https://doi.org/10.32394/pe.74.61

Ewa Niewiadomska[1], Małgorzata Kowalska[2]*

OVERVIEW OF STATISTICAL METHODS FOR ESTIMATING THE RELATIVE RISK OF DELAYED RESPIRATORY EFFECT RELATED TO AMBIENT AIR POLLUTION EXPOSURE

PRZEGLĄD METOD STATYSTYCZNYCH STOSOWANYCH W SZACOWANIU RYZYKA WZGLĘDNEGO OPÓŹNIONEGO EFEKTU ODDECHOWEGO W ZWIĄZKU Z NARAŻENIEM NA ZANIECZYSZCZENIA POWIETRZA ATMOSFERYCZNEGO

[1]Medical University of Silesia, Poland
Faculty of Health Sciences in Bytom, Department of Epidemiology and Biostatistics
Śląski Uniwersytet Medyczny w Katowicach, Polska,
Wydział Nauk o Zdrowiu w Bytomiu, Katedra Epidemiologii i Biostatystyki
[2]Medical University of Silesia, Poland
Faculty of Medical Sciences in Katowice, Department of Epidemiology
Śląski Uniwersytet Medyczny w Katowicach, Polska
Wydział Nauk Medycznych w Katowicach, Katedra i Zakład Epidemiologii

ABSTRACT

INTRODUCTION. . The occurrence of smog episodes and their significant impact on human health have forced research focused on risk assessment. Over the years, methods of exposure measuring have been improved, as well as statistical models necessary to the biological response estimation including the risk of incidence or death.
AIM. The aim of presented study is to review and evaluate possibilities of statistical methods of delayed respiratory health effects risk assessment related to ambient air pollution exposure.
MATERIAL AND METHODS. The review of published data was carried using the PubMed platform from 1994 to the 2020 year. Over 80 references were include in the analysis identifying general characteristics, construction of models estimating the relative risk of respiratory incidents with delayed health effect, and modelling tools available in statistical packages R, SAS, and Statistica.
RESULTS. Among various methods of health risk assessment, the Almon model, the Poisson model, and the Distributed Lag Non-Linear Models (DLNM) were most common used. Initially, the Poisson model was used, close to 60% of the cited works apply this method. The interest in the nonlinear modelling implementation has increased (34% of cited papers) in recent years. Mostly researchers used R or SAS statistical software. Usually, was calculated the relative risk of health effect related to short-term exposure (up to a week). About 75% of available papers concern measurements of relative risk in response to the concentration of pollution increase by unit=10 μg/m3. Other describe the risk associated with the exposure increasing by the interquartile range (IQR).
CONCLUSIONS. Distributed Lag Non-linear Model DLNM is classified as the statistical tool recommended by researchers due to its flexibility in defining, simplicity in interpretation, and increasingly frequent applications to environmental epidemiology.

Key words: *delayed health effect, Almon model, Poisson model, DLNM*

STRESZCZENIE

WSTĘP. Występowanie incydentów smogowych i ich znaczący wpływ na zdrowie i życie ludzi wymuszały inicjowanie badań ukierunkowanych na pomiar ryzyka. Na przestrzeni lat doskonalono metody pomiaru narażenia, a także modele statystyczne niezbędne do szacowania odpowiedzi biologicznej w postaci ryzyka zdrowotnego (zachorowania lub zgonu).

CEL PRACY. Celem pracy jest przegląd i ocena możliwości wykorzystania dostępnych metod statystycznych do szacowania ryzyka zdrowotnego w obrębie układu oddechowego pozostającego w związku z narażeniem na zanieczyszczenia powietrza i uwzględniających opóźnienie efektu zdrowotnego.
MATERIAŁ I METODY. Przeglądu dokonano z wykorzystaniem elektronicznej platformy PubMed w przedziale czasowym 1994-2020. Uwzględniono ponad 80 pozycji piśmienniczych identyfikujących ogólną charakterystykę, budowę modeli szacujących ryzyko względne incydentów oddechowych z opóźnionym efektem zdrowotnym oraz narzędzia modelowania dostępne w pakietach statystycznych R, SAS, bądź Statistica.
WYNIKI. Pośród licznie stosowanych metod oceny ryzyka zdrowotnego najczęściej wykorzystywane były: model Almona, Poissona oraz nieliniowy model rozproszonych opóźnień DLNM. Początkowo używano modelu Poissona, prawie 60% cytowanych prac dotyczy właśnie tej metody. W ostatnich latach wzrosło zainteresowanie implementacją modelowania nieliniowego (34% cytowanych prac). Najczęściej analizy statystyczne prowadzono z wykorzystaniem oprogramowania R lub SAS. W szacowaniu ryzyka względnego zazwyczaj uwzględniano narażenie krótkoterminowe (do tygodnia), około 75% prac dotyczyło pomiaru ryzyka względnego w odpowiedzi na wzrost stężenia zanieczyszczenia o 10 μg/m3. Pozostałe uwzględniały wzrost narażenia o rozstęp międzykwartylowy (IQR).
WNIOSKI. Nieliniowy model rozproszonych opóźnień DLNM z uwagi elastyczność w definiowaniu, łatwość interpretacji oraz coraz częstsze stosowanie w pracach z zakresu epidemiologii środowiskowej staje się rekomendowanym przez badaczy narzędziem statystycznym.

Słowa kluczowe: *opóźniony efekt zdrowotny, model Almona, model Poissona, model DLNM*

INTRODUCTION

Severe smog episodes and their impact on the quality of life and health status of inhabitants have become an important issue of public health and results of realized studies confirm the necessity of quality of air improvement (1). Significant importance had results of APHEA project (*Air Pollution and Health: A European Approach*) realized in years 1991-94 in European countries which documented short-term health effects related to air pollution in adopted standardized methods of estimation (2). Statistical modeling guidelines based on time series analysis and Poisson regression were the basis for the research, however, due to the poor accuracy of statistical estimation approach over the years has changed. Further attempts were made with Almon's model borrowed from econometric sciences (3), and next was tested a quasi-Poisson distribution with elements of non-linear modeling with the use of spline functions (4,5). Nevertheless, the analyses include the general necessity of estimation the risk ratio (relative risk, RR) of delayed health effects related to air pollution exposure. A wide range of possible statistical analyzes justify the need for recognization of their capabilities in proper conclusion and is also an opportunity to assess compliance of results obtained in different statistical methods.

