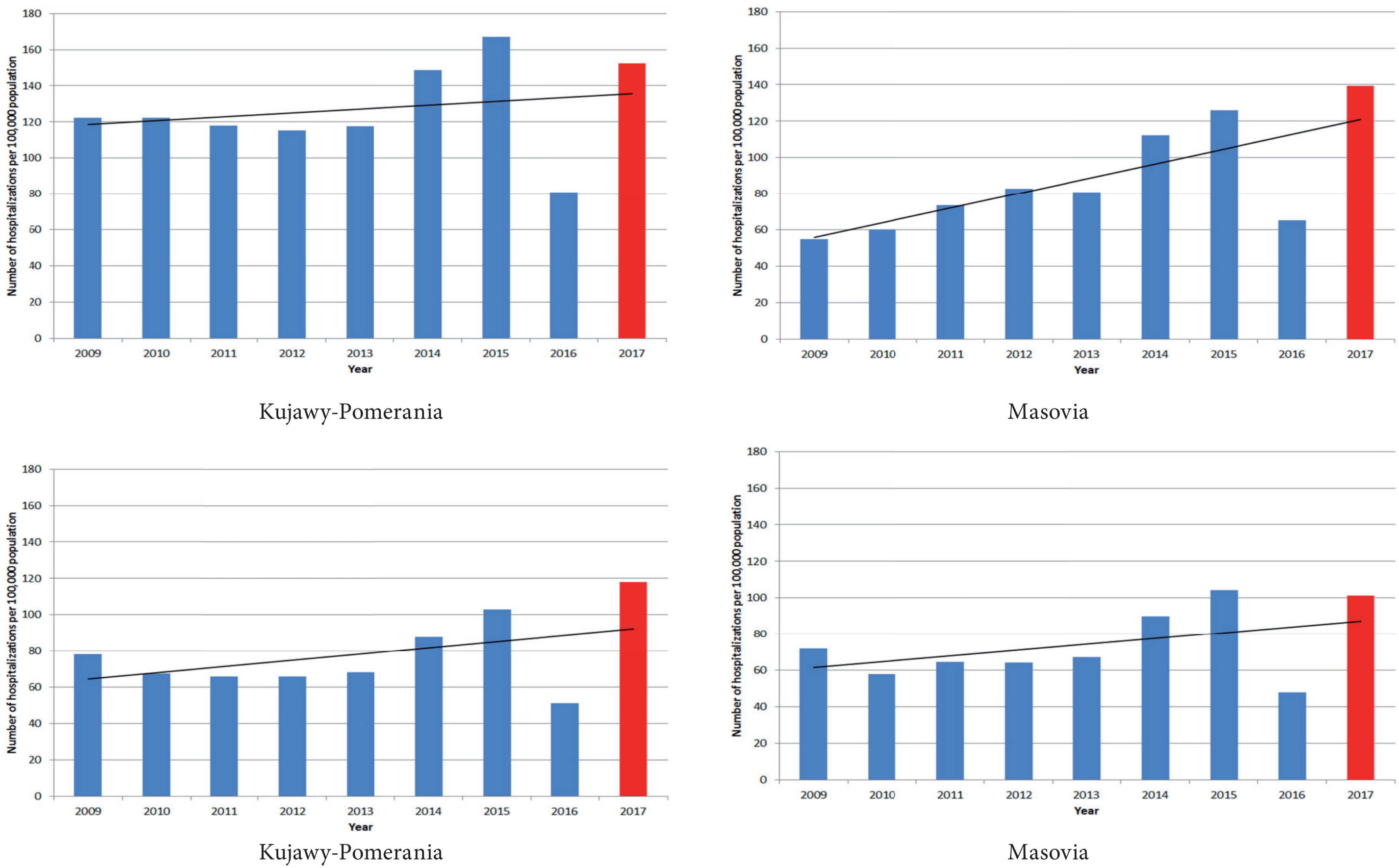


Fig. 2. A set of charts illustrating the number of hospitalizations per 100,000 population due to IBD in individual provinces. Four provinces with the largest hospitalization rate were selected. The trend lines are marked. For 2017, actual data are presented.

