GDAŃSKI UNIWERSYTET MEDYCZNY

WYDZIAŁ FARMACEUTYCZNY

Z ODDZIAŁEM MEDYCYNY LABORATORYJNEJ

Emilia Leila Daghir-Wojtkowiak

Zastosowanie regresji LASSO i bayesowskich modeli hierarchicznych do analizy danych „omicznych” i chromatograficznych

Zakład Biofarmacji i Farmakokinetyki Katedry Biofarmacji i Farmakodynamiki Gdańskiego Uniwersytetu Medycznego

Promotor pracy:

Prof. dr hab. n. farm. Michał Jan Markuszewski

Gdańsk 2018

Praca powstała w wyniku realizacji projektu badawczego nr 2014/13/N/NZ7/00474 finansowanego ze środków Narodowego Centrum Nauki. Praca została współfinansowana z dotacji budżetowych dla Młodych Naukowców, GUMed (MN-01-022/08/529).

"*.... nie ogrom wiedzy napełnia i nasyca duszę,*
lecz raczej wewnętrzne rozumienie rzeczy
i smakowanie w nich."

Św. Ignacy Loyola

SPIS TREŚCI

STRESZCZENIE

Wielowymiarowe dane, generowane z dużą szybkością i łatwością, cechują się zmiennością i złożonością. W takiej strukturze danych często wykrywane są pozorne zależności wynikające z błędów pomiarowych czy wielokrotnego testowania. Regularyzacja pozwala na oddzielenie sygnału od szumu informacyjnego poprzez zapewnienie regularności rozkładu estymatorów. Znajduje ona zastosowanie nie tylko dla danych wielowymiarowych lecz także w kontekście wielopoziomowej struktury danych. Hierarchiczna struktura danych jest konsekwencją schematu eksperymentalnego, w którym obserwacje oraz jednostki eksperymentalne pogrupowane są zgodnie z nadrzędnym schematem. Wielopoziomowość danych skutkuje brakiem niezależności obserwacji oraz większym prawdopodobieństwem korelacji zmiennych w obrębie grupy niż między grupami. W takim układzie, efektywna wielkość próby jest mniejsza niż całkowita liczba obserwacji we wszystkich klastrach. Stąd, zastosowanie regresji liniowej w sytuacji gdy dane mają strukturę hierarchiczną jest podejściem niepoprawnym, gdyż nie są spełnione podstawowe założenia leżące u podstaw regresji liniowej. Alternatywą są modele hierarchiczne w ujęciu bayesowskim, które stanowią użyteczną metodę analizy danych uwzględniając ich wielopoziomową strukturę oraz umożliwiają modelowanie danych zarówno na poziomie pojedynczej obserwacji jak i w obrębie grupy. Podejście bayesowskie umożliwiło zdefiniowanie modelu w kategoriach probabilistycznych uwzględniając dotychczasową wiedzę na temat eksperymentu w formie rozkładu *a priori* nałożonego na parametry modelu.

W przedłożonej pracy przedstawiono zastosowanie regularyzacji do analizy danych chromatograficznych jak również zastosowanie bayesowskich modeli hierarchicznych do analizy danych „omicznych" wygenerowanych w badaniach kliniczno-kontrolnych (celowana analiza metabolomiczna, profilowanie ekspresji mikro RNA) oraz danych zebranych w punktach czasowych (niecelowana analiza metabolomiczna w eksperymencie *in vivo*). Zastosowanie regresji LASSO oraz bayesowskich modeli hierarchicznych umożliwiło (i) selekcję zmiennych z matrycy danych wielowymiarowych jak też redukcję wariancji współczynników modelu umożliwiając jak największą jego generalizację, (ii) uwzględnienie różnych źródeł zmienności w modelu (zmienności międzyosobniczej, między okazjami oraz

analitycznej), (iii) uwzględnienie informacji *a priori*, (iv) ograniczenie występowania wyników fałszywie-pozytywnych, (v) estymację niepewności dla parametrów modelu i przewidywań oraz (vi) modelowanie stężeń powyżej limitu detekcji.

W przypadku analizy zależności struktura-retencja w układzie HILIC dla 16 nukleozydów oraz 11 związków pterynowych, metoda LASSO w połączeniu z regresją krokową umożliwiła selekcję zmiennych oraz otrzymanie modeli o dobrej zdolności predykcyjnej. W przypadku danych pochodzących z celowanej analizy metabolomicznej, obejmujących oznaczanie 13 nukleozydów w moczu pacjentów chorych na nowotwory układu moczowo-płciowego, zaproponowane podejście umożliwiło modelowanie stosunku szybkości wydalania nukleozydów do kreatyniny w funkcji dostępnych kowariant oraz oszacowanie indywidualnego prawdopodobieństwa choroby w oparciu o zmierzone stężenia. Z kolei, zastosowanie bayesowskiego modelowania hierarchicznego dla danych metabolomicznych zebranych w punktach czasowych pochodzących z eksperymentu *in vivo* pozwoliło na określenie zmienności w intensywności zmierzonych sygnałów oraz ocenę trendu zmian „w czasie" w profilu metabolomicznym. Ponadto, dla danych pochodzących z profilowania 50 mikro RNA u zdrowych i chorych na raka jajnika (praca będąca kontynuacją badań, będąca rozwinięciem trzech opublikowanych prac wchodzących w skład rozprawy doktorskiej, aczkolwiek nie stanowiąca podstawy osiągnięcia doktorskiego), zaproponowana metoda umożliwiła wnioskowanie na temat użyteczności ich oznaczania do wykrywania tego nowotworu.

Zastosowanie regresji LASSO do analizy wielowymiarowych danych chromatograficznych pozwoliło na łatwiejsze oddzielenie sygnału od szumu. Z kolei, bayesowskie modelowanie hierarchiczne w dyscyplinach „omicznych" wydaje się być szczególnie atrakcyjne, ponieważ pozwala uwzględnić wpływ fizjologii na mierzone wielkości, wnioskowanie na temat użyteczności potencjalnych wskaźników stanów patofizjologicznych, szacowanie indywidualnego prawdopodobieństwa choroby w aspekcie diagnozy i leczenia oraz modelowanie danych zebranych w punktach czasowych.

ABSTRACT

Multidimensional data, generated quickly and easily, are characterized by variability and complexity. In such a data structure, spurious correlations resulting from measurement errors or multiple testing can often be identified. The regularization-based technique allows separating signal from the information noise by ensuring evenly distributed estimates. This technique is applicable not only for multidimensional data but also in the context of multilevel data. The hierarchical structure of data is a consequence of an experimental scheme in which observations and experimental units are grouped according to the appropriate scheme. Multilevel data results in a lack of independence of observations and a greater probability of correlation between variables within the group than between groups. In addition, the effective sample size is smaller than the total number of observations in all clusters. Thus, the use of linear regression if we deal with hierarchical structure, is an incorrect approach, as the basic assumptions underlying the linear regression are not met. Bayesian hierarchical models constitute an alternative which is a useful method of data analysis considering their multilevel structure. They provide data modeling at the level of both, single observation and within the group as well. The Bayesian approach provided a probabilistic interpretation of the models considering the current knowledge on the experiment in the form of *a priori* distribution imposed on model parameters.

This work presents the use of regularization for the analysis of a chromatographic type of data as well as Bayesian hierarchical models for the analysis of "omics" data generated in case-control studies (targeted metabolomic approach, micro RNA expression profiling) and data collected at time points (untargeted metabolomics approach in an experiment *in vivo*). Application of the LASSO regularization and Bayesian hierarchical modeling provided (i) feature selection from the multidimensional data matrix as well as reduction of variance of model coefficients allowing model generalization, (ii) accounting for the available sources of variability in the model (inter-individual variability, between-subject variability, measurement error), (iii) consideration of *a priori* information, (iv) reduction of false-positives, (v) estimation of uncertainty around model parameters and predictions, and (vi) modeling concentrations below the detection limit.

In terms of the quantitative structure-retention relationship of 16 nucleosides and 11 pterin compounds in HILIC mode, the LASSO in combination with stepwise regression provided feature selection resulting in models with good predictive performance. For data obtained from targeted metabolomics, in which 13 nucleosides were determined in the urine of patients with urinary tract cancer, the proposed approach allowed modeling the rate of excretion of nucleosides to creatinine as a function of available covariates, estimating an individual disease probability based on the concentrations measured. In turn, the application of Bayesian hierarchical modeling for metabolomics data collected at the time points derived from *in vivo* experiment allowed to determine variability in the intensity of measured signals, and to assess the trend of "change over time" in the metabolomic profile. For data obtained from expression profiling of 50 micro RNA in healthy and ovarian cancer patients (paper which is a continuation of on-going research, being an extension of three papers already published and included in the doctoral dissertation, however not as its integral part), the proposed method allowed to make inference on the usefulness of their determination to detect this cancer.

The use of regularization for the analysis of multidimensional data allows separating the signal from noise. In turn, Bayesian hierarchical modeling in "omics" disciplines seems to be particularly advantageous in the light of physiological factors affecting the measured quantity, inferences on the usefulness of potential indicators of pathophysiological states, estimation of individual probability of disease in the aspect of diagnosis and treatment and modeling of data collected at specific time points.

LISTA PUBLIKACJI

Lista publikacji stanowiących rozprawę doktorską:

1. Emilia Daghir-Wojtkowiak, Paweł Wiczling, Szymon Bocian, Łukasz Kubik, Piotr Kośliński, Bogusław Buszewski Roman Kaliszan, Michał Jan Markuszewski, *Least absolute shrinkage and selection operator and dimensionality reduction techniques in quantitative structure retention relationship modeling of retention in hydrophilic interaction liquid chromatography*, Journal of Chromatography A (2015); 1403:54-62. (link)

2. Emilia Daghir-Wojtkowiak, Paweł Wiczling, Małgorzata Waszczuk-Jankowska, Roman Kaliszan, Michał Jan Markuszewski, *Multilevel pharmacokinetics-driven modeling of metabolomics data*, Metabolomics (2017); 13 (3); 31. (link)

3. Paweł Wiczling, Emilia Daghir-Wojtkowiak, Arlette Yumba Mpanga, Damian Szczesny, Roman Kaliszan, Michał Jan Markuszewski, *How to model temporal changes in nontargeted metabolomics study? A Bayesian multilevel perspective*, Journal of Separation Science (2017); 40 (24); 4667-4676. (link)

Lista innych publikacji:

1. Kośliński P, Daghir-Wojtkowiak E, Szatkowska-Wandas P, Markuszewski M, Markuszewski MJ, The metabolic profiles of pterin compounds as potential biomarkers of bladder cancer-Integration of analytical-based approach with biostatistical methodology. *J Pharm Biomed Anal*. 2016; 5;127:256-62.
2. Bujak R, Daghir-Wojtkowiak E, Kaliszan R, Markuszewski MJ, PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-targeted Metabolomics Data. *Front Mol Biosci*. 2016; 26;3:35.
3. Buszewska-Forajta M, Siluk D, Daghir-Wojtkowiak E, Sejda A, Staśkowiak D, Biernat W, Kaliszan R, Studies of the effect of grasshopper abdominal secretion on wound healing with the use of murine model. *J Ethnopharmacol*. 2015; 24;176:413-23.
4. Daghir-Wojtkowiak E, Struck-Lewicka W, Waszczuk-Jankowska M, Markuszewski M, Kaliszan R, Markuszewski MJ, Statistical-based approach in potential diagnostic application of urinary nucleosides in urogenital tract cancer. *Biomark Med.* 2015; 9 (6):577-95.
5. Szatkowska-Wandas P, Koba M, Kuchcicka A, Kurek S, Daghir-Wojtkowiak E, Bączek T, The application of connected QSRR and QSAR strategies to predict the physicochemical interaction of acridinone derivatives with DNA. *Comb Chem High Throughput Screen*. 2014;17(10):820-6.
6. Kośliński P, Bujak R, Daghir E, Markuszewski MJ, Metabolic profiling of pteridines for determination of potential biomarkers in cancer diseases. *Electrophoresis.* 2011; 32 (15):2044-54.
7. Bujak R, Daghir E, Rybka J, Koslinski P, Markuszewski MJ, Metabolomics in urogenital cancer. *Bioanalysis.* 2011;3(8):913-23.
8. Daghir E, Markuszewski MJ, Disposition of drugs of abuse and their metabolites in wastewater as a method of the estimation of drug consumption. *Curr Drug Metab.* 2010;11(8):629-38.

I. CZĘŚĆ TEORETYCZNA

1. Regularyzacja w uczeniu maszynowym

Generowanie dużej ilości danych jest zjawiskiem powszechnym zarówno w świecie nauki, jak też w szeroko pojętej branży biznesowej. Wielowymiarowe dane ($p>>n$, liczba predyktorów znacznie przewyższa liczbę obserwacji), generowane z dużą szybkością i łatwością, cechują się zmiennością (*variability*) i złożonością (*complexcity*). Jedną z popularnych metod ich analizowania jest uczenie maszynowe (*machine learning*), które jest gałęzią sztucznej inteligencji skoncentrowaną na badaniu algorytmów i systemów usprawniających działanie w miarę zdobywania doświadczenia. Uczenie maszynowe, ze względu na postać danych zastosowanych w procesie uczenia, można podzielić na uczenie z nadzorem (*supervised learning*) i bez nadzoru (*unsupervised learning*). Uczenie z nadzorem to przede wszystkim metody regresji, dla których celem jest predykcja zmiennej zależnej na podstawie wartości zmiennych niezależnych. Natomiast uczenie bez nadzoru obejmuje identyfikację wzorców pojawiających się w danych [1].

W przypadku regresji, algorytm uczący na podstawie danych poszukuje funkcji, która najlepiej opisuje dostępne dane, jak również będzie przydatna do przewidywania wartości zmiennej zależnej dla nowych wartości zmiennych niezależnych. Po wybraniu najlepszej funkcji, kolejnym krokiem jest określenie oczekiwanej straty wynikającej z użycia tej funkcji do przewidywania wartości zmiennej zależnej dla każdej zmiennej niezależnej, która minimalizuje oczekiwany błąd (*expected error*). Funkcja straty (*loss function*) powinna wykazywać zdolność do uogólnienia (generalizacji), tzn. różnica pomiędzy błędem empirycznym (z danych) a teoretycznym (z modelu) powinna dążyć do zera, gdy rozmiar zbioru uczącego dąży do nieskończoności.

W procesie minimalizacji błędu empirycznego na zbiorze uczącym bardzo często obserwuje się zjawisko przetrenowania modelu (*model overfitting*). Polega ono na tym, że funkcja straty bardzo dobrze opisuje dane ze zbioru treningowego, jednak dzieje się to kosztem uogólnienia. Jedną z metod zapobiegania przetrenowaniu modelu jest zastosowanie walidacji (np. krzyżowej; *k*-krotnej (*k-fold*), będącej narzędziem do sprawdzania efektywności modelu (*model effectiveness*), jak również jest użyteczna

do weryfikacji, które parametry modelu generują najmniejszy błąd na zbiorze testowym.