AIM

The aim of the presented paper is to review and assess the possibility of selected, available statistical

WSTĘP

Ciężkie epizody smogowe oraz ich wpływ na zdrowie i życie mieszkańców stały się ważnym tematem zdrowia publicznego, a wyniki prowadzonych badań stanowią podstawę działań zmierzających do redukcji stężeń emitowanych zanieczyszczeń (1). Istotne znaczenie miał w tej kwestii projekt APHEA (*Air Pollution and Health: A European Approach*) prowadzony w latach 1991-94 w krajach europejskich. Głównym założeniem była ocena krótkoterminowych efektów zdrowotnych w odpowiedzi na zanieczyszczenie powietrza atmosferycznego przy przyjętej ustandaryzowanej metodologii (2). Podstawę prowadzonych badań stanowiły wytyczne dotyczące modelowania statystycznego w oparciu o analizę szeregów czasowych oraz regresję Poissona. Z uwagi na słabą precyzyjność oszacowań statystycznych podejście to na przestrzeni lat uległo zmianie. Podejmowano próby zapożyczenia metodologii statystycznej z nauk ekonometrycznych poprzez wprowadzenie modelu Almona (3). Testowano również rozkład quasi-Poissona oraz stosowano elementy modelowania nieliniowego w wykorzystaniem funkcji sklejanych (4,5). W analizach tych zachowano jednak podstawową potrzebę, jaką jest konieczność oszacowania ryzyka względnego (RR, ang. *risk ratio*) występowania zdarzeń zdrowotnych w opóźnionej reakcji na zanieczyszczenia powietrza atmosferycznego. Szeroka oferta możliwych do przeprowadzenia analiz statystycznych skłania do podjęcia próby poznania ich możliwości w zakresie prawidłowego wnioskowania, a także jest okazją do poznania zgodności obserwacji odwołujących się do odmiennych metod statystycznych.

methods to estimate delayed health risk in the respiratory system related to air pollution exposure.

MATERIAL AND METHODS

Statistical methodology overview

The paper refers to available descriptive and analytical statistical methods used in the assessment of delayed health risks related to air pollution exposure. The general characteristics and structure of used models were presented as well as computational tools in available statistical packages such as R v.3.6.2 (2019, The R Foundation for Statistical Computing, GNU GPL), SAS v.9.4 (SAS Institute Inc., Cary, North Carolina, USA) and Statistica (data analysis software system) v.13 (2018, StatSoft Poland) are discussed. Presented paper includes only the most frequently used statistical methods relative risk estimation of delayed health effect such us Almon, Poisson, and Distributed Lag Non-linear Models (DLNM).

Literature review and selection

The literature review was performed using the PubMed platform, keywords included delayed health effect and type of dose-relationship assessment model (Poisson regression, Almon regression, Generalized Additive Model GAM). Due to the wide range of topics, the search was limited to modelling the delayed occurrence of acute respiratory events, such as asthma, pneumonia, bronchitis, laryngitis. The following health events were considered: outpatient visits, visits to emergency departments, hospital admissions, and deaths. Moreover, assessment was related to the following environmental pollutants: PM_{10} and $PM_{2.5}$ dust, ozone O_3, sulphur dioxide SO_2, nitrogen oxides NO_2 and NO_X, as well as carbon monoxide CO. The search of the bibliography was made using just English language publications, and was limited to the period of 1990-2020. Initially, 2,498 articles were selected in which there were only 1,264 full-text publications. Then, 441 papers with titles corresponding to the presented topics were selected. After removing duplicates, the content of the papers was analyzed in terms of the methodology and subject. Finally, 82 articles were included in the analysis. Detailed data about the article selection is presented in table I. The complete list of the collected bibliography is available on demand from the corresponding author.

CEL PRACY

Zasadniczym celem pracy jest przegląd i ocena możliwości wybranych, dostępnych metod statystycznych do szacowania ryzyka zdrowotnego w obrębie układu oddechowego pozostającego w związku z narażeniem na zanieczyszczenie powietrza i uwzględniających opóźnienie efektu zdrowotnego.

MATERIAŁ I METODY

Przegląd metodologii statystycznej

W pracy dokonano przeglądu opisowych i analitycznych metod statystycznych wykorzystywanych w szacowaniu środowiskowego ryzyka zdrowotnego w związku z narażeniem ludzi na zanieczyszczenia powietrza, przy uwzględnieniu opóźnienia efektu. Przedstawiono ogólną charakterystykę i budowę stosowanych modeli oraz narzędzia modelowania w dostępnych pakietach statystycznych takich jak R v.3.6.2 (2019, The R Foundation for Statistical Computing, GNU GPL), SAS v.9.4 (SAS Institute Inc., Cary, North Carolina, USA) oraz Statistica (data analysis software system) v.13 (2018, StatSoft Polska). W prezentowanej pracy uwzględniono jedynie te metody statystyczne, które najczęściej stosowano do modelowania ryzyka względnego z opóźnionym efektem. Skoncentrowano się na następujących modelach: Almona, Poissona oraz nieliniowy model rozproszonych opóźnień (ang. *Distributed Lag Non-linear Models* – DLNM).

