BMJ Open

Community determinants of COPD exacerbations in elderly patients in Poland: protocol for a retrospective Big Data observational cohort study

Izabela Zakowska,[1] Katarzyna Kosiek,[2] Anna Kowalczyk,[1] Jacek Grabowski,[1] Maciek Godycki-Cwirko[1,2]

To cite: Zakowska I, Kosiek K, Kowalczyk A, *et al*. Community determinants of COPD exacerbations in elderly patients in Poland: protocol for a retrospective Big Data observational cohort study. *BMJ Open* 2019;**9**:e030524. doi:10.1136/bmjopen-2019-030524

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-030524).

Received 19 March 2019
Revised 6 June 2019
Accepted 7 June 2019

[1]Centre for Family and Community Medicine, Medical University of Lodz, Lodz, Poland
[2]Division of Public Health, Faculty of Medical Sciences, Medical University of Lodz, Lodz, Poland

Correspondence to
Dr Maciek Godycki-Cwirko; maciekgc@uni.lodz.pl

ABSTRACT

Introduction Analyses of large sets of electronic health-related data (Big Data), including local community indicators, may improve knowledge of the outcomes of chronic diseases among patients and healthcare systems. Our study will estimate the prevalence of chronic obstructive pulmonary disease (COPD) and its exacerbations in elderly patients in the Lodz region, Poland; it will also evaluate local community factors potentially associated with disease exacerbations and rank local communities according to health and local community indicators.

Methods and analysis Local community factors, including medical/health, socioeconomic and environmental values potentially associated with COPD exacerbations will be identified. A retrospective analysis of a cohort of about half a million people 65 years old and older, living in local communities of the Lodz region in 2016 will be performed. Relevant data will be extracted from databases, including those of the National Health Fund, Tax Office and National Statistics Centre. This cross-sectional study will include data for a 1 year period, from 1 January until 31 December 2016. The data will first be checked for quality, cleaned and analysed using data mining techniques, and then multilevel logistic regression will be used to discover the community determinants of COPD exacerbations.

Ethics and dissemination The study protocol has been approved by the Bioethical Committee of Medical University of Lodz (RNN/248/18/KE, 10 July 2018). Our findings will be published in peer-reviewed journals and reports.

Strengths and limitations of this study

- This will be a pioneering study in Poland to explore combined big sets of health and local community data extracted from the electronic databases.
- The results will be visualised as maps.
- The main limitations relate to the specificity and sensitivity of the chronic obstructive pulmonary disease coding, gaps in databases, short period of observation.
- The other limitation is the age limit of the patients.
- Additional limitations will be addressed when the study is completed.

INTRODUCTION

The world is currently seeing a growth in the incidence of chronic obstructive pulmonary disease (COPD), a non-reversible lung condition characterised by shortness of breath, chronic cough with sputum production, emphysema and systemic pulmonary inflammation.[1] Its worldwide prevalence in adults has been estimated to be 10%[2] and is among the leading causes of mortality and morbidity worldwide.[3]

The burden of the disease and its exacerbations has been studied globally from different perspectives.[4–10] The prevalence data are limited for Poland, where our study is located. Although the scope of the problem is well recognised worldwide, its impact seems to be poorly reflected in the Polish research data, so little is known of the exact prevalence of COPD in Poland.

Most studies place a strong focus on clinical patterns, but other also consider the socioeconomic and environmental local community status of patients.[11 12]

The community context has been identified as an important determinant of health outcomes.[13] In the proposed study, the term *community* will be used to refer to a local neighbourhood and its inhabitants within certain geographic construct boundaries, and is regarded as being synonymous with a *gmina* in Poland, defined by GUS (Statistics Poland) as the basic unit of the three-tier territorial division of the country. The present analysis covers the whole of the Lodz voivodship, 1 of 16 in the country; it therefore comprises 177 *gminas*, that is, communities.