W kontekście wielowymiarowych zbiorów danych, walidacja *k*-krotna wydaje się mieć szczególne znaczenie. Ta metoda walidacji dzieli zbiór treningowy na *k* równych podzbiorów, a następnie każdy podzbiór jest testowany używając *k*-1 podzbiorów jako zbioru treningowego. Błędy wszystkich iteracji są uśrednione, tak aby otrzymać błąd globalny modelu. W ten sposób każda obserwacja znajduje się w zbiorze walidującym tylko raz. Taka strategia ogranicza obciążenie (*bias*) i wariancję (*variance*) współczynników modelu.

Jednym z nadrzędnych problemów w procesie uczenia maszynowego jest funkcja zbyt dobrze dopasowana do danych. Zmniejsza ona obciążenie współczynników kosztem wariancji, prowadząc do przeszacowania modelu i jego słabej generalizacji na niezależnym zbiorze danych [2].

Problem zbyt dobrego dopasowania można rozwiązać stosując metodę regularyzacji Tikhonova [3], która nakłada karę za zbyt skomplikowaną postać modelu w celu minimalizacji miary jakości dopasowania modelu (np. błędu średniokwadratowego). Zastosowanie regularyzacji prowadzi do redukcji wariancji kosztem obciążenia modelu umożliwiając jego jak największą generalizację. Człon kary (λ) równoważy balans pomiędzy dopasowaniem funkcji (*goodness-of-fit*) do danych, a jej złożonością (*bias-variance trade-off*) określając w ten sposób siłę regularyzacji [4].

Wśród metod regularyzacji można wyróżnić:

(i) regularyzację L1 (LASSO, *Least Absolute Shrinkage and Selection Operator*); [Hastie, Tibshirani and Wainwright (2015); Bühlmann and van de Geer (2011); Pokarowski and Mielniczuk (2015)]
(ii) regularyzację L2 (regresja grzbietowa, *ridge regression*);
(iii) mieszaninę L1 i L2 (sieci elastyczne, *elastic net*) [Zou, Hastie (2005)].

Norma L1 (1) to najczęściej stosowana metoda regularyzacji regresji liniowej, w której wprowadzenie członu kary (2) w miejsce dużych wartości bezwzględnych parametrów skutkuje uzyskaniem estymatorów o mniejszej wariancji, kosztem ich obciążenia.

$$LASSO = \left(\sum_{i=1}^{n} y_i^{'} - \beta_0^{'} - \sum_{j=1}^{p} \beta_j^{'} x_{ij}^{'} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^{'} \right| \qquad (1)$$

$$\sum_{j=1}^{p} \left| \beta_j^{'} \right| \leq \lambda \qquad (2)$$

gdzie y'_i oznacza standaryzowaną *i*-tą zmienną objaśnianą, β'_0 – standaryzowany wyraz wolny, β'_j – standaryzowane współczynniki regresji dla standaryzowanej kowarianty *j*, x'_{ij} oznacza *i*-tą standaryzowaną zmienną dla *j*-tej standaryzowanej kowarianty, *n* – liczba obserwacji, *p* – ilość kowariant, λ – parametr penalizujący.

Wartość λ dostrajana jest z poziomu danych w ten sposób, że im większa wariancja w danych, tym większa kara (standardowe odchylenie szacowane jest na podstawie walidacji krzyżowej). Główną trudnością w metodzie regularyzacji jest ustalenie wartości parametru kary. W celu optymalnego wyznaczenia λ [5], dla różnych wartości λ stosuje się ocenę błędu predykcji poprzez walidację krzyżową (*cross-validation*) lub kryteria informacyjne tj. AIC (*Akaike Information Criterion*) oraz BIC (*Bayesian Information Criterion*). Duże wartości λ bardziej zmniejszają bezwzględne wartości współczynników i więcej z nich zmierza w kierunku zera, zmniejszając tym samym wariancję kosztem obciążenia. Podsumowując, LASSO promuje rzadkie rozwiązania (*sparse solutions*), tzn. wartości współczynników w modelu są ściągane (*shrinkage*) w kierunku zera, z siłą uzależnioną od wartości λ. Dlatego też LASSO, poza regularyzacją, służy również do selekcji zmiennych (*feature selection*) w ten sposób, że parametry z wartością bliską zeru są eliminowane z modelu.

W przypadku kryterium L2 (3), składnik kary uwzględnia sumę kwadratów współczynników (4), ich wielkość zmniejsza się wraz ze wzrostem wartości λ, lecz nie tak bardzo jak w kryterium L1, stąd selekcja zmiennych nie jest możliwa [7].

$$RIDGE = \left(\sum_{i=1}^{n} y_i^{'} - \beta_0^{'} - \sum_{j=1}^{p} \beta_j^{'} x_{ij}^{'} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^{'2} \qquad (3)$$

$$\sum_{j=1}^{p} \beta_j^{'2} \leq \lambda \qquad (4)$$

Metoda sieci elastycznych łączy ideę regresji grzbietowej i LASSO w taki sposób, że możliwa jest selekcja zbioru zmiennych objaśniających [12].

Podsumowując, regularyzacja w uczeniu maszynowym to metoda zapobiegająca przetrenowaniu modelu służąca do jego lepszej generalizacji dzięki wprowadzeniu parametru kary, który ogranicza wariancję parametrów tak aby minimalizować estymowany błąd i którego wartość dobierana jest za pomocą walidacji krzyżowej [8].

2. Regularyzacja we wnioskowaniu bayesowskim

Wykorzystanie podejścia bayesowskiego do analizy dużych zbiorów danych zyskuje na popularności od czasu pojawienia się szybkich komputerów o dużej mocy obliczeniowej. Podejście to umożliwia wykorzystanie zarówno danych eksperymentalnych, jak i *a priori* danych literaturowych w procesie budowania modelu. W zależności od ilości informacji model bayesowski odpowiednio waży oba źródła danych w ten sposób, że konstruowana jest funkcja wiarygodności (*likelihood function*) – rozkład prawdopodobieństwa zaobserwowanych danych warunkowany zestawem parametrów. W przypadku wielowymiarowych zbiorów danych zastosowanie kryterium największej wiarygodności (*maximum likelihood*) prowadzi do obciążenia i dużej wariancji parametrów modelu. Rozwiązaniem w tej sytuacji jest zastosowanie penalizowanego prawdopodobieństwa (*penalized likelihood*), które wprowadza człon kary (i) w formie wiedzy *a priori* na temat rozkładu parametrów (np. średnia zero i standardowe odchylenie równe 1) lub też (ii) szacowanych bezpośrednio z danych (poprzez hiperparametry, odchylenie standardowe oszacowane jest bezpośrednio z danych) [9]. Oba rozwiązania prowadzą do powstania hierarchicznej struktury modelu [10].

Regresja LASSO posiada swoją bayesowską interpretację. W tym przypadku, na współczynniki modelu nałożony jest rozkład podwójnie wykładniczy (*double exponential prior,* tzw. rozkład Laplace'a). Zakłada on parametr lokalizacji (*location parameter*) równy zero i parametr skali (*scale parameter*) większy od zera w ten sposób, że w porównaniu do rozkładu normalnego, masa prawdopodobieństwa jest bardziej skoncentrowana wokół zera. Bayesowska interpretacja LASSO jest przykładem metody opartej na penalizowanym prawdopodobieństwie z rozkładem log-normalnym (*LASSO based on the penalized log-likelihood*) [11].

W badaniach naukowych, w których generowane są dane wielowymiarowe, bardzo łatwo o pozorne zależności (*spurious correlations*) wynikające z błędów pomiarowych czy wielokrotnego testowania (*multiple testing*). Stąd, zastosowanie metod opartych na regularyzacji wykorzystujących parametr kary, którego wartość szacowana jest na podstawie walidacji krzyżowej, jak też tych wykorzystujących penalizowane prawdopodobieństwo, jest podejściem rekomendowanym dla analizy danych wielowymiarowych.

3. Hierarchiczne modele liniowe

Hierarchiczne modele liniowe (*multilevel models, hierarchical linear models, mixed effects models*), znane również jako modele wielopoziomowe [13], podobnie jak inne modele liniowe opisują związek pomiędzy zmienną zależną (objaśnianą), a zmiennymi niezależnymi (objaśniającymi). Znajdują zastosowanie, gdy zbiór danych ma strukturę hierarchiczną, czyli gdy dane pogrupowane są względem jednej lub kilku zmiennych. Dzięki temu, że wpływ grupowania traktowany jest jako efekt losowy, modele hierarchiczne pozwalają uwzględnić strukturę zależności pomiędzy zmiennymi grupującymi obserwacje. Stanowią zatem użyteczną metodę analizy danych uwzględniającą ich wielopoziomową strukturę wynikającą z natury eksperymentu (*study design*), w którym obserwacje (pomiary) oraz jednostki eksperymentalne (*unit of analysis*) pogrupowane są zgodnie z nadrzędnym schematem. Hierarchiczne struktury danych generowane są zarówno w dyscyplinach nauki tj. nauki społeczne, polityczne, ekonomiczne, przyrodnicze, szeroko pojętym biznesie (finanse, zarządzanie, marketing) [14,15], jak też w naukowych badaniach eksperymentalnych. Hierarchiczna struktura danych generowanych w obszarze nauki jest typowa dla badań przekrojowych (*cross-sectional study)*, badań podłużnych (*longitudinal study*) oraz badań z powtarzalnym pomiarem (*repeated-measures study*) [16].

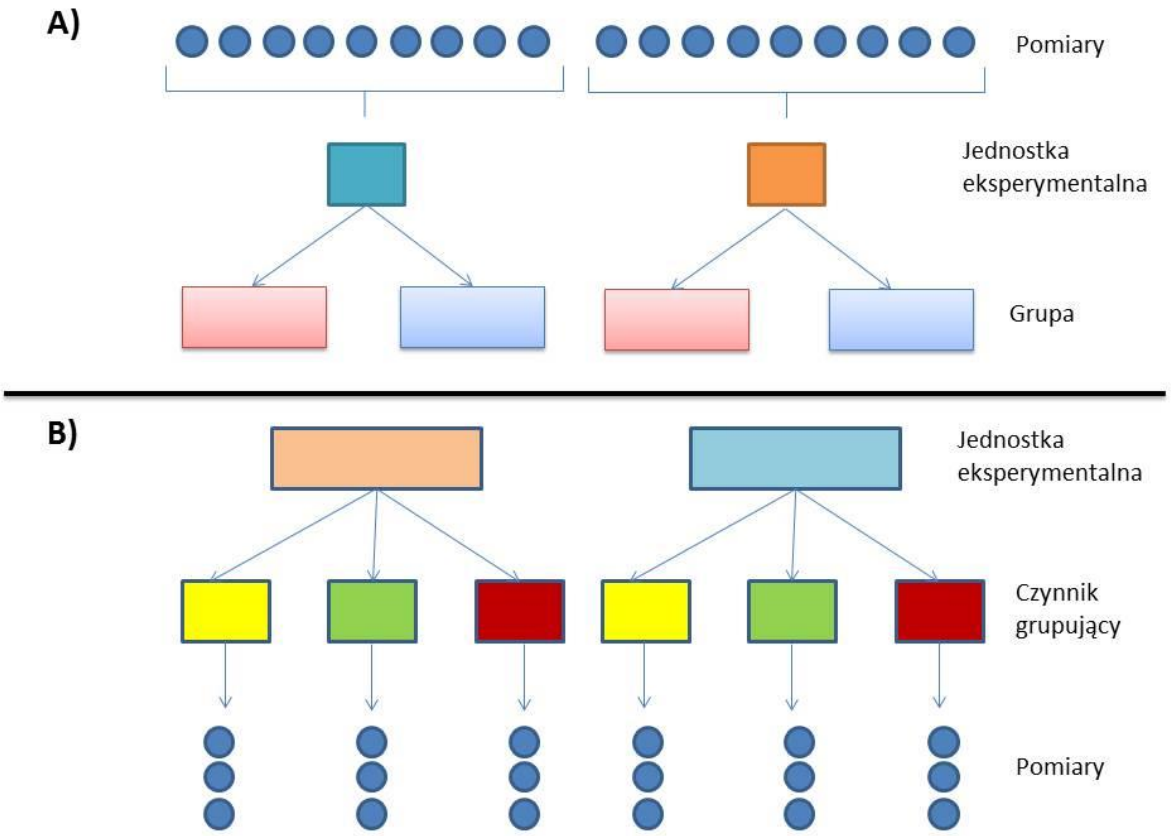
3.1 Hierarchiczna struktura danych w badaniach naukowych

Wielopoziomowość danych w badaniach przekrojowych (*cross-sectional study*) najczęściej jest skutkiem zbierania materiału biologicznego w określonych zbiorowościach (*clustered data*). Zmienna zależna jest mierzona dla każdej jednostki eksperymentalnej (poziom 1) biorącej udział w badaniu, a te z kolei mogą być zgrupowane lub zagnieżdżone w odpowiednich grupach/klastrach (poziom 2) (Rys. 1A).

Hierarchia danych otrzymanych w badaniach z powtarzalnym pomiarem (*repeated-mearures data*) wynika z przeprowadzenia wielu pomiarów w czasie na pojedynczej jednostce eksperymentalnej w obrębie czynnika grupującego (*within-subject factor*) np. czas, lokalizacja (Rys. 1B).

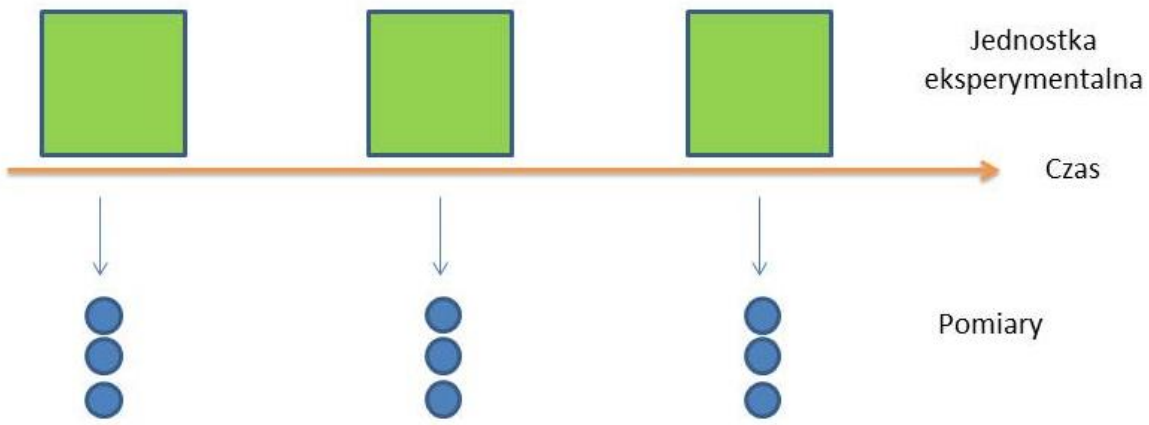
Z kolei, hierarchia w strukturze danych podłużnych (*longitudinal data*) dotyczy tzw. zależności serii (*serial dependency*) i polega na wielokrotnych pomiarach zmiennej zależnej przez dłuższy okres dla każdej jednostki eksperymentalnej (Rys. 2) [17].