Przegląd i selekcja piśmiennictwa

Przeglądu piśmiennictwa dokonano za pomocą platformy PubMed, w której zgromadzono m.in. prace odwołujące się do modelowania ryzyka względnego z opóźnionym efektem zdrowotnym. Z uwagi na szeroki zakres tematyczny, poszukiwania zawężono do modelowania opóźnionego występowania ostrych incydentów ze strony układu oddechowego, takich jak: astma, zapalenie płuc, zapalenie oskrzeli, zapalenie krtani. Wśród zdarzeń zdrowotnych brano pod uwagę: wizyty ambulatoryjne, wizyty na oddziałach ratunkowych, przyjęcia szpitalne oraz zgony. Przy przeszukiwaniu uwzględniono następujące czynniki środowiskowe: pyły PM_{10} i $PM_{2,5}$, ozon O_3, dwutlenek siarki SO_2, tlenki azotu NO_2 i NO_X, a także tlenek węgla CO. Wykorzystano także następujące słowa kluczowe: regresja Poissona, regresja Almona, uogólniony model addytywny (GAM, ang. *Generalized Additive Model*). Przeszukiwanie bazy publikacji wykonano z użyciem terminologii podanej w języku angielskim, zawężając wyniki do okresu publikacji w latach od 1990 do 2020. Wyłoniono 2 498 artykułów, spośród których wzięto pod uwagę jedynie 1 264 publikacje w postaci pełnotekstowej. Następnie wybrano prace o tytułach odpowiadających

Tabela I. Strategia przeglądu piśmiennictwa
Table I. The strategy of literature review

Etapy przeglądu/ review stages	Choroby układu oddechowego/ respiratory diseases N (100%)	Astma/ asthma	Zapalenie oskrzeli/ bronchitis	Zapalenie płuc/ pneumonia	Ostre incydenty oddechowe/ acute respiratory effects
Identyfikacja/ identification	2 498	825	110	339	1 224
Identyfikacja artykułów pełnotekstowych/ full-text articles identification	1 264	381	49	166	668
Przegląd tytułów/ screening by the titles	441	200	28	50	163
Kwalifikacja na podstawie abstraktów i metod/ eligibility by abstracts and methods	107	58	3	16	30
Kwalifikacja na podstawie wybranej metodologii/ eligibility by chosen methodology	**82**	**44**	**2**	**14**	**22**

RESULTS

Statistical methodology review

Evaluation of crude data presenting the level of exposure (pollutant concentrations and meteorological conditions favoring smog developing) is starting point in the analysis of delayed health effects. During the occurrence of smog episodes environmental factors are compared to dependent values y_t, (in this case, y_t represents the number of health events) to assess the potential impact and determine the maximum possible delay of health effect L. The results are presented in the form of scatter plots. It is also important to consider seasonality and the associated likely impact of different environmental factors on human health. Therefore, it is justified to analyze within the entire observation period, but also in intervals reduced to particular seasons.

Almon's model with delayed effects [1] in its general form is represented by the number of events y_t at moment/ t day, which are explained by a linear combination of function of the predictor x_t: $p_l(x)_t$ and express the lag concerning the moment t by period l, where $l=0,...,L$, L – maximum delay, with error ε_t:

$$y_t = \varepsilon_t + \sum_{l=0}^{L} \beta_l p_l(x_t) \qquad [1]$$

Calculation of β_l coefficients, determining the impact of changes in values x_t on the dependent value y_t, using a function [2], makes it possible to reduce the

przedstawionej tematyce w liczbie 441. Po usunięciu zdublowanych artykułów przeanalizowano zawartość prac pod kątem stosowanej metodologii i tematyki. Ostatecznie wyłoniono 82 artykuły. Szczegółowy proces selekcji artykułów przedstawiono w tabeli I. Pełny wykaz zebranego piśmiennictwa jest dostępny u autora korespondencyjnego.

WYNIKI

Przegląd metodologii statystycznej

Punktem wyjścia w analizie opóźnionych efektów zdrowotnych jest ocena surowych danych prezentujących poziom narażenia (stężenia zanieczyszczeń i warunki meteorologiczne sprzyjające formowaniu się smogu). W wyróżnionych okresach czynniki środowiskowe zestawiane są z wartościami objaśnianymi y_t, (w tym przypadku jest to liczba zdarzeń zdrowotnych) celem zaakcentowania potencjalnego wpływu oraz ustalenia maksymalnie możliwego opóźnienia efektu zdrowotnego L. Wyniki prezentowane są w postaci wykresów rozrzutu. Istotne jest również uwzględnienie sezonowości, oraz związanego z nią prawdopodobnego wpływu odmiennych czynników środowiskowych na zdrowie ludzi. Dlatego uzasadnione jest prowadzenie analiz w całym okresie obserwacji, ale także w okrojonych do sezonu ramach czasowych.

Model Almona z opóźnionymi efektami [1] w ogólnej postaci jest reprezentowany przez wartości liczby zdarzeń y_t w momencie/dniu t, które są objaśniane kombinacją liniową funkcji predyktora x_t: $p_l(x)_t$ i wyrażają

impact of the collinearity of variables in the analyzed moment.

$$\beta_l = \alpha_0 + \alpha_1 l + \alpha_2 l^2 + \cdots + \alpha_p l^p \qquad [2]$$
$$l = 1,2,\ldots,L;\ \ p = 1,2,\ldots,n;\ n < l$$

It is assumed in the model, that the influence of explanatory variable x_t on dependent variable y_t lasts at most for L periods. However, the main problem is the difficulty of choosing the degree of polynomial p and the inability of interpreting the delayed multivariable impact. To determine the relative risk (RR), the logarithmic transformation of the number of events y_t in moment/ t day is applied (3). It is worth noticing that, software packages R, SAS, and Statistica have statistical libraries for Almon's linear regression analysis. Table II lists the availability of functions in each of the discussed packages (6,7). Akaike's Information Criterion (AIC), and therefore the degree of the polynomial and the maximum delay is usually used as the selection criterion of the statistical model.

opóźnienie względem momentu t o okres l, gdzie $l=0, \ldots,L$, L – maksymalne opóźnienie, z błędem ε_t:

$$y_t = \varepsilon_t + \sum_{l=0}^{L} \beta_l p_l(x_t) \qquad [1]$$

Wyznaczenie współczynników β_l, określających wpływ zmian wartości x_t na oczekiwaną wartość y_t, z wykorzystaniem funkcji [2], umożliwia redukcję wpływu współliniowości zmiennych w analizowanym momencie.