Certain predictors of exacerbations in COPD are well known[14–17] while some

community and regional factors remain under examination. Although a systematic review of the broad variety of factors to which patients are exposed in their living area has already been conducted by Pleasants *et al*[12] the amount of data generated and collected routinely has increased significantly in the last decade, as has our ability to analyse and interpret it, especially in medicine. For example, a number of such Big Data studies have been performed using the Chinese healthcare system, including large populations and multiple structured and unstructured data sources, with the aim of improving decision-making.[18] It has been proposed that Big Data extracted by combining databases from various sources, including medical records, clinical and diagnostic results, patient medication records and medicine purchases, as well as data concerning costs, diagnostic costs and sports habits, could be used to improve the decision-making process, and thus influence patient health and quality of life.[19] Further analyses of large sets of electronic health records, including indicators among local communities, may improve knowledge about the outcomes of chronic diseases among patients.

The members of the patient cohort will be COPD 'labelled' patients who had been identified by the healthcare system and assigned the code J44 from International Classification of Diseases (ICD-10). We are aware of the limits of this approach, and that some COPD patients may not have been coded, and hence not included in the group, but our area of interest is the healthcare system dataset reflected by coding. A more detailed picture, and a more correct analysis, can be obtained by follow-up studies with more precise coding being applied and verified in the future.

AIMS

1. To estimate the prevalence of J44 coded COPD cases in elderly patients living in the Lodz voivodship, Poland.
2. To evaluate local community factors potentially associated with disease exacerbations in this population.
3. To rank the *gminas* in the region according to health and local community indicators.

Our study is the pioneering of this kind in Poland, the purpose of which is to provide evidence for the potential role of local community factors in the health outcomes of the older population.

METHODS AND ANALYSIS

Study design

This will be a retrospective cohort study involving approximately half a million patients aged 65 years and older living in the Lodz voivodship, Poland, including patients with COPD and its exacerbations. This study will include data for a 1 year period, from 1 January until 31 December 2016. The study reported in the manuscript (data extraction and analysis) will take place from 10 July 2018 until the 29 February 2020.

Data source

Data will be obtained from Big Data databases, such as the electronic health records of patients from the National Health Service (NFZ), US (Tax Office) and GUS. Depersonalised data will be loaded and subjected to quality control and cleaning.

Individual patient data will be anonymised and assigned to the local communities which are the basic units of our analysis. We will collect three categories of data: (1) disease-related data, (2) healthcare services use-related data and (3) data relevant for selected local community indicators from restricted and publically available databases and repositories with limited and unlimited access. Patient consent is not needed since we will not collect any personally sensitive data.

Individual patient data will be matched by patient identifier within a single database. Individual data will not be matched between databases. Data will be matched on the local community level, and these matched local community datasets will be the units of our analyses.

The scope of associations between the well-known patient-level risk factors and triggers of exacerbations of COPD, including local community factors, will be identified with a literature review. Local community status factors will then be listed and selected with brainstorming (BS) and focus group discussion (FGD), with the participation of researchers, experts and decision makers in the field of medicine and public health based on the methods described by Osborn[20 21] and Kitzinger,[22] respectively. During the BS and FGD, experts will select and classify factors into three main groups at *gmina* level, according to the Remington and Catlin methodology, as follows: (1) health factors, (2) socioeconomic factors and (3) community environmental factors.[23] The group of experts will decide on the outline/framework of available databases and the collection of Big Data sets, and this outline will be filled with depersonalised data.

Population

Residents of the Lodz voivodship aged 65 years and over between 1 January and 31 December 2016 will be identified from NFZ electronic health record systems, US and GUS using a residence code and assigned to a local community (*gmina*). Patients with COPD will be identified by the ICD-10 code J44 in their medical records; exacerbations will be defined as cases 'hospitalised with the J44 code as a main reason for admission'.

Study variables

This study will reveal a possible association between COPD exacerbations in elderly and local community factors: demographic, healthcare use, social, economic and environmental factors. It will take into account patient demographic and characteristics, including age, gender, residence code, as well as the number of visits to the general practitioner (GP) in 2016, number of GP visits due to COPD in 2016, hospitalisation, hospitalisation with the J44 code as a main reason for admission,

number of deaths, costs of care, patient income per *gmina* and number of GPs per *gmina*.