We wszystkich powyższych układach eksperymentalnych, hierarchiczna struktura danych klasyfikuje jednostki eksperymentalne (poziom 1) na najniższym poziomie hierarchii. Pomiary (poziom 2) są następnie zagnieżdżone w jednostkach eksperymentalnych, a te z kolei na kolejnym wyższym poziomie hierarchii są zgrupowane w obrębie czynnika grupującego (poziom 3). W modelu wielopoziomowym grupy nie muszą być równoliczne, a dane nie muszą być zrównoważone (*balanced data*), oraz możliwe są różne poziomy zmienności [18,19].



Rys. 1 Struktura badania przekrojowego (A), w którym jednostki eksperymentalne przynależą do określonego czynnika grupującego (np. jednostka chorobowa, płeć, wiek), na podstawie którego klasyfikuje się je do odpowiednich grup (np. zdrowi/chorzy). Struktura badania z powtarzalnym pomiarem (B), w którym pomiary

znajdują się na najniższym poziomie hierarchii (poziom 1), i które są następnie zagnieżdżone w obrębie czynnika grupującego (np. czas, lokalizacja) (poziom 2) przynależącego do danej jednostki eksperymentalnej.



Rys. 2 Struktura badania podłużnego, w którym dla grup jednostek eksperymentalnych wykonywane są pomiary w równych odstępach czasowych.

Wielopoziomowa struktura danych pozwala zatem założyć większe podobieństwo otrzymanych wyników w obrębie danej grupy (lub czynnika grupującego) niż pomiędzy grupami. Można zatem zaobserwować brak niezależności obserwacji oraz większe prawdopodobieństwo korelacji zmiennych w obrębie grupy (tzw. wewnątrzgrupowa korelacja) niż między grupami. Ponadto, informacje zawarte w klastrach nie są całkowicie niezależne, a efektywna wielkość próby jest mniejsza niż całkowita liczba obserwacji we wszystkich klastrach [20].

W konsekwencji, do analizy danych o strukturze hierarchicznej nie można zastosować klasycznego modelu regresji liniowej, ponieważ zastosowanie tej metody, w przypadku wielopoziomowej struktury danych, podważa podstawowe założenia regresji liniowej takie jak te dotyczące [21]:

a) identycznej wariancji reszt składnika losowego (ε) dla każdej *i*-tej obserwacji (homoscedastyczność) (wariancja reszt składnika losowego jest taka sama dla wszystkich obserwacji, opisuje na ile dobrze model charakteryzuje dane; błędy przewidywania powinny być podobne w każdym przedziale zmiennej niezależnej; niespełnienie tego założenia skutkuje obniżeniem precyzji oszacowania).

$$\mathrm{var}(\varepsilon_i) = \sigma^2$$

b) kowariancji pomiędzy błędami losowymi pochodzącymi od różnych obserwacji (kowariancja równa zero).

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$$

c) braku autokorelacji reszt składnika losowego (błędy przewidywania rzeczywistej wartości zmiennej zależnej są niezależne od siebie a ich rozkład jest losowy).

c) rozkładu normalnego reszt składnika losowego (średnia równa zero i stała wariancja).

$$\varepsilon = \mathrm{N}(0, \sigma^2)$$

Zastosowanie regresji liniowej, w sytuacji gdy dane charakteryzują się wielopoziomową strukturą (a więc gdy powyższe założenia nie są spełnione), wpływa na otrzymane wyniki, a w konsekwencji na wnioski wysuwane z eksperymentu. Z kolei, ignorowanie wielopoziomowej struktury danych często prowadzi do dezagregacji danych albo ich agregacji na poziomie grupowym. W przypadku dezagregacji, dane opisują zależności w próbie złożonej z wielu obserwacji, dla których zakłada się niezależność, nie uwzględniając przy tym korelacji związanej ze zgrupowaniem danych. Agregacja danych opisuje natomiast związek pomiędzy średnimi grupowymi wszystkich predyktorów oraz zmiennej zależnej w grupach. Pozwala to na oszacowanie międzygrupowych różnic predyktora wobec zmiennej zależnej, jednakże nierówne liczebności obserwacji w klastrach mogą faworyzować wyniki na stronę klastra o większej liczebności (należy im się większa waga w analizie). Ponadto, rozważanie jedynie średnich grupowych powoduje utratę informacji pochodzących z indywidualnych obserwacji [20]. Oba podejścia generują odmienną estymację współczynników równania a wielopoziomowa struktura danych w tym przypadku, wyklucza zastosowanie regresji liniowej w procedurze wnioskowania statystycznego prowadząc do błędnych wniosków [22].

Alternatywą dla metod regresji liniowej, gdy dane mają rozkład normalny, są modele hierarchiczne, które umożliwiają modelowanie danych zarówno na poziomie jednostkowym (pojedynczej obserwacji), jak i grupowym. Uogólnione liniowe modele hierarchiczne (*Hierarchical Generalized Linear Models*, HGLM), jak też uogólnione

równania estymujące (*Generalized Estimating Equations*, GEE) są natomiast rozszerzeniem modeli hierarchicznych dla różnego typu zmiennych zależnych, których rozkład odbiega od rozkładu normalnego [20]. W niniejszej rozprawie doktorskiej zostanie przedstawiona m.in. koncepcja modeli hierarchicznych oraz jej zastosowania do modelowania danych „omicznych".

4. Liniowy model hierarchiczny dla efektów stałych

4.1 Idea wielopoziomowości w strukturze danych

Hierarchiczne modele liniowe cechują się większą złożonością w porównaniu do tradycyjnych modeli regresji liniowej, gdyż uwzględniają zarówno wielopoziomową strukturę danych, jak też indywidualny i grupowy związek między predyktorami. W zależności od poziomu danych, klasyczny model hierarchiczny identyfikuje trzy rodzaje równań w zbiorze danych:

1) równanie dla każdej grupy (poziom 1),

2) równanie opisujące grupową (populacyjną) strukturę danych (poziom 2),

3) równanie opisujące łączne efekty poziomu 1 i 2 (model hierarchiczny).

Efekt stały najczęściej opisywany jest przez zmienną kategoryczną obejmującą wszystkie klastry w badaniu i opisującą zależności między zmienną objaśnianą a czynnikami stałymi. Efekty stałe są reprezentowane przez nieznane stałe parametry.

Postać liniowego modelu hierarchicznego na poziomie 1 (dla jednego czynnika) można zapisać jako:

$$y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \text{ dla } i = 1, 2, \ldots, n_j \qquad j = 1, 2, \ldots, m \qquad (5)$$

gdzie:

y_{ij} – zmienna objaśniana dla i-tej obserwacji w j-tej grupie (klastrze),

β_{0j} – współczynnik przecięcia (stała) w j-tej grupie na poziomie 1,

β_{1j} – współczynnik nachylenia w j-tej grupie na poziomie 1,

ε_{ij} – błąd dla i-tej obserwacji w j-tej grupie na poziomie 1,

m – liczba klastrów (grup),

n_j – liczba obserwacji w j-tym klastrze.

Przy założeniu, że jednostki eksperymentalne stanowią losową próbę z populacji, zarówno β_{0j} jak i β_{1j} obserwowane na poziomie 1 przyjmują postać zmiennych losowych związanych z międzyosobniczym zróżnicowaniem.

Ponieważ w hierarchicznych modelach liniowych wnioskowanie na temat współczynników modelu odbywa się z poziomu populacji, postać liniowego modelu hierarchicznego na poziomie 2 opisuje związek grupowych estymatorów β_{0j} i β_{1j} (poziom 1) z ich populacyjnymi odpowiednikami estymowanymi na poziomie 2. Przy założeniu, że wartości stałej β_{0j} z poziomu 1 dla wszystkich zaobserwowanych grup tworzą rozkład losowy wokół populacyjnej wartości stałej y_{00}, a wartości współczynników β_{1j} z poziomu 1 dla tych grup tworzą losowy rozkład wokół populacyjnej wartości współczynnika nachylenia y_{10}, postać liniowego modelu hierarchicznego na poziomie 2 (dla jednego czynnika) można zapisać jako:

$$\beta_{0j} = y_{00} + \mu_{oj} \qquad (6)$$

$$\beta_{ij} = y_{10} + \mu_{1j} \qquad (7)$$

gdzie:

y_{00} – populacyjny współczynnik przecięcia (stała),

y_{10} – populacyjny współczynnik nachylenia,

μ_{0j} – losowy błąd odchylenia współczynnika przecięcia w *j*-tej grupie od populacyjnego współczynnika przecięcia,

μ_{1j} – losowy błąd odchylenia współczynnika nachylenia w *j*-tej grupie od populacyjnego współczynnika przecięcia.

Grupowa wartość β_{0j} jest efektem addytywnym złożonym z populacyjnej wartości stałej (y_{00}) oraz losowego błędu odchylenia μ_{0j} od y_{00}. Podobnie, grupowa wartość β_{ij} jest efektem addytywnym, w skład którego wchodzi populacyjna wartość współczynnika nachylenia (y_{10}) oraz losowy błąd odchylenia μ_{1j} od y_{10}.

Równanie opisujące grupowe i populacyjne efekty obserwowane na poziomie 1 i 2 tworzą hierarchiczny model liniowy postaci:

$$y_{ij} = y_{00} + \mu_{10} X_{ij} + \left(\mu_{0j} + \mu_{1j} X_{ij} + \varepsilon_{ij}\right) \qquad (8)$$

Addytywny charakter wariancji reszt modelu łączy składniki błędu pochodzące z poziomu 1 (ε_{ij}) i poziomu 2 (μ_{0j}i $\mu_{1j}X_{ij}$). Wartości ε_{ij} są różnicą pomiędzy obserwowaną a przewidywaną wartością zmiennej zależnej y_{ij}, natomiast μ_{0j} i μ_{1j} obrazują różnice między grupowymi a populacyjnymi wartościami współczynników przecięcia i nachylenia [22].

5. Liniowy model hierarchiczny z efektem stałym oraz losowym

5.1 Efekty stałe i losowe w modelu

Efekty stałe (*fixed effects*) są to zmienne niezależne, których wszystkie możliwe poziomy są zdefiniowane przez badacza. Określają zatem wpływ zdefiniowanych poziomów zmiennej niezależnej na zmienną zależną. Celem wprowadzenia efektów stałych do modelu jest porównanie średnich wartości zmiennej niezależnej na różnych poziomach.

W liniowych modelach mieszanych, oprócz efektów stałych znajdują się również efekty losowe (*random effects*). Są to współczynniki modelu, dla których ocena wartości nie jest istotna, natomiast istotna jest ocena ich zmienności. Zmienność ta jest analizowana zakładając, że współczynniki te są realizacjami pewnej zmiennej losowej. Efekty losowe nie podlegają kontroli badacza. Zakłada się, że poziomy z nimi związane są dobierane w sposób losowy z nieskończonej populacji możliwych poziomów. Celem wprowadzenia efektów losowych do modelu hierarchicznego jest oszacowanie zmienności (wariancji, δ^2) tych efektów, a wnioski rozciągane są na całą populację możliwych poziomów efektów losowych. W modelu hierarchicznym, efekty losowe odpowiedzialne są za zmienność międzyosobniczą i korelację wewnątrzosobniczą, jednak nie zmieniają się w obrębie obserwacji dla danej osoby [23]. Przykładem efektu losowego w modelu hierarchicznym jest np. wpływ odmiany pszenicy (zmienna niezależna) na wielkość szkód (zmienna zależna) wyrządzonych przez szkodniki na różnych poletkach (zmienna niezależna). Z uwagi na to, że w eksperymencie trudno jest uwzględnić wszystkie możliwe odmiany pszenicy, zmienna to ma charakter losowy.

Zaklasyfikowanie danego efektu jako stałego bądź losowego jest trudne oraz zależne od charakteru eksperymentu i sposobu doboru poziomów danego czynnika.

5.2 Źródła zmienności

5.2.1 Zmienność wewnątrzosobnicza

Zmienność wewnątrzosobnicza (*inter-individual variability*) charakteryzuje zmienność określonego pomiaru dla pojedynczej osoby w badaniu. Zmienność wewnątrzosobnicza jest charakterystyczna dla badań z powtarzalnym pomiarem,

opisując jak bardzo pomiar dla danej osoby zmienia się w czasie (innymi słowy zmienność ta opisuje średnią zmianę w wartości pomiaru dla pojedynczej osoby w próbie). W badaniach z powtarzalnym pomiarem, uwzględnienie zmienności wewnątrzosobniczej jest możliwe jedynie poprzez ujęcie w modelu efektu losowego, w celu modelowania korelacji pomiędzy obserwacjami. Struktura zmienności pomiędzy obserwacjami pochodzącymi od tej samej osoby ma określony charakter [24].

5.2.2 Zmienność międzyosobnicza

Zmienność międzyosobnicza (*intra-individual variability*; *between-subject variability*) charakteryzuje zmienność określonego pomiaru między osobami, (jeśli zmienna zależna jest zmienną ciągłą) lub między grupami (gdy zmienna zależna jest zmienną kategoryczną). Tego rodzaju zmienność obserwowana jest zarówno w badaniach z powtarzalnym pomiarem, jak i badaniach obserwacyjnych. Zmienność międzyosobnicza określa czy istnieją różnice między wartościami zmiennej zależnej (gdy pomiar wyrażony jest na skali numerycznej) dla poszczególnych osób lub pomiędzy grupami (np. kobiety *versus* mężczyźni; zdrowi *versus* chorzy) [24].