$$\beta_l = \alpha_0 + \alpha_1 l + \alpha_2 l^2 + \cdots + \alpha_p l^p \qquad [2]$$
$$l = 1,2,\ldots,L;\ \ p = 1,2,\ldots,n;\ n < l$$

W tak zadanym modelu zakłada się, że wpływ zmiennej objaśniającej x_t na zmienną objaśnianą y_t trwa przez co najwyżej L okresów. Zasadniczym problemem jest jednak trudność w wyborze stopnia wielomianu p oraz brak możliwości interpretacji opóźnionego wpływu wielu zmiennych. W celu wyznaczenia ryzyka względnego (RR) stosowana jest transformacja logarytmiczna wartości liczby zdarzeń y_t w momencie/dniu t (3). Warto zauważyć, iż oprogramowanie R, SAS oraz Statistica zawierają pakiety statystyczne umożliwiające przeprowadzenie analizy liniowej regresji Almona. Dla ułatwienia samodzielnej pracy w tym zakresie, w tabeli II zestawiono dostępność poszczególnych funkcji w każdym z omawianych pakietów (6,7). Jako kryterium wyboru modelu statystycznego, a zatem stopnia wielomianu oraz maksymalnego opóźnienia, zazwyczaj stosowane jest Kryterium Informacyjne Akaikego (AIC).

Tabela II. Funkcje i pakiety umożliwiające analizę regresji Almona dostępne w oprogramowaniu R, SAS i Statistica
Table II. Functions and packages enabling Almon regression analysis available in R, SAS and Statistica software.

Model Almon'a/ Almon model	R	package: dLagM output ← polyDlm(x, y, l, p, show.beta=TRUE) output ← koyckDlm(y,x,show.summary=TRUE)
	SAS	PROC PDLREG; model y = x(N lags, Polynomial degree) z ; z - covariates without lags distribution
	Statistica	Statystyka; Zaawansowane modele liniowe i nieliniowe; Szeregi czasowe i prognozowanie; Analiza z uwzględnieniem opóźnień/ Statistics; Advanced Models; Time Series/Forecasting; Lag analysis

Due to the distribution of the dependent variable, which is similar to the Poisson distribution for the number of health events a model applied its properties is used (4,5). The Poisson model with delayed effects in general [3] is represented by the logarithm of the expected values of the number of events y_t in moment/ t day, which is explained by a linear combination of i-th functions of predictors x_t: $p_{i,l}(x)_t$ expressed by the

Z uwagi na rozkład zmiennej objaśnianej, który w przypadku liczby zachorowań jest zbliżony do rozkładu Poissona, stosowany jest model wykorzystujący jego własności (4,5). W ogólnej postaci Model Poissona z opóźnionymi efektami [3] jest reprezentowany przez logarytm oczekiwanych wartości liczby zdarzeń y_t w momencie/dniu t, która jest z kolei objaśniana kombinacją liniową i-tych funkcji predyktorów x_t: $p_{i,l}(x)_t$ wyrażonych przez opóźnienie względem momentu t o okres

delay related to moment *t* by period *l*, where *l=0,...,L*. *L* maximum delay, with error ε_t:

$$\log\left(E(y_t)\right) = \varepsilon_t + \sum_{i=1}^{m} \beta_{i,l} p_{i,l}(x_t) \qquad [3]$$

Statistical software R, SAS, and Statistica offer possibilities of statistical analysis. The availability of individual functions in each of the enlisted packages are presented in table III (8-11). Due to the overdispersion of the model, when the residual deviation is larger than the number of degrees of freedom, the so-called *quasi-Poisson* regression model is used. The counterpart of Akaike's Information Criterion (AIC) – quasi-AIC, is used as the selection criterion of a statistical model – the goodness of fit. The possibility to estimate the relative risk (RR) of the health effect related to an increase of exposure by selected unit $\Delta x_i = const$ is an advantage of the Poisson regression model, while the risk is calculated by the formula [4].

l, gdzie *l=0,...,L*. *L* oznacza maksymalne opóźnienie, z błędem ε_t:

$$\log\left(E(y_t)\right) = \varepsilon_t + \sum_{i=1}^{m} \beta_{i,l} p_{i,l}(x_t) \qquad [3]$$

Jak poprzednio, kolejne pakiety statystyczne (R, SAS oraz Statistica) oferują możliwości prowadzenia analizy. W tabeli III ujawniono dostępność poszczególnych funkcji w każdym z wymienionych pakietów oprogramowania (8-11). Z uwagi na nadmierną dyspersję modelu, gdy odchylenie resztkowe jest większe od liczby stopni swobody, stosowany jest model regresji quasi-Poissona. Jako kryterium wyboru modelu statystycznego, a zarazem dobroci dopasowania modelu stosowany jest odpowiednik Kryterium Informacyjnego Akaikego (AIC) – quasi-AIC. Zaletą modelu regresji Poissona jest możliwość oszacowania ryzyka względnego (RR) efektu zdrowotnego przy wzroście narażenia o wybraną jednostkę $\Delta x_i = const$, przy czym ryzyko to oblicza się ze wzoru [4].

Tabela III. Funkcje i pakiety umożliwiające samodzielne wykonanie analizy regresji liniowej Poissona dostępne w oprogramowaniu R, SAS i Statistica

Table III. Functions and packages enabling Poisson linear regression analysis available in R, SAS and Statistica software.