Expected and real values of the number of hospitalizations due to IBD in 2017 are given in Table 3.

Table 3. Expected and real numbers of hospitalization cases per 100,000 population due to IBD in individual provinces in Poland in 2017 with prediction errors (K50 and K51 cases are possible to compute only for the whole country).

Province	Cases in 2017	Linear Regression	LR error	ARMA model	ARMA error	GARCH model	GARCH error
Whole Poland, K50	21	17	19%	16	23%	15	29%
Whole Poland, K51	21	16	25%	23	11%	18	15%
Whole Poland, IBD	99	72	27%	60	39%	60	39%
Lower Silesia	118	77	35%	77	35%	72	39%
Kujawy-Pomerania	152	126	18%	138	10%	125	18%
Lublin province	108	67	38%	61	44%	59	46%
Lubusz	59	27	54%	25	58%	25	58%
Lodz province	102	66	35%	58	43%	54	47%
Lesser Poland	101	79	22%	72	29%	69	32%
Masovia	139	110	21%	74	47%	82	41%
Opole province	63	57	10%	36	44%	37	42%
Podkarpackie	109	74	33%	51	53%	54	51%
Podlaskie	86	65	25%	86	1%	70	19%
Pomerania	72	57	22%	49	32%	46	37%
Silesia	87	66	24%	60	31%	55	36%
Swietokrzyskie	84	75	11%	46	45%	48	43%
Warmia-Masuria	51	39	23%	31	38%	33	35%
Greater Poland	67	43	36%	39	43%	39	42%
West Pomerania	76	55	28%	45	40%	45	41%

Following prediction using linear regression, it was calculated that in 2017 in Poland, for both K50 and K51 disease, the number of hospitalizations may occur in respectively 17 and 16 cases out of 100,000 population (Fig. 3).

Analyzes carried out in the area of 16 provinces showed that the average accuracy is: for the linear regression model — 73%, ARMA — 63% and GARCH — 61%. At the same time, the linear regression model was the best for the Opole province (10% error) and Swietokrzyskie (11%), the ARMA model for Podlaskie (1%) and Kujawy-Pomerania (10%), and the GARCH model for Kujawy-Pomerania (18%) and Podlaskie (19%). Based on the whole country, we noticed that good data accuracy for K51 cases is generated by ARMA model (11% error) and GARCH model (15%).