5.3 Postać modelu

Model hierarchiczny z efektem stałym i losowym jest zapisywany w postaci generycznej:

$$Y_j = X_j\beta + Z_j b_j + \varepsilon_j \text{ dla} j = 1, 2, \ldots, m \qquad (9),$$

co odpowiada poniższemu zapisowi w postaci macierzy:

$$\begin{bmatrix} Y_{j1} \\ Y_{j2} \\ \vdots \\ Y_{jn_j} \end{bmatrix} = \begin{bmatrix} 1 & X_{j1.1} & \cdots & X_{j1.p} \\ 1 & X_{j2.1} & \cdots & X_{j2.p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{jn_j.1} & \cdots & X_{jn_j.p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} Z_{j1.1} & \cdots & Z_{j1.q} \\ \vdots & \ddots & \vdots \\ Z_{jn_j.1} & \cdots & Z_{jn_j.q} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{jq} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{jn_j} \end{bmatrix}$$

gdzie:

Y_j – wektor wartości zmiennej objaśnianej o wymiarze n_j x 1, gdzie n_j to liczba obserwacji w j - tym klastrze $i = 1,2, \ldots n_j$,

β – wektornieznanych parametrów efektów stałych o wymiarze (p + 1) x 1,
X_j – macierz p predyktorów związanych z efektami stałymi o wymiarze n_j x (p + 1),
b_j – wektor nieznanych parametrów efektów losowych o wymiarze q x 1,
Z_j – macierz$q$ predyktorów związanych z efektami losowymi o wymiarze n_j x q,
ε_j – wektor dla reszt o wymiarze n_j x 1.

Macierz X_j *i* Z_j są macierzami znanych wartości q predyktorów, natomiast b_j to nieobserwowane realizacje zmiennej losowej $b_j \sim N(0, D)$. Zakłada się także, że $\varepsilon_j \sim N(0, R_j)$ oraz $b_j \perp \varepsilon_j$ dla każdego j= 1, 2, ... , m; (b_j jest niezależne od ε_j w każdej grupie). Założenie niezależności efektów losowych b_j oraz błędów losowych ε_j jest typowym założeniem dla klasycznych liniowych modeli wielopoziomowych.

Blokowa postać macierzy D i R_j przyjmuje postać:

$$D = Var(b_j) = \begin{bmatrix} Var(b_{1j}) & \text{cov}(b_{1j}, b_{2j}) & \dots & \text{cov}(b_{1j}, b_{qj}) \\ \text{cov}(b_{1j}, b_{2j}) & Var(b_{2j}) & \dots & \text{cov}(b_{1j}, b_{qj}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_{1j}, b_{qj}) & \text{cov}(b_{1j}, b_{qj}) & \dots & Var(b_{qj}) \end{bmatrix}$$

$$R_j = Var(\varepsilon_j) = \begin{bmatrix} Var(\varepsilon_{1j}) & \text{cov}(\varepsilon_{1j}, \varepsilon_{2j}) & \dots & \text{cov}(\varepsilon_{1j}, \varepsilon_{mj}) \\ \text{cov}(\varepsilon_{1j}, \varepsilon_{2j}) & Var(\varepsilon_{2j}) & \dots & \text{cov}(\varepsilon_{1j}, \varepsilon_{mj}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\varepsilon_{1j}, \varepsilon_{mj}) & \text{cov}(\varepsilon_{1j}, \varepsilon_{mj}) & \dots & Var(\varepsilon_{mj}) \end{bmatrix}$$

gdzie:

D – dodatnio określona i symetryczna macierz kowariancji dla efektów losowych o wymiarze q x q,
R_j – dodatnio określona i symetryczna macierz kowariancji dla reszt w j-tym klastrze o wymiarach n_j x n_j.

Podsumowując, liniowy model hierarchiczny z efektami losowymi posiada w swojej strukturze część odpowiedzialną za wyjaśnienie zależności zmiennej Y_j od wybranych zmiennych niezależnych X_j w populacji oraz część, która odpowiada za wyjaśnienie zależności zmiennej Y_j od wybranych zmiennych Z_j w danym klastrze.

Innymi słowy, powyższa postać liniowego modelu hierarchicznego zakłada istnienie określonej zależności pomiędzy zmienną zależną a zmiennymi niezależnymi w populacji, jak też istnienie losowych zaburzeń tych zależności między klastrami [25].

5.4 Macierze kowariancji

Obecność efektów losowych w liniowym modelu hierarchicznym z efektami losowymi, wymusza założenie odpowiedniej struktury macierzy w celu zbadania zależności pomiędzy parametrami (efektami). Wyróżniamy matrycę kowariancji dla efektów losowych oraz dla błędu [26].

Najbardziej popularnymi macierzami kowariancji dla efektów losowych są (i) macierz niestrukturalna (*Unstructured Covariance Matrix*) oraz (ii) macierz komponentów wariancji (diagonalna) (*Variance Components*) [27]. W celu uproszczenia estymacji nakłada się na macierz *D* określoną strukturę po to, żeby ograniczyć liczbę parametrów do oszacowania jak też ułatwić interpretację wyników. Założenie o postaci macierzy zależy przede wszystkim od rodzaju eksperymentu oraz wygenerowanej struktury danych. Najbardziej popularnymi macierzami kowariancji dla błędu są macierz komponentów wariancji (*Variance Components*), macierz symetrii związku (*Compound Symmetry*), macierz autoregresji I rzędu (*Autoagressive (1)*), macierz Toeplitza (*Toeplitz matrix*).

W przypadku estymacji zmienności międzyosobniczej dla matrycy kowariancji danych wielowymiarowych często używany jest rozkład Wishart'a (*Wishart distribution*). Za pomocą tego rozkładu możliwe jest modelowanie jedynie matrycy precyzji (a nie matrycy kowariancji). Przy założeniu, że parametry θ są opisane przy pomocy wielowymiarowego rozkładu normalnego ze średnią μ (p x 1) i matrycą precyzji $\Omega = \sum^{-1}$ (p x p) wokół zmienności międzyosobniczej, tj. $\theta \sim MVN(\mu, \Omega)$, odwrotność matrycy kowariancji zmienności międzyosobniczej (Ω^{-1}) pochodzi z rozkładu Wisharta (*wish*), tj., $\Omega^{-1} \sim wish(\rho\Sigma, \rho)$, gdzie Σ to p x p-elementowa symetryczna, dodatnio określona macierz wartości oczekiwanych matrycy kowariancji zmienności międzyosobniczej, a ρ oznacza liczbę stopni swobody [28].

W badaniach przy użyciu symulacji, szczególnie ważne jest, aby generowana przestrzeń losowych wartości parametrów była zgodna z rzeczywistością, tj. nie zawierała wartości ujemnych (była dodatnio określona). Poza rozkładem Wishart'a, który estymuje matrycę kowariancji dla zmienności międzyosobniczej, istnieją też

takie rozkłady jak rozkład gamma, który służy do modelowania zmienności resztkowej lub wielowymiarowy rozkład log-normalny za pomocą którego możliwe jest modelowanie efektów stałych w modelu [29].

6. Idea wnioskowania bayesowskiego

Metody bayesowskie definiują model statystyczny w kategoriach probabilistycznych uwzględniając dotychczasową wiedzę na temat modelowanego zjawiska/procesu w formie rozkładu *a priori* nałożonego na parametry modelu. Innymi słowy, wnioskowanie bayesowskie polega na uaktualnieniu wiedzy o nieznanym rozkładzie parametrów θ dysponując modelem opartym o dane rzeczywiste (obserwacje y), dla którego uaktualnienie wiedzy przedstawione jest w postaci rozkładu *a posteriori* $P(\theta|y)$ [30]. Uaktualniona wiedza na temat modelowanego zjawiska $P(\theta|y)$ w świetle posiadanych danych y, pochodzi z: (i) rzeczywistego rozkładu danych $P(y)$, (ii) założonego przez eksperymentatora, hipotetycznego rozkładu nałożonego na parametry modelu $P(\theta)$ oraz (iii) ich łącznego rozkładu $P(y|\theta)$ [31] i ma swoje odbicie w twierdzeniu Bayesa (10).

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)} \qquad (10)$$

Łączny rozkład obserwacji y i parametrów modelu $P(y|\theta)$ to funkcja wiarygodności (*likelihood function*), która odzwierciedla prawdopodobieństwo danych na podstawie ich parametrów, natomiast $P(y)$ to stała normalizująca (niezależna od θ) której zadaniem jest sumowanie estymowanych wartości prawdopodobieństwa do 1. Stąd, twierdzenie Bayesa można zapisać również w formie uproszczonej (11):

$$P(\theta|y) \propto P(\theta)P(y|\theta) \qquad (11)$$

Dzięki twierdzeniu Bayesa, w miarę zbierania nowych dowodów (danych) oraz rozkładowi *a priori* wyrażającemu naszą wiedzę na temat modelowanego zjawiska, prawdopodobieństwo *a posteriori* może być wielokrotnie uaktualnione. Wykorzystanie informacji spoza obserwowanych danych stanowi podstawową różnicę odróżniającą metody bayesowskie od klasycznych metod częstościowych (*frequentist methods*).

6.1 Dobór rozkładu *a priori*

Założenie na temat rozkładu *a priori* parametrów modelu powinno obejmować wszystkie prawdopodobne wartości θ, aczkolwiek bez koncentrowania się wokół prawdziwych wartości. Idea wnioskowania bayesowskiego zakłada, że informacja o θ

zawarta w rzeczywistych danych powinna przeważyć „każde rozsądne zdefiniowanie rozkładu *a priori*" [32]. Innymi słowy, wpływ założonego rozkładu *a priori* jest tym mniejszy im więcej informacji zawierają dane. Z kolei, gdy dane którymi dysponujemy zawierają dużo szumu informacyjnego, wpływ naszej dotychczasowej wiedzy o modelowanym zjawisku będzie większy, zatem założony rozkład *a priori* będzie miał większy wpływ na uzyskane wyniki. Wśród rozkładów *a priori* można wyróżnić: nieinformatywne, mało informatywne i informatywne.

6.1.1 Nieinformatywne rozkłady *a priori*

Nieinformatywne rozkłady *a priori* (*non-informative, diffuse, flat,objective, vague, reference priors*) „*pozwalają danym mówić za siebie*", stosowane są gdy nie zakładamy żadnej wiedzy na temat parametrów przed zebraniem danych, stąd rozkłady te charakteryzują się małym wpływem na wyniki oszacowania parametrów modelu. Są powszechne w przypadku, gdy nie możemy albo nie chcemy przyjmować założeń na temat parametrów θ. Przykładami nieinformatywnych rozkładów *a priori* są rozkłady jednostajne, które obejmują wszystkie możliwe wartości parametrów θ, np. rozkład normalny o średniej μ i dużej wariancji δ [33].

Mało informatywne rozkłady *a priori* (*weekly-informative priors*) koncentrują się wokół pewnej wartości centralnej dopuszczając duży zakres rozproszenia. Stanowią pewnego rodzaju alternatywę dla nieinformatywnych rozkładów *a priori* dzięki założeniu, że zarówno wartości typowe (centralne) jak i nietypowe, są tak samo prawdopodobne. Założenie tego typu rozkładów *a priori* w danych nie zakłada żadnej konkretnej informacji, a jedynie „nakierowuje" rozkład danych w stronę wnioskowania opartego na niekompletnej wiedzy o modelowanym procesie. Gelman sugeruje stosowanie połowy rozkładu t (*half-t distribution*) lub połowy rozkładu Cauchy'ego (*half-Cauchy distribution*), jako mało informatywnych rozkładów jednostajnych ciągłych, aczkolwiek o ograniczonym zakresie rozproszenia [34].

6.1.2 Informatywne rozkłady *a priori*

Założenie informatywnych rozkładów *a priori* opiera się na dotychczasowej wiedzy na temat modelowanego zjawiska i stanowi dodatkową informację w modelu. Informacja na temat informatywnych rozkładów *a priori* opisujących modelowane zjawisko zazwyczaj pochodzi albo z literatury albo z wcześniejszej analizy podobnego

typu danych [35]. Niemniej, aby korzystać w pełni z informatywnego rozkładu *a priori*, trzeba posiadać wiedzę o rozkładzie interesujących nas, lecz nieznanych parametrach modelu (pochodzącą najlepiej z meta-analizy lub hierarchicznego modelowania bayesowskiego) lub posiadać wiedzę na temat wyraźnych granic prawdopodobieństwa (*probability bounds*), dzięki którym można uzyskać rozkład parametrów. Informacja *a priori* w modelu statystycznym często uważana jest za subiektywną lub wręcz „*nie-naukową czy też pozbawioną postaw naukowych*". Jednakże, jeśli dysponujemy pewną informacją opisującą parametry modelu jeszcze przed zebraniem danych, to taka wiedza powinna zostać przez nas uwzględniona w modelu [36].

Możliwe jest również losowanie parametrów rozkładu *a priori* z innych rozkładów *a priori*, tzw. super-priorów (*hyperpriors*). Stosowanie super-priorów pozwala na określenie niepewności (*uncertainty*) wokół super-parametrów (*hyperparameters*). Nałożenie kolejnego rozkładu na parametry rozkładu *a priori* jest dodaniem kolejnego poziomu złożoności w hierarchicznej strukturze danych, gdy przypuszczamy, że istnieje wspólna cecha zaobserwowanych danych [32].

6.2 Funkcja wiarygodności

Funkcja wiarygodności (*likelihood function*) jest skonstruowana po zaobserwowaniu danych. Funkcja ta jest wspólną funkcją rozkładu prawdopodobieństwa (*joint probability function*) danych w funkcji parametrów, traktując dane zaobserwowane jako ustalone wartości. Zakładając niezależność obserwacji, $y = (y_1, \ldots, y_n)$, funkcja wiarygodności przyjmuje postać:

$$L(\theta|y) = P(y_1 \ldots y_n|\theta) = \prod_{i=1}^{n} P(y_i|\theta) \qquad (12)$$

W ujęciu bayesowskim, informacja o parametrach θ pochodząca z danych jest zawarta w funkcji wiarygodności, $P(y|\theta)$ (12). Celem jest znalezienie takich wartości parametrów θ dla których funkcja ta osiąga największą wartość [36].

6.3 Estymacja parametrów modelu

Metody bayesowskie służą do estymacji parametrów modelu oraz mogą być traktowane jako metoda komplementarna lub równoległa dla metod częstościowych. Estymacja parametrów w ujęciu bayesowskim opiera się na metodach symulacyjnych, u podstaw których leży losowanie dużej ilości prób z rozkładu charakteryzującego określone zjawisko/proces [37,38]. Główną zaletą symulacyjnych metod estymacji parametrów jest mniejsze ograniczenie liczebnością próby w porównaniu do metod częstościowych. Natomiast ograniczeniem może być fakt, że wynik może być obciążony przyjętymi wcześniej założeniami na temat modelowanego zjawiska. Podstawą metod symulacji stosowanych do bayesowskiej estymacji parametrów są tzw. łańcuchy Markova (*Markov chain Monte Carlo, MCMC*), które są obecnie najpowszechniej stosowaną techniką próbkowania przestrzeni parametrów. Wśród popularnych algorytmów losujących opartych na łańcuchach Markova można wyróżnić: algorytm Metropolis-Hastings oraz próbkowanie Gibbsa (*Gibbs sampling*) [39].