Model Poissona/ Poisson model	R	package: stats output ← glm ($y \sim \sum_{i=1}^{m} p_{i,l}$, data = data, family = poisson(link = „log"))
	SAS	PROC GENMOD; MODEL $y \sim \sum_{i=1}^{m} p_{i,l}$, / dist=poisson dscale ;
	Statistica	Statystyka; Zaawansowane modele liniowe i nieliniowe; Uogólnione modele liniowe i nieliniowe; Model log Poissona/ Statistics; Advanced Models; Generalized Linear/Nonlinear; Log Poisson Model
Model quasi-Poissona/ quasi-Poisson model	R	output ← glm ($y \sim \sum_{i=1}^{m} p_{i,l}$, data = data, family = quasipoisson(link = „log"))
	SAS	PROC GLIMMIX; MODEL $y \sim \sum_{i=1}^{m} p_{i,l}$, / dist=log solution; _variance_ = _mu_; random _residual_;
	Statistica	Model log Poissona; Szacuj rozrzut/ Log Poisson Model; Estimate scatter
kryterium quasi-AIC quasi AIC criterion	R	package: stats package: genomaths/usefr LogLike ← sum(dpois(y_t, lambda=exp(predict(output)), log=TRUE)) QAIC = 2 * (length(coef(output)) – LogLike)
	SAS	AIC (Akaike Information Criterion), AICc (The corrected Akaike's Information Criterion), BIC (Bayesian Information Criterion)
	Statistica	Model log Poissona; Dobroć dopasowania/ Log Poisson Model; Goodness of fit

$$RR_{i,l} = e^{const \cdot \beta_{i,l}} \qquad [4]$$

Due to the abovementioned limitations (limitation to one variable, overdispersion), the method using

$$RR_{i,l} = e^{const \cdot \beta_{i,l}} \qquad [4]$$

Z uwagi na wcześniej wymienione ograniczenia (ograniczenie do jednej zmiennej, nadmierna dyspersja)

spline functions (splines) in the analysis of time series was introduced (5). The Distributed Lag Non-linear Model (DLNM) in its general form [5] is represented by the logarithm of expected values of the number of events y_t in moment/ t day, which is explained by a linear combination of i-th functions of predictors x_t: $p_i(x)_t$ in moment t, and splines expressing the delay with the moment t by period l, where $l=0,...,L$. L is a maximum delay, with error ε_t:

$$\log\big(E(y_t)\big) = \varepsilon_t + \sum_{i=1}^{m} \beta_i p_i(x_t) + \sum_{i=m+1}^{p} s_i(x_{t,l}) \qquad [5]$$

The method is based on the Poisson model, however, linearity is partially replaced by nonlinear and lagging effects, defined by spline functions. The method was implemented by *A. Gasparrini* and *B. Armstrong* in the R *dlnm* package (12). This is the only broadly and freely available (GNU GPL license) statistical package that enables performing nonlinear Poisson regression with delayed effects in the analysis of time-series data. The applied functions related to the quasi-Poisson model as well as functions for implementation of splines and graphical representations of results are presented in table IV (12). Due to the overdispersion of the model, when the residual deviation is larger than the number of degrees of freedom, the quasi-Poisson regression model is used. The counterpart of Akaike's Information Criterion (AIC) – quasi-AIC, is used as the selection criterion of statistical model, hence – the goodness of fit (tab. III). The possibility to estimate the relative risk (RR) of the health effect related to an increase in exposure by selected unit Δx_i concerning setpoint x_0 is an advantage of the Poisson regression model while the risk is calculated by the following formula [6].

wprowadzono metodę wykorzystującą funkcje sklejane (splajny) w analizie szeregów czasowych (5). W ogólnej postaci model nieliniowy rozproszonych opóźnień (ang. *Distributed Lag Non-linear Models* – DLNM) [5] jest reprezentowany przez logarytm oczekiwanych wartości liczby zdarzeń y_t w momencie/dniu t, która jest objaśniana kombinacją liniową predyktorów x_t: $p_i(x)_t$ w momencie t, oraz splajnów wyrażających opóźnienie względem momentu t o okres l, gdzie $l=0,...,L$. Podobnie L oznacza maksymalne opóźnienie, z błędem ε_t:

$$\log\big(E(y_t)\big) = \varepsilon_t + \sum_{i=1}^{m} \beta_i p_i(x_t) + \sum_{i=m+1}^{p} s_i(x_{t,l}) \qquad [5]$$

Zasadniczo metoda bazuje na wykorzystaniu modelu Poissona, jednak liniowość jest tu częściowo zastąpiona efektami nieliniowymi i opóźnionymi, zdefiniowanymi poprzez funkcje sklejane. Metoda została zaimplementowana przez *A. Gasparrini i B. Armstrong* w pakiecie R *dlnm* (12). Jest to jedyny powszechnie dostępny (licencja GNU GPL) pakiet statystyczny umożliwiający wykonanie nieliniowej regresji Poissona z opóźnionymi efektami w analizie danych szeregów czasowych. Stosowane funkcje związane z wykorzystaniem modelu quasi-Poissona oraz funkcje pozwalające na implementację splajnów i graficzną prezentację wyników przedstawiono w tabeli IV (12). Z uwagi na nadmierną dyspersję modelu, gdy odchylenie resztkowe jest większe od liczby stopni swobody, stosowany jest model regresji quasi-Poissona. Jako kryterium wyboru modelu statystycznego, a zarazem dobroci dopasowania modelu stosowany jest odpowiednik Kryterium Informacyjnego Akaikego (AIC) – quasi-AIC (tab. III). Zaletą modelu DLNM jest możliwość oszacowania ryzyka względnego (RR) konkretnego efektu zdrowotnego przy wzroście narażenia o jednostkę Δx_i w odniesieniu do zadanej wartości x_0, przy czym ryzyko to oblicza się ze wzoru [6].

Tabela IV. Funkcje pakietu *dlnm* dostępne w oprogramowaniu R
Table IV. Functions of the *dlnm* package available in R software

Deklaracja węzłów interpolacyjnych i funkcji dla predyktora/ declaration of interpolation nodes and predictor functions	argvar ← list(fun="bs", knots=quantile(x, c(x_1,x_2,x_3)/100, na.rm=T), Bound=range(x, na.rm=T))
Deklaracja węzłów interpolacyjnych i funkcji dla opóźnień/ declaration of interpolation nodes and lag functions	maxlag=L arglag ← list(fun="bs", knots=logknots(maxlag, nk=n_k))
Typy funkcji/ function types	ns – naturalny splajn kubiczny/ natural spline bs – B-splajn/ simple B-splines strata – stratyfikacja dla określonych wartości/ strata through dummy variables poly – funkcja wielomianowa/ polynomial function lin – funkcja liniowa/ linear function