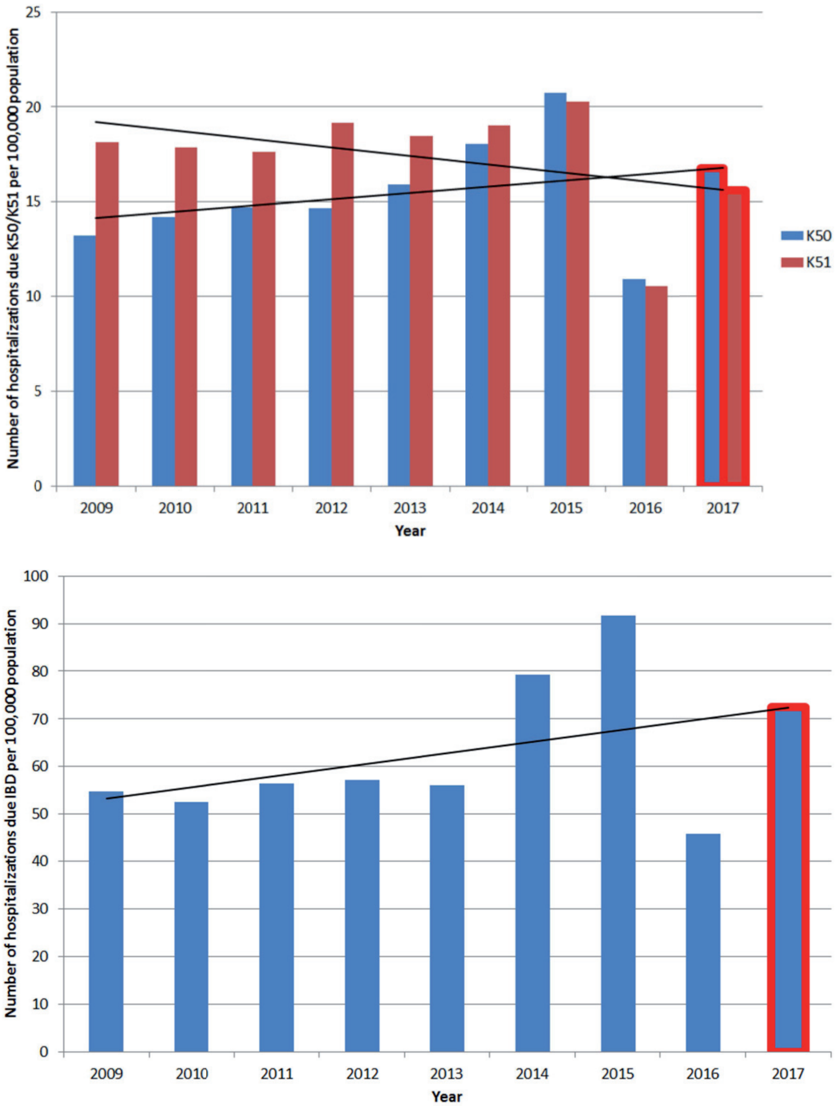


Fig. 3. Upward trend for cases K50 and K51 (left) and all IBD cases (right) in 2009–2016 with prediction of number of hospitalizations per 100,000 population in 2017 (linear regression).

Discussion

It is estimated that in Europe, approximately 0.3% of the population (2.5–3 million people) suffer from IBD. Morbidity in CD (K50 in ICD-10) varies from 1.5 to 213 cases per 100,000 people, and in the case of UC (K51 in ICD-10) from 2.4 to 294 cases per 100,000 people [7].

In Poland, no precise epidemiological studies have been conducted so far. The National Register of Crohn's Diseases [8] contains IBD data for the years 2005–2018. It is reported therein that the number of registered centers is 95, the number of patients in the register is 6324, but there is no division into years or provinces. The assessment of IBD incidence based on national hospital registries showed that in 2007, 4.2 new cases of K50 and 12.8 new cases of K51 per 100,000 population were found [9]. Noteworthy, in the years 2004–2007 the number of new hospitalized cases of K50 significantly increased, especially in the youngest age groups. However, the number of new cases of K51 decreased, mainly in older age groups [10]. Starting from 2009, data on the number of patients and hospitalization in Poland are provided by the National Health Fund. In the group of IBD, in the years 2009–2013 the number of patients remained stable: for minors, it did not exceed 3,000 per year, and the average number of adult patients did not exceed 15,000.

Current analysis used a public database of the Polish National Health Fund with data provided for the years 2009–2017. Of the three methods of prediction tested in the study (ARMA, GARCH, and linear regression), GARCH and ARMA envisaged the most optimistic scenario for Poland in 2017 (with 60.2 and 60.3 hospitalizations, respectively) and linear regression gave the most pessimistic outcome with 72 hospitalized IBD patients per 100,000. However, it was the latter that was chosen as the most accurate among the tools tested (the accuracy of 81, 75 and 73% for CD, UC and IBD in general, respectively). The results suggest that the linear regression method could be now used to estimate the number of hospitalizations for IBD as well as other causes. Healthcare administrators around the world are trying to reduce the cost of care, while improving its quality. Anticipating outcomes, such as the number of hospitalizations (and their length), with our tool may contribute to making better, both clinical and administrative decisions.

The analysis also proved that in 2009–2015 in all Polish provinces there was an increase in the number of hospitalizations due to IBD. Noteworthy, in 2016 the number of cases dropped dramatically by half, so that in 2017 it could return to the state of 2015. One of the reasons for this could be the start from 2016 of budesonide reimbursement in the form of prolonged-release tablets. In 2017, the medicine was discontinued for pediatric patients. The explanation of this phenomenon could also be seen in a brief change in classification of IBD cases.