6.3.1 Łańcuchy Markova

W kontekście analizy bayesowskiej, generowanie łańcuchów Markova prowadzi do powstania wielu zbiorów parametrów o różnych wartościach. Zgodność parametrów wygenerowanych w każdym losowaniu z rzeczywistymi danymi jest

oceniana poprzez funkcję wiarygodności $P(y|\theta)$. W przypadku gdy wygenerowany łańcuch Markova osiąga tzw. stan stacjonarny, czyli taki w którym częstość występowania poszczególnych stanów parametrów θ jest zgodna z założoną wiedzą *a priori* na temat wartości tych parametrów można przypuszczać, że rozkład parametrów próby z rozkładu *a posteriori* jest reprezentatywny (łańcuchy Markova efektywnie eksplorują przestrzeń parametrów) i może zostać użyty do wnioskowania. Stan stacjonarny może zostać osiągnięty (łańcuch Markova osiąga zbieżność) pod warunkiem odpowiedniej ilości iteracji algorytmu. W praktyce, zaleca się stosowanie od kilku do kilkunastu tysięcy iteracji w obrębie co najmniej trzech łańcuchów, które inicjowane są za pomocą różnych wartości inicjujących. Należy podkreślić, że liczba iteracji zależy również od ilości parametrów w modelu [39]. Ocena zbieżności łańcucha Markova, dzięki której możemy ocenić czy wygenerowany ciąg liczb jest reprezentatywny dla rozkładu docelowego, opiera się na wizualnej ocenie (wykres typu *trace plot* lub wykres autokorelacji) lub wykorzystaniu statystyk temu dedykowanych.

Wśród wspomnianych statystyk, najbardziej powszechną jest metoda Brooksa-Gelmana-Rubina (*Brooks-Gelman-Rubin method*), u podstaw której leży oszacowanie stosunku wariancji wewnątrz łańcuchów do wariancji pomiędzy łańcuchami. W metodzie tej, łańcuch uznaje się za zbieżny, gdy stosunek ten jest bliski jedności ($R \leq 1.1$). Inną znaną statystyką jest wyliczenie efektywnej liczby losowań z symulacji n_{eff}, która charakteryzuje ilość iteracji algorytmu pochodzących z niezależnych losowań. Z uwagi na autokorelację w łańcuchach Markova, ilość ta jest zawsze mniejsza niż liczba przeprowadzonych symulacji, a jej optymalna wartość powinna wynosić co najmniej 100 [40]. Kolejną statystyką pomocną w ocenie zbieżności łańcucha jest błąd Monte Carlo (*Monte Carlo error*), który opisuje zmienność pomiędzy iteracjami dla konkretnej zmiennej będąc miarą precyzji jej estymacji [37].

6.4 Bayesowska ocena dopasowania modelu

Celem oceny dopasowania modelu jest sprawdzenie czy model jest dobrze dopasowany do danych (*goodness-of-fit*) oraz czy przewidywania na podstawie modelu mają sens.

Wykres dla reszt (*residual plot*) jest powszechną i znaną miarą oceny dopasowania modelu do danych. Reszty są definiowane jako różnica między wartością

obserwowaną a przewidzianą przez model i służą do lokalizacji obserwacji odstających, które w znacznym stopniu wpływają na wartości oszacowanych parametrów równania. Wizualna ocena własności predykcyjnych modelu (*Visual Posterior Predictive Check, VPC*) polega na wyznaczeniu z danych symulowanych 5, 50 i 95 percentyla z wykorzystaniem oszacowanych parametrów modelu, a następnie policzeniu tej samej statystyki dla danych eksperymentalnych. Z uwagi na dużą liczbę iteracji danych symulowanych, rozrzut wyników wyrażony jest poprzez 5 i 95 percentyl dla każdego z parametrów. Z otrzymanych statystyk konstruuje się wykres i wizualnie określa ich zgodność. Model o dobrych właściwościach predykcyjnych to taki, dla którego wartości eksperymentalne zawierają się w 95% przedziale ufności [38,41]. Kolejną bayesowską miarą oceny dopasowania modelu są tzw. bayesowskie wartości *p* (*Bayesian p-values, posterior predictive p-values*). Definiowane są one jako prawdopodobieństwa wyliczone na podstawie danych pochodzących z symulacji i informują o tym, jak często wartość konkretnej statystyki testowej (średnia, mediana, rozbieżność między przewidywaniami a wartościami obserwowanymi) jest większa lub mniejsza od wartości tej samej statystyki oszacowanej na podstawie obserwowanych danych. Innymi słowy, oszacowanie tej statystyki opiera się na policzeniu jak często wartość *a posteriori* badanej statystyki we wszystkich losowaniach MCMC jest wyższa względem statystyki zaobserwowanej. Otrzymana wartość zazwyczaj jest bliska 0,5 i jest interpretowana jako prawdopodobieństwo wystąpienia danego wyniku (w odróżnieniu do klasycznej interpretacji wartości *p*). Należy jednak podkreślić, że kryterium to nie powinno decydować o odrzuceniu czy akceptacji modelu ze względu na probabilistyczną naturę analizy bayesowskiej, która w kontekście częstościowej wartości *p* nie może być tak samo interpretowana [42,43].

Aby odpowiedzieć na pytanie czy przewidywania otrzymanego modelu mają sens, model należy zwalidować za pomocą walidacji zewnętrznej (*external validation*) lub krzyżowej (*cross-validation*). W przypadku tej ostatniej, podział zbioru danych w proporcjach np. 8:2 pozwala na estymację parametrów modelu na większym zbiorze, a następnie ocenę czy zbudowany model dobrze przewiduje wyniki na mniejszym zbiorze.

6.4.1 Porównanie modeli

Porównanie modeli stosowane jest w przypadku, gdy porównujemy ze sobą kilka alternatywnych modeli bądź analizę rozpoczynamy od zbudowania prostszego modelu stopniowo go rozbudowując. Istniej kilka miar, za pomocą których można ocenić dopasowanie modelu do danych [27].

Najpopularniejszą miarą jest dewiancja (*deviance*) (13), która jest iloczynem -2 i logarytmu z funkcji wiarygodności (prawdopodobieństwo danych przy założeniu parametrów modelu) [40] i opisuje błąd dopasowania (im mniejsza wartość tym mniejszy błąd dopasowania).

$$D(y,\theta) = -2\log P(y|\theta) \qquad (13)$$

Dewiancja jest również miarą, na podstawie której obliczane jest kryterium informacyjne Akaikego (*Akaike information criterion, AIC*) (14)

$$AIC = D(y,\theta) + 2k \qquad (14)$$

gdzie: k – liczba predyktorów w modelu.

Uwzględnienie ich w modelu obniża dewiancję o ilość odpowiadającą rozkładowi χ^2 z z-liczbą stopni swobody. Stąd, jeśli dodanie k – predyktorów do modelu redukuje dewiancję więcej niż k-razy, można przypuszczać, że nastąpiła znacząca poprawa zdolności predykcyjnych modelu.

Dodatkowo, dewiancja stanowi podstawę do obliczenia bayesowskiego kryterium informacyjnego (*Bayesian information criterion, BIC*) (15).

$$BIC = 2\log P(y|\theta) + k\log(N) \qquad (15)$$

gdzie N – liczebność próby.

Zastosowanie miar dewiancji i AIC do oceny dopasowania modeli hierarchicznych wiąże się z pewną niedogodnością wyznaczenia ilości parametrów. Ilość parametrów w modelu hierarchicznym zależy od stopnia uwspólniania (w przypadku całkowitego uwspólniania, zbiór parametrów θ odpowiada jednemu parametrowi; w przypadku braku uwspólniania są to θ niezależne parametry; w przypadku częściowego uwspólniania jest to wartość pomiędzy).

Obecnie, najpopularniejszą miarą dopasowania bayesowskiego modelu hierarchicznego (*Bayesian measure of fit or adequacy*) jest dewiacyjne kryterium informacyjne (*deviance information criterion, DIC*) (16), które charakteryzuje błąd predykcji zbioru walidującego i jest uznawane za generalizację AIC [44]:

$$DIC = \hat{D}_{avg}(y,\theta) + 2p_D \qquad (16)$$

gdzie: $\hat{D}_{avg}$ to uśredniona dewiancja z liczby wszystkich symulacji parametrów; p_D to efektywna liczba parametrów.

Pomimo przydatności DIC do porównania modeli, uważa się ją za miarę dość niestabilną o niejasnych podstawach teoretycznych i sugeruje się, że kryterium to powinno być raczej wskazówką niż miarą ostateczną. Ponadto, dobrą praktyką jest przedstawienie wyników porównania modeli za pomocą komplementarnych do DIC miar dopasowania modelu [45].

6.5 Programy do analizy bayesowskiej

Programy do bayesowskiej analizy danych oparte na łańcuchach Markova są darmowe i dostępne *on-line*. Wśród nich można wyróżnić (i) program *BUGS* (**B**ayesian inference **U**sing **G**ibbs **S**ampling), (Lunn, Jackson, Best, Thomas & Spiegelhalter, 2013; Lunn, Thomas, Best & Spiegelhalter, 2000) wersji *WinBUGS* i *OpenBUGS;* (ii) *JAGS* (**J**ust **A**nother **G**ibbs **S**ampler) (Plummer 2003, 2012), który jest rozszerzeniem programu *BUGS;* oraz (iii) *Stan* (Sampling Through Adaptive Neighbourhoods) [39], który w odróżnieniu od programu JAGS, używa hamiltonowskich łańcuchów Markowa (*Hamiltonian Monte Carlo, HMC*), i który uznaje się za bardziej efektywny niż próbniki programu JAGS czy BUGS zwłaszcza dla rozbudowanych modeli. Wszystkie powyższe programy działają efektywnie w połączeniu ze środowiskiem R [47] i Matlab (MATLAB 6.1, The MathWorks Inc., Natick, MA, 2000).

II. CELE PRACY

Celem niniejszej pracy było zastosowanie regresji LASSO oraz bayesowskich modeli hierarchicznych do modelowania danych chromatograficznych, danych omicznych pochodzących głównie z metabolomiki oraz wnioskowanie na temat (i) zdolności predykcyjnych modeli QSRR, (ii) dostępnych źródeł zmienności obserwowanych w profilach metabolomicznych w funkcji dostępnych kowariant, oraz (iii) zmian w intensywności metabolitów „w czasie" wraz z zaproponowaniem sposobu modelowania wielowymiarowych danych metabolomicznych zebranych w punktach czasowych.

Badania mają stanowić także wstęp do analizy innego typu danych „omicznych" jak np. profilowania miRNA do diagnostyki raka jajnika.

Cele szczegółowe:

1. Porównanie zdolności predykcyjnych modeli QSRR zbudowanych przy użyciu metod tj. PLS, LASSO oraz LASSO w połączeniu z regresją krokową dla nukleozydów i pteryn rozdzielanych na kolumnach typu HILIC wraz z wyznaczeniem modeli zależności struktura-retencja (QSRR) (publikacja nr 1).
2. Zaproponowanie bayesowskiego modelu hierarchicznego do modelowania danych pochodzących z celowanej analizy metabolomicznej, wnioskowanie na temat zmienności międzyosobniczej metabolitów w funkcji dostępnych kowariant tj. wiek, płeć, stan zdrowia (zdrowy/chory) oraz ich wpływ na stosunek szybkości wydalania nukleozydów do kreatyniny (publikacja nr 2).
3. Zaproponowanie strategii modelowania hierarchicznej struktury danych zebranych w punktach czasowych, pochodzących z niecelowanej analizy metabolomicznej w oparciu o eksperyment *in vivo,* z jednoczesną oceną zmian „w czasie" w profilu metabolomicznym na skutek wprowadzenia komórek nowotworowych do pęcherza moczowego (publikacja nr 3).

Ostatni wymieniony w pracy aspekt badawczy, który jest realizowany jako rozszerzenie zaplanowanych wcześniej badań obejmuje zaproponowanie bayesowskiego modelu hierarchicznego do modelowania danych pochodzących z profilowania mikro RNA wraz z oceną użyteczności wybranych mikro RNA do diagnostyki raka jajnika (praca będąca kontynuacją badań, będąca rozwinięciem

trzech opublikowanych prac wchodzących w skład rozprawy doktorskiej, aczkolwiek nie będąca podstawą osiągnięcia doktorskiego).

III. METODOLOGIA

1. Metodologia analityczna

Opisana w niniejszym rozdziale metodologia oznaczeń analitycznych dotyczy eksperymentów opisanych w publikacjach 1, 2, 3. Stanowi ona analityczne zaplecze, dzięki któremu wygenerowano dane, które w późniejszym etapie analizowano przy pomocy regularyzacji i bayesowskich modeli hierarchicznych. Szczegółowe rozwinięcie przedstawionej poniżej metodologii analitycznej znajduje się w opublikowanych pracach.

Do badania zależności struktura-retencja (*Quantitative Structure-(Chromatographic) Retention Relationships*, QSRR) wykorzystano 16 nukleozydów i 11 związków pterynowych. Synteza faz stacjonarnych typu HILIC tj. N,O-diaminofosfoamidowe (*N,O-dialkylphosphoramidate*) APC-10 i APC-18 (250 x 4,6 mm) została przeprowadzona w laboratorium Katedry Chemii Środowiska i Bioanalityki UMK. Trzecią fazę typu HILIC stanowiła komercyjnie dostępna faza z immobilizowaną sztuczną membraną IAM.PC.DD2 (150 x 4,6 mm). Analizy chromatograficzne standardów badanych związków zostały przeprowadzone przy użyciu zestawu Dionex Ultimate 3000 (Sunnyvale, CA, USA) wyposażonego w automatyczny podajnik próbek i detektor UV-Vis, w trybie elucji izokratycznej przy 4 testowanych składach fazy ruchomej: acetonitryl (95%, 90%, 85% i 80%) i woda. Każdy analit wprowadzano na kolumnę w trzech powtórzeniach. Wykorzystano równanie kwadratowe do modelowania zależności między współczynnikiem retencji a zawartością organicznego modyfikatora. Wartość $\log k_{ACN}$ przy 95% zawartości acetonitrylu użyto jako zmienną zależną w modelach QSRR. Do numerycznego opisu właściwości 27 badanych związków wykorzystano 1304 deskryptory wygenerowane w programie HyperChem 7.5 (HyperChem(TM) Professional 7.5, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA) i Dragon 5.0 (Taletesrl, DRAGON (Software for Molecular Descriptor Calculation) Version 5.0 – 2004 – http://www.talete.mi.it/).