Funkcja macierzy dla predyktora i opóźnień/ matrix function for the predictor and lags	$s_i \leftarrow$ crossbasis(x, lag, argvar=list(), arglag=list(), group=NULL)
Regresja quasi-Poissona/ quasi-Poisson regression	output ← glm ($y\sim \sum_{i=1}^{m} p_i + y\sim \sum_{i=1}^{m} s_i$, data = data, family = quasipoisson(link = „log"))
Funkcja generująca prognozy/ forecasts function	prediction ← crosspred(s_i, output, coef=NULL, vcov=NULL, model.link=NULL, at=NULL, from=NULL, to=NULL, by=NULL, lag, bylag=1, cen=NULL, ci.level=0.95, cumul=FALSE)
Wykresy/ plots	plot(prediction, „3d", xlab="x", ylab="y", zlab="z", col=gray(), main="3D")
	plot(prediction, "slices", type="p", cex=1, var=x_0, ci="bars", ylab="RR", main="Lag-specyfic effects")
	plot(prediction, "overall", xlab="x", xlim=c(x_{min},x_{max}), ylim=c(y_{min},y_{max}), main="Overall effect of temperature", cumul=FALSE)

$$RR_{i,l} = e^{\beta_{i,l}} \qquad [6]$$

$$RR_{i,l} = e^{\beta_{i,l}} \qquad [6]$$

Review of the literature using models with delayed health effect

The review of the literature allowed us to select 82 articles, in which modeling with a delayed effect was used. The Almon model was used in 4 works, in over half of the studies (n=48; 58.5%) the Poisson model, and in every third (n=29; 35.4%) the non-linear modeling (DLNM) were used. Only one paper referred to all the modeling methods. The works published in the years 1994-2020 were analyzed, while the authors used data collected in individual research periods between 1985 and 2016. Moreover, obtained results represent various regions of the world, except for African countries; the detailed data are presented in table V.

Przegląd piśmiennictwa wykorzystującego modele z opóźnionym efektem zdrowotnym

Dokonany przegląd piśmiennictwa wyłonił 82 artykuły, w których zastosowano modelowanie z opóźnionym efektem, przy czym w 4 pracach wykorzystano model Almona. W ponad połowie prac (n=48; 58,5%) stosowano model Poissona, w co trzeciej (n=29; 35,4%) wykorzystano modelowanie nieliniowe (DLNM). Tylko jedna praca odnosiła się do wszystkich wymienionych sposobów modelowania. Wzięto pod uwagę prace opublikowane w latach 1994-2020, przy czym autorzy analizowali dane zebrane w poszczególnych okresach badawczych pomiędzy 1985 i 2016 rokiem. Warto dodać, że wyniki badań reprezentują różne regiony świata z wyjątkiem krajów afrykańskich a szczegółowe dane przedstawiono w tabeli V.

Tabela V. Statystyka opisowa dla zebranej bibliografii (N=82) ujawniająca rok i okres badawczy, region i metodologię
Table V. Basic statistic for used bibliography (N=82) by year, period of the study, region and methodology

Charakterystyka/ characteristic		Ogółem/ total N=82 (100%)	Regresja Almon'a/ Almon regression N=4 (100%)	Regresja Poissona/ Poisson regression N=48 (100%)	DLNM N=29 (100%)
Rok publikacji/ year of publication	1994-2000	14 (17,1%)	1	13	0
	2001-2010	14 (17,1%)	1	11	2
	2011-2020	54 (65,8%)	3*	25*	28*
Okres badawczy (lata)/ study period	<=1	21 (25,6%)	2*	17*	4*
	2-4	28 (34,1%)	2	18	8
	5-7	20 (24,4%)	1	10	9
	8-10	9 (11%)	0	2	7
	>10	4 (4,9%)	0	2	2

Lokalizacja badania/ place of study	Ameryka Południowa/ South America	17 (20,7%)	1	14	2
	Ameryka Północna/ North America	18 (21,9%)	1	8	9
	Australia/ Australia	3 (3,7%)	0	1	2
	Azja/ Asia	25 (30,5%)	1	13	11
	Europa/ Europe	19 (23,2%)	2*	13*	6*
Oprogramowanie/ software	R	24 (29,3%)	2*	7*	17*
	SAS	14 (17,1%)	0	11	3
	S-Plus	6 (7,3%)	0	4	2
	SPSS	4 (4,9%)	0	4	0
	Stata	5 (6,1%)	1	3	1
	Statistica	5 (6,1%)	1	4	0
	Brak/ lack	24 (29,3%)	1	16	7
Zakres badanego opóźnienia efektu zdrowotnego [dni]/ lag of health effect [days]	0	4 (4,9%)		4	
	0-2	8 (9,8%)	0	5	3
	0-3	8 (9,8%)	1	4	3
	0-4	10 (12,2%)	0	7	3
	0-5	15 (18,3%)	0	12	3
	0-6	11 (13,4%)	1	2	8
	0-7	10 (12,2%)	0	6	4
	>7	10 (12,2%)	2*	6*	4*
	Brak/ lack	6 (7,3%)	1	3	2
Jednostka przyrostu narażenia Δx_i w szacowaniu ryzyka względnego RR / increase by unit of exposure in the RR assessment	1	3 (3,7%)	0	3	0
	10	29 (35,4%)	1	13	15
	100	7 (8,5%)	0	7	0
	IQR	17 (20,7%)	2*	7*	10*
	Inne/ other	10 (12,2%)	0	9	1
	Brak/ lack	16 (19,5%)	2	10	4

*- łącznie z publikacją powstałą w oparciu o każdy rodzaj modelowania/ together with the publication based on each type of modeling
IQR – rozstęp międzykwartylowy/ interqurtile range

Mostly the cited publications was concentrated on the delayed effects related to short-term exposure to the following air pollutants: PM_{10} (52 papers; 62,7%), $PM_{2,5}$ (43 papers; 51,8%), O_3 (49 papers; 59%), SO_2 (38 papers; 45,8%), CO (20 papers; 24,1%), NO_2 (44 papers; 53%), NO_X (4 papers; 4,8%). The review was based on the occurrence of health events from the respiratory system: asthma (44 papers; 53.7%), bronchitis (2 papers; 2.4%), pneumonia (14 papers; 17.1%), acute respiratory incidents (22 papers; 26.8%). The analysis used registry data on hospital admissions (36 papers; 43.9%), emergency visits (19 papers; 23.2%), deaths (11 papers; 13.4%), outpatient visits (5 papers; 6.1%) and combined health benefits (11 papers; 13.4%). Usually publications referred to the general population (51 papers; 62.2%), but also children (25 papers; 30.5%) and some of them only adults (6 papers; 7.3%).