Patient and public involvement

The priorities, experience and preferences of the patients and other health professionals were identified by individual interviews, BS and FGD technique in an earlier work.[24] The suggestions regarding COPD determinants were discussed and will be taken into consideration in the planned research.

Patients were not directly involved in the design of this study. As this is a protocol for a retrospective cohort study and no participant recruitment will take place, their involvement in the recruitment and dissemination of findings was not applicable.

The results of the study will be available for the public through internet and local media.

Statistical analysis

The data obtained from the Big Data databases will be used to characterise patient health status, patient status related to the healthcare system, and the characteristics of the local community. Descriptive statistics for the total group, and the presence of COPD exacerbation will be calculated, aggregated at *gmina* level and categorised. Health characteristics and health outcomes will be aggregated by *gmina* in the Lodz voivodship, standardised and categorised.

Data mining techniques will be used to examine the relationships between patients and *gmina*; on the basis of which, indexes for each *gmina* will be calculated and normalised. Cross-sectional, case–control multilevel multivariable logistic regression models (adjusted by demographics and health factors) will be used to test variables significantly associated with exacerbations of COPD. Health outcomes, such as the numbers of non-hospitalised patients awarded the code J44 and the numbers of hospitalised patients with exacerbation within the *gmina*, will be categorised as a dependent variable in the regression analysis.

The obtained data will be visualised on a map of 177 *gminas* located in the Lodz region. Local community factors significantly associated with exacerbations of COPD will be shown on the *gmina* map according to each statistically significant factor.[23] The occurrence of exacerbations of COPD 65+ patients will be shown at *gmina* level using colours.

Complex variables will be calculated for each group of determinants using the weights from the BS and FGD and literature review results. Additionally, the obtained complex variables related to health outcome and health determinants (health behaviours; clinical care; social and economic factors; and physical environment) for patients aged 65 years and above with COPD exacerbations will also be visualised on the maps. It is planned therefore to obtain five maps, one for each of the five complex variables, illustrating the Lodz voivodship in terms of COPD exacerbations resolved at the *gmina* level.

All the data will be analysed using the SAS V.9.4 statistical package, MLwiN V.2.24 and STATISTICA V.13.1.

DISCUSSION

Our proposed methods will enable quantitative findings to be obtained that can be used to better understand the factors associated with exacerbations of COPD in communities.

The community-level contribution identified in the findings might be useful for future planning and resource allocation. This will be particularly useful if the obtained body of data is regularly updated by ongoing Big Data analysis of the *gminas* and healthcare systems.

Combining community and medical data can allow recommendations to be prepared for improving the quality of patient life in the local community.

This will be a pioneering study in Poland exploring combined sets of blinded health and local community Big Data, extracted from the electronic databases of health-related records. The results will be visualised as maps.

The main limitations relate to the specificity and sensitivity of the COPD coding, gaps in databases and short period of observation. We will select our study population based on the codes used by the national health service. We are aware of the bias related to this approach, such as limited code sets, mistaken coding and errors related to the coding within the public datasets and repositories. Another limitation is fact that the 177 gminas were not randomly selected and were chosen based on their location within the Lodz voivodship.

Additional limitations and bias may be related to incompleteness and inaccuracy of data in databases; some variables of potential interest might not be available, as well as indicator selection might not be complete. The advantage is an ability to set a pilot framework for study disease in real world community environment.

Ethics and dissemination

Our findings will be published in peer-reviewed journals and reports. Recommendations will disseminated to key stakeholders including local leaders, decision makers, managers of prevention programmes and local community media.

Acknowledgements Special thanks to mgr. Edward Lowczowski for English language corrections.

Contributors The study concept and design was conceived by MG-C, IZ, KK, AK and JG. Analysis will be performed by IZ (SAS, MLwiN and STATISTICA statistical analyses) and MG-C. MG-C, IZ, KK, AK and JG prepared the first draft of the manuscript. All authors provided edits and critiqued the manuscript for intellectual content.