Oznaczenia chromatograficzne, dotyczące celowanej analizy metabolomicznej nukleozydów (przedłożone w publikacji nr 2), zostały przeprowadzone z zastosowaniem techniki wysokosprawnej chromatografii cieczowej w odwróconym

układzie faz (RP HPLC) przy użyciu zestawu Agilent Technologies 1200 (Waldbronn, Germany) wyposażonego w pompę, automatyczny podajnik próbek i detektor fotodiodowy (*diode array detector*, DAD). Przed właściwą analizą chromatograficzną, w celu zatężenia związków z ugrupowaniami cis-diolowymi w próbkach moczu, przeprowadzono ekstrakcję do fazy stałej (*solid phase exctraction*, SPE) (Affi-gel 601, Bio-Rad, CA, USA) oraz liofilizację (ChristfreezedryerAlpha 1–2LD, Martin Christ, Osterodeam Harz, Germany). Otrzymaną pozostałość rozpuszczono w wodzie i poddano analizie chromatograficznej. Oznaczenia ilościowe 14 modyfikowanych i niemodyfikowanych nukleozydów (10 mM) (lista A) zostały przeprowadzone przy długości fali 254 nm i przy użyciu kolumny chromatograficznej Gemini C18™ (250 × 4,6 mm, 5 μm; Phenomenex,CA, USA). Analizy przeprowadzono w trybie elucji gradientowej przy składzie fazy ruchomej: A – metanol, B – 0,1% wodny roztwór kwasu mrówkowego o pH 2,8 (od 2 % A do 20% A). Czas analizy wynosił 50 minut. Dokładny opis parametrów zastosowanej metody analitycznej znajduje się w publikacji Waszczuk-Jankowska i wsp. [48].

Niecelowana analiza metabolomiczna próbek moczu przedłożona w publikacji nr 3 została przeprowadzona metodą wysokosprawnej chromatografii cieczowej sprzężonej ze spektrometrem mas z analizatorem czasu przelotu i jonizacją metodą elektrorozpylania (HPLC-ESI-TOF-MS, zestaw 1200 HPLC oraz 6224 TOF-MS, Agilent Technologies, Waldbronn, Germany) przy użyciu kolumny chromatograficznej Ascentis Express C18 (150 x 4,6 mm, 2,7 μm; Supelco Analytical, USA). Analizy chromatograficzne przeprowadzono z zastosowaniem fazy ruchomej o następującym składzie: 0,1% roztwór kwasu mrówkowego w metanolu (A) i 0,1% roztwór kwasu mrówkowego w wodzie (B). Zastosowano następujący program elucji gradientowej: 5% A od 0 do 1 min, 5-65% A od 1 do 9 min i 65-99% A od 9 do 17 min, 99 % A od 17 do 26 min, a następnie 12-minutowy czas ekwilibracji. Szybkość przepływu fazy ruchomej wynosiła 0,5 ml/min, a całkowity czas analizy 26 min. Próbki moczu analizowano w zrandomizowanej kolejności. Dane analityczne zbierano w trybie przemiatania (*scan mode)* w zakresie 100-1100 *m/z* w trybie jonizacji dodatniej i ujemnej. Przetwarzanie uzyskanych danych przeprowadzono przy użyciu algorytmu MFE (*Molecular Feature Extractor*) w programie MassHunter Qualitative Analysis (Agilent Technologies, Waldbronn, Germany). Wyrównanie sygnałów analitycznych przeprowadzono przy użyciu

programu Mass Profiler Professional B.02.01 (Agilent Technologies, Waldbronn, Germany). Dane zostały w następnej kolejności poddane procedurze filtracji w oparciu o kryteria zapewnienia jakości, tj. obecność danego sygnału w przynajmniej 50% próbek kontrolnych i współczynniku zmienności (*coefficient of variation*, CV) < 20%, a także obecności tych sygnałów w 95% próbek w co najmniej jednej z porównywanych grup. W wyniku niecelowanej analizy metabolomicznej otrzymano 603 sygnały w jonizacji dodatniej i 361 w ujemnej. Każda zmienna opisana była za pomocą masy monoizotopowej, czasu retencji i intensywności sygnału analitycznego.

2. Metodologia obliczeniowa

2.1 Modelowanie retencji chromatograficznej nukleozydów i związków pterynowych

Publikacja nr 1 pt. "*Least absolute shrinkage and selection operator and dimensionality reduction techniques in quantitative structure retention relationship modeling of retention in hydrophilic interaction liquid chromatography*" opisuje zastosowanie takich metod analizy danych jak (i) metoda cząstkowych najmniejszych kwadratów (PLS, *Partial Least Squares*), (ii) metoda regularyzacji LASSO oraz (iii) LASSO w połączeniu z regresją krokową, do analizy zależności struktura-retencja w układzie HILIC (*hydrophilic interaction liquid chromatography*) dla nukleozydów oraz związków pterynowych.

Wymienione związki, zaproponowane jako potencjalne wskaźniki stanów patofizjologicznych, mają charakter polarny, są więc trudne do analizy w klasycznym układzie RP-HPLC (*reverse-phase high-performance liquid chromatography*). W pracy analizowano zależności struktura-retencja wykorzystując 3 fazy stacjonarne typu HILIC. Dwie fazy stacjonarne N,O-diaminofosfoamidowe (*N,O-dialkylphosphoramidate*) tj. APC-10 i APC-18 zostały zsyntetyzowane w laboratorium Katedry Chemii Środowiska i Bioanalityki UMK [49,50], natomiast trzecią stanowiła komercyjnie dostępna faza z immobilizowaną sztuczną membraną (*Immobilized Artificial Membrane*, IAM.PC.DD2). Wszystkie 3 fazy stacjonarne wykazywały niską hydrofobowość. Do numerycznego opisu właściwości 27 badanych analitów wykorzystano 1304 deskryptory wygenerowane w programie HyperChem i Dragon.

Celem pracy było porównanie zdolności predykcyjnych modeli QSRR zbudowanych przy użyciu metod tj. PLS, LASSO oraz LASSO w połączeniu z regresją krokową, wraz z wyznaczeniem modeli QSRR, dla nukleozydów i pteryn.

Postać modelu

W pracy wykorzystano trzy rodzaje metod regresji do wyznaczenia zależności pomiędzy deskryptorami a eksperymentalnie wyznaczonymi współczynnikami retencji.

Metoda PLS stanowi nadzorowaną metodę redukującą wielkowymiarowość danych, która pozwala identyfikować zbiór zmiennych będących liniową kombinacją oryginalnych zmiennych.

Regresja LASSO [51] to technika redukcji wielowymiarowej matrycy danych oraz selekcji zmiennych objaśniających. Metoda zapewniła redukcję wariancji współczynników modelu kosztem ich większego obciążenia w wyniku czego zmniejszył się błąd średniokwadratowy. Funkcja straty (*loss function*) w metodzie LASSO wygląda tak jak w metodzie najmniejszych kwadratów, aczkolwiek uwzględnia tzw. człon kary $\sum_{j=1}^{p}\left|\beta_j^{'}\right| \le \lambda$. Regresja LASSO estymuje wartości współczynników $\beta_j^{'}$ poprzez minimalizację funkcji straty, $LASSO = \left(\sum_{i=1}^{n} y_i^{'} - \beta_0^{'} - \sum_{j=1}^{p}\beta_j^{'} x_{ij}^{'}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j^{'}\right|$. Redukcja wariancji kosztem większego obciążenia współczynników poprawia dokładność modelu (*model accuracy*) zwłaszcza w przypadku danych wielowymiarowych, gdy występuje więcej zmiennych niż obserwacji. Proces selekcji zmiennych jak też estymacja parametrów zachodzi jednocześnie. W modelu zostają jedynie te współczynniki, które są w największym stopniu związane ze zmienną objaśnianą.

Zastosowanie metody LASSO w połączeniu z regresją krokową pozwoliło na eliminację problemu współliniowości zmiennych oraz pozwoliło na obliczenie błędu standardowego współczynników, których z założenia nie oblicza się przy zastosowaniu jedynie metody LASSO ze względu na ich duże obciążenie.

Diagnostyka i dopasowanie modelu

Zdolność predykcyjną modeli sprawdzono za pomocą walidacji krzyżowej, tj. walidacji *k*-krotnej walidacji „wyrzuć jeden obiekt" (*leave-one-out crossvalidation*), walidacji „wyrzuć wiele obiektów" (*leave-many-out crossvalidation*) oraz walidacji zewnętrznej (*external validation*). Aby oszacować zdolność predykcyjną modeli (*model performance*) wykorzystano miary takie jak Q^2 (*Predictive Squared Correlation Coefficient*) i RMSE (*Root Mean Squared Error*). W celu oceny współliniowości zmiennych objaśniających wykorzystano miarę VIF (*Variance Inflation Factor*).

Wyniki i wnioski

Wyniki analizy PLS pokazały, że model dla fazy APC-10 charakteryzował się największą zdolnością predykcyjną (Q^2_{LOOCV} = 0,687), natomiast niższą zaobserwowano dla fazy stacjonarnej IAM.PC.DD2 (Q^2_{LOOCV} = 0,503). W przypadku fazy stacjonarnej APC-18 obserwowano negatywną wartość Q^2, co sugeruje brak zdolności predykcyjnych modelu opisującego retencję przy użyciu tej fazy.

Dla metody LASSO wyselekcjonowano 5, 7 i 4 deskryptory odpowiednio dla APC-10, APC-18 i IAM.PC.DD2, dla których wartości R^2 i Q^2 wskazywały na dobre dopasowanie modelu do danych (R^2> 0,81) i brak przeszacowania modelu (różnica między R^2 i Q^2 nie była większa niż 0,3). W przypadku fazy stacjonarnej APC-18 nie obserwowano negatywnej wartości Q^2 tj. w modelu PLS, co może być spowodowane obecnością wartości wpływowych (*influential points*), czyli takich które mają duży wpływ na współczynnik kierunkowy prostej regresji.

Zastosowanie LASSO w połączeniu z regresją krokową miało na celu ograniczenie liczby deskryptorów w modelu wyselekcjonowanych jedynie przy pomocy techniki LASSO oraz zmniejszenie prawdopodobieństwa korelacji miedzy nimi z powodu wprowadzenia dużej ilości zmiennych do modelu. Najlepsze zdolności predykcyjne modelu opisującego retencję otrzymano dla fazy APC-10 (Q^2_{ext} = 0,866; $RMSE^2_{ext}$ = 0,197).

Na podstawie wygenerowanych równań QSRR, można przypuszczać, że mechanizm retencji nukleozydów i związków pterynowych badany przy użyciu 3 faz stacjonarnych opiera się na (i) mechanizmie podziału między fazą ruchomą a warstwą wodną zebraną blisko warstwy krzemionkowej oraz (ii) oddziaływaniach polarnych między analitami a powierzchnią fazy stacjonarnej. Związek miedzy procentową zawartością modyfikatora organicznego a współczynnikiem retencji może być opisany funkcją kwadratową. Niemniej jednak, zaznaczyć należy, że dla otrzymanych równań QSRR mierzono retencję jedynie dla 3-4 punktów. Metoda LASSO efektywnie redukuje wielowymiarowość danych, zmniejsza ryzyko przeszacowania modelu w porównaniu do metody PLS. Jednakże, w przypadku danych wielowymiarowych należy zaznaczyć, że promuje rozwiązania rzadkie, czyli takie dla których

współczynniki modelu są bliskie zeru (prawdziwa wartość współczynników jest obniżana).

Otrzymane równania QSRR nie powinny być jednak generalizowane na inne polarne anality. Jedynie osobne modele zbudowane dla nukleozydów i pteryn mogą upoważnić do generalizacji otrzymanych równań przy użyciu zewnętrznej walidacji.

2.2 Modelowanie danych metabolomicznych z badań obserwacyjnych

Publikacja nr 2 pt. "*Multilevel pharmacokinetics-driven modeling of metabolomics data*" przedstawia koncepcję modelowania hierarchicznego danych metabolomicznych znanego w farmakokinetyce pod nazwą modelowania efektów mieszanych (*mixed effects modeling*) dla danych zebranych w punktach czasowych. Celem populacyjnego modelowania farmakokinetycznego jest charakterystyka profilu stężenie-czas w celu zrozumienia procesów towarzyszących absorpcji i dystrybucji leku w organizmie [52,53]. Podejście populacyjne do modelowania farmakokinetyki leku opisuje związek między parametrami farmakokinetycznymi a indywidualnymi cechami pacjenta (tj. wiek, płeć, masa ciała, funkcja nerek), które w modelu traktowane są jako efekty stałe oraz uwzględnia zmienność parametrów, które pełnią rolę efektów losowych [54,55].

Z punktu widzenia farmakokinetyki oraz metabolomiki, klirens kreatyniny jest kluczową zmienną odpowiednio do (i) ustalenia dawkowania leków wydalanych przez mocz oraz (ii) normalizacji stężeń metabolitów wydalanych do moczu [56]. Klirens kreatyniny może być estymowany m.in. poprzez równanie MDRD (*Modification of Diet in Renal Disease Study*) uwzględniając takie kowarianty jak wiek, płeć, osoczowe stężenie kreatyniny. Stąd, można się spodziewać, że różnice międzyosobnicze będą miały wpływ na zmienność produkcji kreatyniny.

Z perspektywy modelowania stężeń substancji w organizmie, zmierzone stężenie jest definiowane jako ilość substancji w jednostce objętości oraz (jak każdy pomiar) jest mierzone z błędem. Stąd obserwowane stężenia metabolitów wydalanych z moczem i kreatyniny można zdefiniować jako szybkość ich produkcji do szybkości diurezy (jednakowa dla metabolitów i kreatyniny). Dzięki normalizacji stężeń metabolitów, można zaobserwować niezależny od diurezy stosunek szybkości wydalania metabolitów do kreatyniny. Obydwie wielkości charakteryzują się typową wartością w populacji z którą związana jest określona zmienność międzyosobnicza. Co więcej, stosunek szybkości wydalania metabolitów do kreatyniny ma również swoją wartość typową, której towarzyszy zmienność międzyosobnicza oraz błąd pomiarowy. Stąd też, na normalizowane stężenia metabolitów w moczu wpływają takie aspekty fizjologii jak szybkość wydalania metabolitów i kreatyniny, a reszty

modelu zależą od błędu pomiarowego i zmienności międzyosobniczej zarówno metabolitów jaki i kreatyniny.

Farmakokinetyczne aspekty modelowania danych mają zatem ogromny wpływ na obserwacyjne badania metabolomiczne, których domeną jest zmienność (międzyosobnicza, wewnątrzosobnicza) oraz związana z nią niepewność (*uncertainty*) wokół estymowanych parametrów.