Głównym celem prezentowanych w bibliografii badań było sprawdzenie opóźnionego wpływu na zdrowie krótkoterminowego narażenia na następujące zanieczyszczenia powietrza: PM_{10} (52 prace; 62,7%), $PM_{2,5}$ (43 prace; 51,8%), O_3 (49 prac; 59%), SO_2 (38 prac; 45,8%), CO (20 prac; 24,1%), NO_2 (44 prace; 53%), NO_X (4 prace; 4,8%). W przeglądzie z założenia brano pod uwagę występowanie zdarzeń zdrowotnych ze strony układu oddechowego: astmy (44 prace; 53,7%), zapalenia oskrzeli (2 prace; 2,4%), zapalenia płuc (14 prac; 17,1%), ostrych incydentów oddechowych (22 prace; 26,8%). W analizach wykorzystywano dane z oficjalnych rejestrów i dotyczące: przyjęć szpitalnych (36 prac; 43,9%), wizyt na oddziałach ratunkowych (19 prac; 23,2%), zgonów (11 prac; 13,4%), wizyt ambulatoryjnych (5 prac; 6,1%) oraz łączonych świadczeń zdrowotnych (11 prac; 13,4%). Badania dotyczyły zazwyczaj populacji ogólnej (51 prac; 62,2%), ale także

Statistical analyses were frequently carried out using R or SAS software. In the case of 43 (52.4%) papers, the quasi-Poisson model was used due to the observed overdispersion. The use of a nonlinear model was associated with the use of natural splines (29 papers) rather than B-splines (only one paper). The relative risk of health effect was estimated for delays in the next days after environmental exposure, while the short-term effect, i.e. up to one week, was the most often used, n=62 (75.6%). The maximum effect was related to month delay and dose-response was calculated for an increase of exposure by unit (10 mg/m^3) or the value IQR (interquartile range).

DISCUSSION

The paper presents particularly available and accepted statistical models usually used to estimate the relative risk of delayed respiratory health effect in response to the increase of air pollution concentration. Additionally, were presented practical applications of the indicated methods in selected statistical packages.

Review of bibliography shows that the most frequently used models are: Almon's model, Poisson's model, and the DLNM model. Each of the above listed methods has some certain limitations (13). In Almon's model, the analysis must be preceded by establishing initial conditions, including the maximum delay *L* (influencing the number of variables in the model). It is also necessary to define the degree of the polynomial *p* for the approximation. The obtained system of *L* equations of the *p* degree requires an appropriately large data set. A significant limitation of the method is the possibility of assessing the impact of just one environmental factor. An alternative solution is the Poisson model based on the classical formula of the GLM model (Generalized Linear Model). This method enables multivariable analysis, however, the model tends to be overdispersed. Relatively the newest one, the DLNM method, use spline functions to define predictors, which makes it easier to precisely define the initial conditions in the relative risk analysis. The model enables multivariable analysis and each of the predictors can be implemented by selecting an appropriate function sensitive to individual initial conditions via the criterion quasi-AIC. Flexibility in parametrizing the model and ease of its interpretation points at applying this method to environmental epidemiology studies aimed at estimating health risks in response to increased exposure.

The analysis of available reviewed literature confirms that the non-linear DLNM method is increasingly used, although the classic Poisson model remains also a popular method. This is probably related to the guidelines introduced by the APHEA

dzieci (25 prac; 30,5%) oraz wyłącznie osób dorosłych (6 prac; 7,3%).

Analizy statystyczne najczęściej prowadzone były z wykorzystaniem oprogramowania R lub SAS. W przypadku 43 (52,4%) prac zastosowano model quasi-Poissona z uwagi na zaobserwowaną nadmierną dyspersję modelu. Z kolei wykorzystanie modelu nieliniowego związane było raczej ze stosowaniem splajnów naturalnych (29 prac) niż B-splajnów (tylko jedna praca). Ryzyko względne efektu zdrowotnego szacowane było dla opóźnień w kolejnych dniach od wystąpienia narażenia środowiskowego, przy czym najczęściej brano pod uwagę krótkoterminowy efekt, tj. do jednego tygodnia, n=62 (75,6%). Maksymalnie uwzględniano miesięczne opóźnienie efektu w odpowiedzi na wzrost narażenia o 10 jednostek pomiarowych ($\mu g/m^3$) lub o wartość IQR (rozstępu międzykwartylowego).

DYSKUSJA

W pracy przedstawiono uznane modele matematyczne zazwyczaj stosowane do szacowania względnego ryzyka opóźnionego oddechowego efektu zdrowotnego w odpowiedzi na wzrost zanieczyszczenia powietrza atmosferycznego. Zaprezentowano również sposób praktycznego dostępu do wskazanych metod w wybranych pakietach statystycznych.