There are some limitations of the validated tool that need further attention. It should be taken into account that NHF data concentrate on the numbers of hospitalizations and therapies, and these may refer to the same patient. Moreover, the number of patients does not concern exclusively the first admission. One should thus be careful with the interpretation of the data obtained, as the number of hospitalizations does not coincide with the incidence rate. However, future studies are warranted, as the tool may help explore different aspects of IBD epidemiology, such as environmental or the age-specific effects of air pollution exposure on IBD risk [11].

Conclusions

1. Based on the analysis of medical data obtained from the National Health Fund, predictive models have been validated to support the process of planning hospital treatment for IBD patients.
2. The linear regression was chosen as the best method from the tools tested as part of the work. This method proved to be used to estimate the number of hospitalizations for CD with an accuracy of 81%, UC 75%, and IBD 73%.
3. The authors observed an anomaly in NHF data which consisted in a decrease in the number of hospitalizations due to IBD in 2016 compared to 2015 by more than a half. The number of hospitalizations increased again in 2017. The explanation of this phenomenon should be seen in a brief change in classification of cases of IBD, or the increase in the availability or reimbursement of a specific drug.

Acknowledgements

Supported by the Medical University of Lodz (#503/1-156-04/503-11-001 to JF) and National Science Center (#UMO-2017/25/B/NZ5/02848 to JF).

Conflict of interest

None declared.

References

1. *Sobczak M., Fabisiak A., Murawska N., et al.*: Current overview of extrinsic and intrinsic factors in etiology and progression of inflammatory bowel diseases. Pharmacol Rep. 2014; 66 (5): 766–775.
2. *Fichna J. (ed.)*: Introduction to gastrointestinal diseases. Vol. 1, Springer International Publishing AG; 2017.

3. *Dahlhamer J.M., Zammitti E.P., Ward B.W., Wheaton A.G., Croft J.B.*: Prevalence of inflammatory bowel disease among adults aged ≥18 years –— United States, 2015. Morb Mortal Wkly Rep. 2016; 65 (42): 1166–1169.
4. *Molodecky N.A., Soon I.S., Rabi D.M., Ghali W.A., Ferris M., Chernoff G., Benchimol E.I., Panaccione R., Ghosh S., Barkema H.W., Kaplan G.G.*: Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 2012; 142 (1): 46–54.
5. *Moćko P., Kawalec P., Pilc A.*: Inflammatory bowel diseases (IBD) as public health problem — review of the literature. Medycyna Rodzinna. 2016; 19 (4): 244–254, http://www.medrodzinna.pl/wp-content/uploads/2017/02/mr_2016_244-254.pdf.
6. https://prog.nfz.gov.pl/app-jgp/AnalizaPrzekrojowa.aspx (Access: 05.09.2018).
7. *Rydzewska G.M., Głuszek-Osuch M., et al.*: Epidemiological and social report "Inflammatory bowel disease — an opponent growing in strength". Warsaw 2016, http://www.ippez.pl/wp-content/uploads/2016/12/Raport_choroba_jelit_bez-jelita.pdf (Access: 05.09.2018).
8. http://rejestr.chorobacrohna.pl/index/rejestr/wyniki (Access: 05.09.2018).
9. *Wejman J., Bartnik W.*: Atlas kliniczno-patologiczny nieswoistych chorób zapalnych jelit [Clinical-pathological atlas of inflammatory bowel disease], Poznań, Termedia; 2011.
10. *Jakubowski A., Bartnik W., Kraszewska E., et al.*: Trends of hospitalization due to inflammatory bowel diseases in Poland. Gastroenterologia Polska. 2010; 17: 50.
11. *Ananthakrishnan A.N., McGinley E.L., Binion D.G., Saeian K.*: Ambient air pollution correlates with hospitalizations for inflammatory bowel disease: An ecologic analysis. Inflammatory Bowel Diseases. 2011; 17 (5): 1138–1145.