Celem pracy było więc zaproponowanie modelu do modelowania niezbalansowanych i niedobranych pod względem płci i wieku danych pochodzących z celowanej analizy metabolomicznej 13 nukleozydów w obrębie 248 osób z nowotworami układu moczowo-płciowego (153 pacjentów i 95 zdrowych ochotników) oraz próba odpowiedzi na pytanie czy kowarianty tj. wiek, płeć, stan zdrowia (zdrowy/chory) wpływają na stosunek szybkości wydalania nukleozydów do kreatyniny.

Postać modelu

Zakładając stan stacjonarny, zaproponowany bayesowski model hierarchiczny uwzględniał efekt wywierany przez wiek, płeć, stan zdrowia (zdrowy/chory) na stosunek szybkości wydalania nukleozydów do kreatyniny. Na parametr opisujący zmienną „stan zdrowia" nałożono mało informatywny rozkład *a priori*, tj. rozkład normalny ze średnią zero i odchyleniem standardowym z rozkładu jednostajnego. Natomiast na parametr opisujący zmienną „wiek" i „płeć" nałożono informatywne rozkłady *a priori*, tj. rozkład normalny ze średnią 0,203 i odchyleniem standardowym z rozkładu jednostajnego oraz rozkład normalny ze średnią 0,293 i odchyleniem standardowym z rozkładu jednostajnego (obie wartości średnich pochodziły z równania MDRD). Punkt przecięcia równania (*intercept*) był modelowany przy użyciu rozkładu normalnego ze średnią zero i dużą wariancją. Zmienność międzyosobniczą dla rozkładu zmiennej losowej pochodzącej z wielowymiarowego rozkładu normalnego (*Multivariate Normal Distribution*) modelowano przy pomocy rozkładu Wishart'a ze sferyczną matrycą kowariancji $\Lambda_0 = 0{,}05\, I_{13}$ (*I – identity matrix*) i liczbą stopni swobody równą liczbie zmiennych ($\rho = 13$).

Diagnostyka łańcuchów i dopasowanie modelu

Zaproponowany model został sparametryzowany w programie JAGS 4.0.0 w środowisku programistycznym języka R z wykorzystaniem oprogramowania RStudio. Zastosowano 3 łańcuchy MCMC, dla których liczba użytych iteracji wynosiła 6000. Pierwsze 3000 iteracji z każdego łańcucha zostało pominięte, a następnie, co trzecia iteracja została zachowana. Dla wszystkich parametrów zbudowano wykres typu *traceplot* w celu oceny czy stan stacjonarny został osiągnięty. Zbieżność łańcuchów oceniono metodą Brooksa-Gelmana-Rubina ($R < 1,2$) oraz wizualnie za pomocą wykresu obrazującego kolejne stany łańcucha Markova, które powinny w stanie stacjonarnym przypominać "włochatą gąsienicę" (*fuzzy caterpillar)*. Dopasowanie modelu oceniono wizualnie przy pomocy VPC oraz za pomocą wykresu dla reszt. Porównania dokładności modelu z/bez zmiennej „stan zdrowia", w celu oceny wpływu tej zmiennej na stosunek szybkości wydalania nukleozydów do kreatyniny, dokonano wykorzystując miarę DIC. Zbiór walidujący służył do oceny zdolności predykcyjnych modelu.

Wyniki i wnioski

Zaproponowany model charakteryzował się zadowalającymi właściwościach predykcyjnymi, dla którego wartości eksperymentalne zawierały się w 90% przedziale ufności. Uzyskano lepsze (w porównaniu do modelu zerowego) dopasowanie modelu, gdy zmienna „stan zdrowia" była jedną ze zmiennych w modelu ($\Delta DIC = 31$). Na tej podstawie wnioskowano, że informacja o stanie zdrowia wyjaśnia część zmienności międzyosobniczej w stężeniach nukleozydów. W konsekwencji obecności nowotworu, normalizowane stężenie metylotioadenozyny (MTA) zwiększyło się średnio o 1,42 (1,09–2,03), natomiast dla I o 1,25 (1,04–1,6) przy założeniu zaproponowanego modelu i dostępnych danych. Dla pozostałych nukleozydów średni efekt był bliski zeru. Wiek wpływał na stosunek szybkości wydalania nukleozydów do kreatyniny dla wszystkich nukleozydów w tym samym kierunku, co prawdopodobnie jest konsekwencją obniżenia klirensu kreatyniny z wiekiem. Zaobserwowano różny efekt płci wywierany na stosunek szybkości wydalania nukleozydów do kreatyniny. Dla takich nukleozydów jak pseudourydyna (pseu), 1-metyloadenozyna (1mA), 3-metylocytydyna (3mC), ksantozyna (X), adenozyna (A), N_4-acetylocytydyna (N4Ac), N_2N_2-dimetyloguanozyna (N2N2), 7-metyloguanozyna

(7mG), inozyna (I), guanozyna (G), 5-metylourydyna (5mU), 6-metyloadenozyna (6mA), stężenie było wyższe u kobiet w porównaniu do mężczyzn, co może być efektem wpływu płci na klirens kreatyniny. W oparciu o dystrybucję *a posteriori* opisującą prawdopodobieństwo wystąpienia choroby wnioskowano na temat ograniczonej użyteczności oznaczania 13 nukleozydów do przewidywania stanu zdrowia, AUC = 0,57 (0,5–0,67), co zostało potwierdzone poprzez wizualizację indywidualnego prawdopodobieństwa rozwoju choroby dla wybranych osób ze zbioru walidującego.

Wzrost stężenia MTA, chociaż nieistotny z klinicznego punktu widzenia, zasługuje na uwagę, gdyż doniesienia literaturowe wskazują na jego potencjalną zależność ze stopniem rozwoju nowotworu [57].

Podejście populacyjne zastosowane do modelowania danych metabolomicznych pozwoliło na zbudowanie modelu opisującego stosunek szybkości wydalania nukleozydów do kreatyniny opartego na wybranych aspektach fizjologicznych i zmienności międzyosobniczej. Poza tym, dzięki zastosowaniu podejścia populacyjnego możliwe było przewidywanie indywidualnego prawdopodobieństwa choroby. Z praktycznego punktu widzenia umożliwia ono racjonalne podejmowanie decyzji czy dodatkowa wiedza, dostarczona przez badane związki, dodaje informację do tej założonej *a priori*. Zatem, zastosowanie bayesowskiego modelu hierarchicznego może mieć praktyczne zastosowanie w metabolomice, szczególnie w aspekcie diagnozy i leczenia indywidualnego pacjenta.

2.3 Modelowanie danych metabolomicznych zebranych w punktach czasowych

Publikacja nr 3 pt. "*How to model temporal changes in nontargeted metabolomics study? A Bayesian multilevel perspective.*" przedstawia koncepcję modelowania danych pochodzących z niecelowanej analizy metabolomicznej zebranych w punktach czasowych w badaniu *in vivo*.

Idea przeprowadzonego eksperymentu *in vivo* zakładała wywołanie nowotworu pęcherza moczowego u szczurów poprzez zaszczepienie komórkami nowotworowymi ścian pęcherza moczowego. Badanie przeprowadzono w obrębie grupy badanej i kontrolnej. Po zaszczepieniu, mocz zwierząt pobierano w 4 punktach czasowych: dzień przed szczepieniem, dzień po szczepieniu, a następnie 3 i 8 tygodni po szczepieniu komórkami nowotworowymi.

Aby modelować zmiany intensywności metabolitów „w czasie", zaproponowano metodologię obliczeniową zaczerpniętą z koncepcji populacyjnego modelowania farmakokinetycznego. Jej celem było (i) określenie zmienności w intensywności zmierzonych sygnałów uwzględniając zmienność międzyosobniczą, wewnątrzosobniczą, jak też zmienność resztkową oraz (ii) ocena dynamiki zmian „w czasie" w profilu metabolomicznym na skutek wprowadzenia komórek nowotworowych do pęcherza moczowego.

Postać modelu

Zmienność w intensywnościach 964 zmierzonych sygnałów została opisana bayesowskim modelem hierarchicznym postaci:

$$y_{\mathrm{i}} = \mu_z + \beta_z COV_i + \eta_{z,k}^{MET*RAT} + \eta_{z,l}^{MET*OCC} + \eta_{k,l}^{RAT*OCC} + \varepsilon_i$$

gdzie y_i to logarytm zarejestrowanego sygnału, μ_z to średnia wartość sygnału dla metabolitu, β_z to efekt wywierany przez szczepienie komórkami nowotworowymi (*treatment effect*), natomiast zmienność międzyosobniczą, między okazjami oraz wspólną zmienność dla metabolitów między szczurami a okazjami opisano przez $\eta^{MET*RAT}$, $\eta^{MET*OCC}$, $\eta^{RAT*OCC}$. Na powyższe parametry oraz towarzyszące im super-parametry nałożono odpowiednie rozkłady. Efekt szczepienia komórkami nowotworowymi (β_z) modelowany był przy założeniu (i) braku efektu wywołanego

szczepieniem komórek nowotworowych do pęcherza zwierząt (Model 001), (ii) efektu pochodzącego z rozkładu Laplace'a dla którego parametr skali estymowany był z danych (Model 002) oraz (iii) efektu pochodzącego z rozkładu Laplace'a, dla którego parametr skali był równy jedności (Model 003).

Diagnostyka łańcuchów i dopasowanie modelu

Zaproponowany model został sparametryzowany w programie JAGS 4.0.0 w środowisku programistycznym programu MATLAB 8.4. Zastosowano 3 łańcuchy MCMC, dla których liczba użytych iteracji wynosiła 8550. Pierwsze 7500 iteracji z każdego łańcucha zostało pominięte, a następnie, co trzecia iteracja została zachowana. Zbieżność łańcuchów oceniono metodą Brooksa-Gelmana-Rubina ($R < 1,2$) oraz wizualnie za pomocą wykresu obrazującego kolejne stany łańcucha Markova, które powinny w stanie stacjonarnym przypominać "włochatą gąsienicę" (*fuzzy caterpillar)*. Dopasowanie modelu sprawdzono przy pomocy wykresu GOF oraz VPC. Porównania dokładności modeli przy założeniu różnych rozkładów nałożonych na efekt wywierany przez szczepienie komórkami nowotworowymi dokonano przy użyciu kryterium informacyjnego WAIC (*Wantanabe-Akaike information criteron*), który aproksymuje walidację typu „wyrzuć jeden obiekt" dla krzyżowej walidacji modeli bayesowskich. Miara ta została użyta również do oceny użyteczności składowych modelu.

Wyniki i wnioski

Model 001 oraz Model 002 najlepiej opisywały efekt wywierany przez zaszczepienie komórek nowotworowych do ściany pęcherza zwierząt badanych. W przypadku obu modeli wartości β_z były bliskie zeru co oznaczało brak bądź znikomy wpływ zaszczepienia na zmiany w profilu metabolomicznym szczurów. Zaszczepienie komórek nowotworowych powodowało zmiany (zwiększenie lub zmniejszenie intensywności) w średnich sygnałach metabolitów w zakresie 0,8–1,25. Największą zmienność zaobserwowano między okazjami (% CV = 153%) sugerując, że sygnały pochodzące od niektórych metabolitów mają mniejszą zmienność u danego szczura, a z kolei inne wykazują dużą zmienność w obrębie szczura. Zmienność między szczurami była mniejsza w obrębie metabolitów (% CV = 74%), natomiast zmienność resztkowa (% CV = 71%) wskazywała na względną stałość intensywności sygnałów w czasie. Obserwowane sygnały metabolitów zmieniały się

około ±20% w obrębie wszystkich okazji. 36% metabolitów zawierało obserwacje odstające, stąd zastosowanie rozkładu Laplace'a o „ciężkich końcach" (*heavy tailed*) do modelowania efektu wywieranego przez zaszczepienie komórek nowotworowych, wydawało się być słusznym podejściem. Nie zaobserwowano trendu w czasie (rosnącego lub malejącego) na skutek zaszczepienia komórek nowotworowych, co potwierdzono brakiem wywołania nowotworu u badanych zwierząt w badaniu histopatologicznym.

Zastosowanie bayesowskiego modelu hierarchicznego do modelowania zmian w intensywności metabolitów w czasie pozwoliło uwzględnić hierarchiczną strukturę danych metabolomicznych oraz wszelkie wynikające z niej źródła zmienności. Jednakże, w modelu nie uwzględniono matrycy kowariancji wyjaśniającej korelację międzyosobniczą pomiędzy metabolitami z uwagi na jej zbyt obszerny rozmiar. Brak trendu w czasie dla obserwowanych sygnałów metabolitów mógł być skutkiem zbyt krótkich odstępów czasowych, w których zbierane były dane. Z drugiej strony, biorąc pod uwagę fizjologiczną odpowiedz organizmu na obecność komórek nowotworowych w ścianie pęcherza, rozwój nowotworu pęcherza moczowego w tak krótkim przedziale czasowym jak 8 tygodni mógł nie zostać osiągnięty.

2.4 Modelowanie danych transkryptomicznych z badań obserwacyjnych

Modelowanie danych transkryptomicznych z badań obserwacyjnych jest kolejnym etapem badawczym, nie stanowiącym podstawy rozprawy doktorskiej ale wykorzystującym omówione wcześniej aspekty naukowo-badawcze i kolejno je rozwijające. Badania dotyczą zastosowania modeli hierarchicznych w analizie danych „omicznych" polegające na analizie danych transkryptomicznych pochodzących z profilowania ekspresji mikro RNA (miRNA) w grupie osób zdrowych i chorych na raka jajnika przy użyciu bayesowskiego modelu hierarchicznego. Celem pracy była ocena użyteczności profilowania 49 miRNA do diagnostyki raka jajnika. Otrzymane wyniki zostały przedyskutowane w kontekście zastosowania testu FDR (*False Discovery Rate*). Wyniki badań zostały zebrane w manuskrypcie pt. „*Bayesian multilevel model of micro RNA levels in ovarian-cancer and healthy subjects*" autorstwa Paweł Wiczling, Emilia Daghir-Wojtkowiak, Roman Kaliszan, Michał Jan Markuszewski, Magdalena Ratajska, Magdalena Koczkowska, Maciej Stukan, Alina Kuźniacka, Janusz Limon i przesłane do recenzji.