Dokonany przegląd piśmiennictwa wskazuje, że wśród najczęściej stosowanych dotychczas modeli znajdują się: model Almona, model Poissona oraz model DLNM. Każda z wymienionych metod posiada pewne ograniczenia (13). W przypadku modelu Almona analiza musi być poprzedzona ustaleniem warunków początkowych, w tym maksymalnego opóźnienia *L* (mającego wpływ na liczbę zmiennych w modelu), niezbędne jest również określenie stopnia wielomianu *p* dla aproksymacji. Uzyskany układ *L* równań stopnia *p* wymaga odpowiednio licznego zbioru danych. Istotnym ograniczeniem metody jest możliwość oceny wpływu tylko jednego czynnika środowiskowego. Alternatywnym rozwiązaniem jest model Poissona oparty o klasyczną formułę modelu GLM (ang. *Generalized Linear Model*). Metoda ta umożliwia wprawdzie przeprowadzenie analizy wieloczynnikowej, jednakże model ma tendencję do nadmiernego rozproszenia. Stosunkowo najnowsza, metoda DLNM wykorzystuje funkcje sklejane w definiowaniu predyktorów, co ułatwia dokładne sprecyzowanie warunków początkowych w analizie ryzyka względnego. Model umożliwia analizę wieloczynnikową a jednocześnie każdy z predyktorów może zostać zaimplementowany przy wyborze odpowiedniej funkcji wrażliwej na indywidualne warunki początkowe dzięki kryterium dopasowania modelu quasi-AIC. Duża elastyczność w definiowaniu modelu oraz łatwość interpretacji skłania do stosowania właśnie tej metody

project imposing this pattern of statistical analysis of data, and this automatically implies relatively numerous publications using this method (1,2). As we already mentioned, most researchers still use Poisson or quasi-Poisson regression in their studies, but more recent publications indicate an increasing interest in the use of spline functions in the GAM model (14). Unfortunately, it should be noticed that currently, only R software have tools dedicated to DLNM regression analysis.

SUMMARY

The Poisson model is the most frequently used method of statistical analysis in time-series studies with a delayed health effect. However, distributed Lag Non-linear Model DLNM is classified as the statistical tool recommended by researchers due to its flexibility in defining, simplicity in interpretation, and increasingly frequent applications to environmental epidemiology.

REFERENCES

1. Katsouyanni K i in. Short-term effects of air pollution on health: a European approach using epidemiological time-series data. The APHEA project: background, objectives, design. Eur Respir J 1995;8:1030–1038. DOI: 10.1183/09031936.95.08061030.
2. Katsouyanni K i in. Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. J Epidemiol Comm Health 1995;50(1):S12-S18.
3. Schwartz, Joel The Distributed Lag between Air Pollution and Daily Deaths, Epidemiology: 2000;11(3):320-326.
4. Bhaskaran K, Gasparrini A, Hajat S i in. Time series regression studies in environmental epidemiology. Int J Epidemiol. 2013;42(4):1187-95. DOI: 10.1093/ije/dyt092.
5. Gasparrini A. Modeling exposure-lag-response associations with distributed lag non-linear models. Stat Med. 2014;33(5):881-99. DOI: 10.1002/sim.5963. Epub 2013 Sep 12.
6. The Comprehensive R Archive Network. dLagM package. Available online: https://cran.r-project.org/web/packages/dLagM/dLagM.pdf (dostępny 11 lipca 2020).
7. SAS/ETS® 13.2 User's Guide. The PDLREG Procedure. Available online: https://support.sas.com/documentation/onlinedoc/ets/132/pdlreg.pdf (dostępny 11 lipca 2020).

w badaniach z zakresu epidemiologii środowiskowej zmierzających do szacowania ryzyka zdrowotnego w odpowiedzi na wzrost narażenia.

W trakcie analizy zebranego piśmiennictwa potwierdzono, że coraz częściej stosowana jest metoda nieliniowa DLNM, choć klasyczny model Poissona pozostaje nadal popularną metodą. Prawdopodobnie jest to związane z wytycznymi wprowadzonymi w ramach projektu APHEA narzucającymi właśnie ten schemat statystycznej analizy danych, co automatycznie przekłada się na stosunkowo liczne publikacje odwołujące się do tej metody (1,2). Jak już wielokrotnie wspomniano większość badaczy wykorzystywała w swoich pracach regresję Poissona lub quasi-Poissona ale nowsze publikacje wskazują na wzrost zainteresowania możliwością wykorzystania funkcji sklejanych w modelu GAM (14). Niestety, należy zauważyć, że aktualnie jedynie oprogramowanie R dysponuje narzędziami dedykowanymi analizie regresji DLNM.

WNIOSKI

Najczęściej stosowaną metodą analiz statystycznych w badaniach szeregów czasowych z opóźnionym efektem zdrowotnym jest model Poissona, przy czym w ostatnich latach wzrasta zainteresowanie możliwością wykorzystania modelowania nieliniowego DLNM.

8. The Comprehensive R Archive Network. Stats package. Available online: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (dostępny 11 lipca 2020).
9. SAS/ETS® 14.3 User's Guide. The GENMOD Procedure. Available online: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_genmod_syntax01.htm&docsetVersion=14.3&locale=en (dostępny 11 lipca 2020).
10. SAS/ETS® 14.3 User's Guide. The GLIMMIX Procedure. Available online: https://support.sas.com/rnd/app/stat/procedures/glimmix.html (dostępny 11 lipca 2020).
11. Stanisz, A. Logistic regression models. Applications in medicine, natural and social sciences. StatSoft: Kraków, Poland, 2016.
12. The Comprehensive R Archive Network. Dlnm package. Available online: https://cran.r-project.org/web/packages/dlnm/ (dostępny 11 lipca 2020).
13. Niewiadomska E, Kowalska M, Niewiadomski A i in. Assessment of Risk Hospitalization due to Acute Respiratory Incidents Related to Ozone Exposure in Silesian Voivodeship (Poland). Int J Environ Res Public Health. 2020;17(10):3591. DOI: 10.3390/ijerph17103591

14. Chisato I, Masahiro HA Systematic Review of Methodology: Time Series Regression Analysis for Environmental Factors and Infectious Diseases. Tropical Medicine and Health 2015;43(1):1–9. DOI: 10.2149/tmh.2014-21.

Received. 13.07.2020
Accepted for publication: 23.11.2020
Otrzymano: 13.07.2020 r.
Zaakceptowano do publikacji: 23.11.2020 r.

Address for correspondence:
Adres do korespondencji:
Ewa Niewiadomska
Faculty of Health Sciences in Bytom,
Department of Epidemiology and Biostatistics,
Medical University of Silesia, Poland
Piekarska 18 Str.
41-902 Bytom
e-mail: eniewiadomska@sum.edu.pl