Funding This article was prepared within the research project no 2016/21/B/NZ7/02052 funded by Narodowe Centrum Nauki (National Science Centre Poland).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study protocol has been approved by the Bioethical Committee of Medical University of Lodz (RNN/248/18/KE, 10 July 2018).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Diaz-Guzman E, Mannino DM. Epidemiology and prevalence of chronic obstructive pulmonary disease. *Clin Chest Med* 2014;35:7–16.
2. Buist AS, McBurnie MA, Vollmer WM, *et al*. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *Lancet* 2007;370:741–50.
3. GBD 2015 Chronic Respiratory Disease Collaborators. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respir Med* 2017;5:691–706.
4. Donaldson GC, Seemungal TA, Bhowmik A, *et al*. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. *Thorax* 2002;57:847–52.
5. Flattet Y, Garin N, Serratrice J, *et al*. Determining prognosis in acute exacerbation of COPD. *Int J Chron Obstruct Pulmon Dis* 2017;12:467–75.
6. Lozano R, Naghavi M, Foreman K, *et al*. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2095–128.
7. Seemungal TA, Donaldson GC, Paul EA, *et al*. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998;157:1418–22.
8. Seemungal TA, Hurst JR, Wedzicha JA. Exacerbation rate, health status and mortality in COPD--a review of potential interventions. *Int J Chron Obstruct Pulmon Dis* 2009;4:203–23.
9. Vogelmeier CF, Criner GJ, Martinez FJ, *et al*. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *Am J Respir Crit Care Med* 2017;195:557–82.
10. Wedzicha JA, Seemungal TA. COPD exacerbations: defining their cause and prevention. *Lancet* 2007;370:786–96.
11. Grigsby M, Siddharthan T, Chowdhury MA, *et al*. Socioeconomic status and COPD among low- and middle-income countries. *Int J Chron Obstruct Pulmon Dis* 2016;11:2497–507.
12. Pleasants RA, Riley IL, Mannino DM. Defining and targeting health disparities in chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2016;11:2475–96.
13. Marmot MGB, Davey Smith G. *et al*Explanations for social inequalities in health. In: IaB A, Levine S, Tarlov AR, Chapman WD, . eds. *Society and health*. New York: Oxford University Press, 1995:172–210.
14. Cardoso J, Coelho R, Rocha C, *et al*. Prediction of severe exacerbations and mortality in COPD: the role of exacerbation history and inspiratory capacity/total lung capacity ratio. *Int J Chron Obstruct Pulmon Dis* 2018;13:1105–13.
15. Chiba H, Abe S. [The environmental risk factors for COPD--tobacco smoke, air pollution, chemicals]. *Nihon Rinsho* 2003;61:2101–6.
16. Halpin DMG, Miravitlles M, Metzdorf N, *et al*. Impact and prevention of severe exacerbations of COPD: a review of the evidence. *Int J Chronic Obstr* 2017;12:2891–908.
17. Viniol C, Vogelmeier CF. Exacerbations of COPD. *Eur Respir Rev* 2018;27:170103.
18. Zhang L, Wang H, Li Q, *et al*. Big data and medical research in China. *BMJ* 2018;360:j5910.
19. Chen P-T. Medical big data applications: Intertwined effects and effective resource allocation strategies identified through IRA-NRM analysis. *Technol Forecast Soc Change* 2018;130:150–64.
20. Osborn AF. *Applied imagination : principles and procedures of creative problem solving / by Alex F. Osborn*. 3rd edn, 1963.
21. iMindQ. What is brainstorming and how is it helpful? https://www.imindq.com/uses/brainstorming
22. Kitzinger J. Qualitative research. Introducing focus groups. *BMJ* 1995;311:299–302.
23. Remington PL, Catlin BB, Gennuso KP. The County Health Rankings: rationale and methods. *Popul Health Metr* 2015;13:11.
24. Krause J, Van Lieshout J, Klomp R, *et al*. Identifying determinants of care for tailoring implementation in chronic diseases: an evaluation of different methods. *Implement Sci* 2014;9:102.