Profilowanie ekspresji 50 mikro RNA (miRNA) przedłożone w powyższym manuskrypcie zostało przeprowadzone metodą *real-time* PCR przez firmę Exiqon (Qiagen Company, Denmark). W pierwszym etapie, z próbek osocza wyizolowano RNA, które w dalszym etapie zostało przepisane na komplementarne DNA (cDNA). Otrzymane cDNA zostało następnie poddane reakcji PCR w termocyklerze (LightCycler® 480 Real-Time PCR System, Roche) przy użyciu odpowiedniej mieszaniny reakcyjnej (*master mix*) na płytce 384-dołkowej. Dane w postaci krzywych amplifikacji zebrane zostały za pomocą odpowiedniego oprogramowania (Roche LC software). Dla każdego miRNA określono tzw. wartość CT (*thresholdcycle*), czyli numer cyklu przy którym sygnał fluorecencji przecina poziom obserwowanego sygnału (*threshold*) wskazujący na wzrost powyżej linii bazowej (*baseline*) (wartość CT jest odwrotnie proporcjonalna do ilości matrycy w próbce). Jakość wyizolowanego RNA, syntezy cDNA, amplifikacji PCR, jak również obecność inhibitorów, była kontrolowana dzięki dodaniu tzw. *spike-ins* przed etapem izolacji RNA z materiału biologicznego. W przypadku niskiej wydajności reakcji, nieprawidłowego przebiegu krzywej topnienia (*melting curve*) czy oznaczeń kontroli negatywnej, wartość CT dla danej próbki i danego miRNA była usuwana z matrycy danych.

Profilowanie miRNA wykonano w dwóch oddzielnych powtórzeniach. Badania użyteczności miRNA w obserwacyjnym eksperymencie transkryptomicznym obejmowało 2 etapy. W pierwszym etapie, z 752 cząsteczek miRNA profilowanych w obrębie 58 osób (9 pacjentek z mutacją +/+ BRCA1 lub BRCA2; 33 pacjentki z mutacją -/- BRCA1/2; 16 zdrowych ochotników), wyselekcjonowano 49 cząsteczek miRNA, których ekspresja w grupie badanej i kontrolnej różniła się istotnie. W drugim etapie badań, wybrane miRNA profilowano w grupie 178 osób (59 pacjentek oraz 119 zdrowych ochotników) w dwóch powtórzeniach otrzymując wartości CT (*quantification threshold*) opisujące numer cyklu przy którym sygnał fluorescencji przecina wartość progową (*threshold*).

Postać modelu

Metodologia obliczeniowa obejmowała następujące etapy:

a) odtworzenie procesu generowania danych – zdefiniowanie zależności deterministycznej między odczytem CT a stężeniem miRNA przy założeniu 40 cykli PCR oraz wydajności reakcji na poziomie 100%.

c) budowa modelu – w celu zbadania jaki wpływ na mierzone wartości miRNA ma obecność choroby, zbudowano model uwzględniający tzw. zmienne zakłócające (*confounding variables*) tj. wiek, masa ciała. Na parametry modelu nałożone zostały odpowiednie rozkłady wraz z uwzględnieniem dostępnych źródeł zmienności. Ocena „*a posteriori*" siły efektu z towarzyszącą niepewnością umożliwiła wnioskowanie na temat potencjalnej użyteczności miRNA.

Diagnostyka łańcuchów i dopasowanie modelu

Zaproponowany model został sparametryzowany w programie JAGS 4.0.0 w środowisku programistycznym języka R z wykorzystaniem oprogramowania RStudio. Zastosowano 3 łańcuchy MCMC, dla których liczba użytych iteracji wynosiła 100000. Pierwsze 1000 iteracji z każdego łańcucha zostało pominięte, a następnie, co trzecia iteracja została zachowana. Dla wszystkich parametrów modelu zbudowano wykres typu *traceplot* w celu oceny czy stan stacjonarny został osiągnięty. Zbieżność łańcuchów oceniono metodą Brooksa-Gelmana-Rubina ($R < 1,2$) oraz wizualnie za pomocą wykresu obrazującego kolejne stany łańcucha Markova, które powinny w stanie stacjonarnym przypominać "włochatą gąsienicę" (*fuzzy caterpillar).*

Dopasowanie modelu oceniono wizualnie przy pomocy wykresu zależności reszt ważonych od wartości przewidzianych.

Wyniki i wnioski

W oparciu o dystrybucję *a posteriori*, istnieją różnice w średnich wartościach ekspresji miRNA. Mikro RNA tj. miR-101-3p, miR-142-5p, miR-148a-3p charakteryzowały się wysokim współczynnikiem korelacji (> 0,95) oraz podobną siłą efektu [0,45 – 1]. Dla powyższych miRNA estymowana wartość AUC znajdowała się w zakresie 0,63 – 0,67 co świadczy o ich ograniczonym potencjale diagnostycznym. Dla większości miRNA obserwowano dużą niepewność wokół estymowanych parametrów, w szczególności miR-604 i miR-221-5p wymagają dalszych badań w celu oceny ich użyteczności diagnostycznej.

W kontekście analizy częstościowej (*Frequentists paradigm*), zaobserwowano dwie cząsteczki miRNA (miR-221-5p i miR-346), których poziomy nie różniły się istotnie w kontekście testowania istotności hipotezy zerowej jednakże wykazywały dużą siłę efektu (> 1). Jest to prawdopodobnie związane z obecnością braków w danych (odpowiednio 94,38% i 95,51%). Wyniki te potwierdzają duże prawdopodobieństwo błędnego wnioskowania w przypadku ignorowania niepewności wokół parametrów czy przewidywań.

Zaproponowany model jest użyteczny do oceny różnic w ekspresji miRNA między pacjentami a zdrowymi ochotnikami. Selekcja potencjalnych wskaźników choroby na podstawie wartości *p* jest związana z dużym prawdopodobieństwem przeszacowania efektów, dużego odsetka wyników fałszywie-pozytywnych wynikających z charakteru eksperymentu (*study design*), zbyt mało licznej próby czy obecności zmiennych zakłócających. Prowadzi to do ograniczonej zdolności predykcyjnej stosowanego modelu oraz braku powtarzalności wyników. Z tego względu zasadne wydaje się być zastosowanie bardziej dopasowanych metod analizy danych tj. modele hierarchiczne. Zastosowany model hierarchiczny może być użyty to każdego typu danych „omicznych" zarówno w kontekście eksploracji danych jak również do weryfikacji hipotez.

IV. DYSKUSJA

Zastosowanie regresji LASSO oraz bayesowskich modeli hierarchicznych, zaprezentowane na przykładzie analiz chromatograficznych oraz badań „omicznych", dotyczyło analizy (i) danych wielowymiarowych, (ii) danych zebranych z badań obserwacyjnych oraz (iii) danych zebranych w punktach czasowych.

Głównym celem prac było zaproponowanie strategii modelowania (i) wielowymiarowej macierzy danych z wykorzystaniem techniki regularyzacji, oraz (ii) hierarchicznej struktury danych pochodzącej z analizy celowanej (celowana analiza metabolomiczna, profilowanie ekspresji miRNA) oraz niecelowanej (niecelowana analiza metabolomiczna) z wykorzystaniem wnioskowania bayesowskiego. Zaproponowane strategie obliczeniowe umożliwiły (i) otrzymanie optymalnych modeli struktura-retencja QSRR, (ii) uwzględnienie zmienności międzyosobniczej, zmienności między okazjami oraz błędu związanego z pomiarem, (iii) uwzględnienie informacji *a priori* w modelu, (iv) estymację niepewności wokół parametrów i przewidywań [27].

Regularyzacja jest użyteczną techniką analizy danych wielowymiarowych oraz znajduje zastosowanie w procesach odtwarzania sygnałów. Metoda ta pozwala na oddzielenie sygnału od szumu informacyjnego poprzez zapewnienie regularności rozkładu estymatorów, co stanowi jej główną zaletę. Wadą regularyzacji jest natomiast złożoność obliczeniowa oraz obciążenie estymatorów (poprzez nałożenie kary) [59]. W kontekście analizy danych wielowymiarowych, metoda regularyzacji zabezpiecza przed zbyt dobrym dopasowaniem modelu do danych poprzez regularyzowanie wartości o dużej wariancji. W konsekwencji maleje szansa na znalezienie pozornych zależności w danych a modele zyskują zdolność do generalizacji [60].

Regularyzacja ma na celu „ściąganie" (*shrinking*) estymowanych parametrów w kierunku zera co prowadzi do ograniczenia wariancji współczynników poprawiając zdolności predykcyjne modelu. Ma to ogromne znaczenie zarówno w przypadku małej ilości obserwacji/dużej ilości zmiennych, jak też wtedy gdy poszukujemy różnic między grupami w wielowymiarowej przestrzeni danych [64].

Z kolei wielopoziomowa struktura danych typowa dla badań obserwacyjnych oraz badań ze zmienną „czas" (badania z powtarzalnym pomiarem, badania podłużne)

wymaga zastosowania takiej metodologii obliczeniowej, która uwzględnia zastosowany schemat doświadczenia (kryteria włączania i wykluczania) pod kątem różnych poziomów hierarchii danych. Modelowanie takiej struktury danych możliwe jest dzięki tzw. częściowemu uwspólnianiu (*partial pooling)* estymowanych parametrów. Podejście takie opiera się na ściąganiu estymowanych parametrów w kierunku średniej populacyjnej z siłą określoną przez zmienność (szum) obecną w danych. Stąd, częściowe uwspólnianie prowadzi do redukcji wyników fałszywie pozytywnych zapewniając bardziej precyzyjną estymację parametrów dzięki tzw. dzieleniu informacji (*information sharing*) między zmiennymi, zapewniając jednoczesną estymację indywidualnych parametrów modelu [61].

Uwspólnianie parametrów jest charakterystyczne dla bayesowskich modeli hierarchicznych i zapewnia równowagę pomiędzy strategią opartą na braku uwspólniania (*no pooling*) a całkowitym uwspólnianiem (*complete pooling*) parametrów. Brak uwspólniania prowadzi do obciążenia estymowanych parametrów na skutek ignorowania wariancji, natomiast całkowite uwspólnianie skutkuje dużą wariancją. Można zatem powiedzieć, że częściowe uwspólnianie to pewnego rodzaju kompromis między ignorowaniem struktury grupowej w danych a dopasowaniem osobnego modelu dla każdej z grup [62]. Gdy dane zawierają dużo informacji, uwspólnianie estymowanych parametrów objawia się w mniejszym stopniu, oraz odwrotnie, jeśli dane zawierają mało informacji, estymowane parametry w większym stopniu ulegają uwspólnianiu [61]. Stąd częściowe uwspólnianie jest specyficznym kompromisem między wielkością obciążenia a wariancją parametrów [63]. Odpowiednia parametryzacja modelu hierarchicznego prowadzi do otrzymania rozkładu *a posteriori* parametrów modelu, które są zgodne z danymi oraz założonym modelem.

Podejście bayesowskie do modelowania danych oparte o prawdopodobieństwo warunkowe, różni się od klasycznego podejścia do modelowania danych, gdyż traktuje parametry modelu jako zmienne losowe opisywane rozkładem prawdopodobieństwa, dzięki czemu możliwy jest ich opis z zastosowaniem wielu różnych typów rozkładów (bez ograniczenia tylko do rozkładu normalnego). Z tego względu, bayesowska estymacja parametrów jest dużo bardziej odporna na wartości odstające. Rozkład *a posteriori* należy rozumieć jako wypadkową między wiedzą na temat parametrów modelu przed zebraniem danych (wiedza *a priori*) a informacją na temat parametrów

zawartą w zaobserwowanych danych, reprezentowaną przez funkcję wiarygodności [36]. Dodatkowo, niepewność wokół parametrów jest opisywana zarówno na podstawie dowodów eksperymentalnych, jak i subiektywnej opinii badacza i determinuje precyzję przewidywań uzyskanych na podstawie opracowanych modeli. Z drugiej strony, częstościowe podejście do estymacji parametrów posiada szereg wad, które stanowią problem podczas analizy niezbalansowanych danych z badań obserwacyjnych. Należą do nich: duża wrażliwość na wartości odstające, szacowanie parametrów tylko na podstawie danych eksperymentalnych bez uwzględnienia dotychczasowej wiedzy literaturowej czy sztywne założenie rozkładu normalnego dla parametrów modelu. Problematyczne jest także uwzględnienie w modelu dostępnych źródeł zmienności związanych z zaplanowanym eksperymentem, jak też zmienności wynikającej z łączenia różnych badań.

Obecnie w badaniach „omicznych" niejednokrotnie nie ma możliwości przeprowadzenia analizy dużej liczby próbek w jednej sekwencji. Dlatego też, zastosowanie metod obliczeniowych opartych na symulacjach z wielowymiarowych rozkładów *a posteriori* z wykorzystaniem metody Monte Carlo opartej na łańcuchach Markowa wydaje się być szczególnie atrakcyjne. Jednakże, zwłaszcza w analizie metabolomicznej, do wnioskowania statystycznego używa się metod typu *black box* opartych na analizie częstościowej, która oferuje gotowe algorytmy obliczeniowe szeroko stosowane do wielu matryc danych. Poważnym ograniczeniem zastosowania gotowych pakietów jest to, że nie są one dedykowane konkretnym danym, nie pozwalają na dokładne odtworzenie danych oraz często generują fałszywie pozytywnie wyniki. Konkurencyjnym narzędziem dla metod typu *black box* może być modelowanie hierarchiczne. Wykorzystuje ono zjawisko losowości, co oznacza, że implementacja danego algorytmu korzysta z generatora liczb losowych i służy do wyliczenia najbardziej prawdopodobnych parametrów dla konkretnego zestawu danych. Zastosowanie modelowania hierarchicznego, jak i regularyzacji w badaniach „omicznych" i do analizy danych wielowymiarowych eliminuje problem wielokrotnych porównań, pozwala modelować wartości odstające oraz ogranicza prawdopodobieństwo generowania fałszywie pozytywnych wyników.

Aby zweryfikować przewagę modelowania hierarchicznego nad metodami częstościowymi tj. metody oparte na PLS czy testowanie istotności hipotez szeroko stosowanymi do analizy danych „omicznych", należałoby przeprowadzić analizę

porównawczą otrzymanych wyników z wynikami uzyskanymi klasycznymi metodami częstościowymi. Pewnym ograniczeniem przeprowadzonych badań w przedłożonej rozprawie doktorskiej jest brak ww. porównania w odniesieniu do wszystkich prac wchodzących w skład rozprawy doktorskiej. O ile idea, przedstawione założenia oraz zalety bayesowskich modeli hierarchicznych mogą być przekonywujące, „dobra praktyka" modelowania danych wymaga porównania ich do metod powszechnie stosowanych. W dalszych badaniach będących kontynuacją pracy doktorskiej ale nie wchodzących w skład rozprawy doktorskiej („*Bayesian multilevel model of micro RNA levels in ovarian-cancer and healthy subjects*") zaproponowano model umożliwiający odtworzenie procesu generowania danych.

Jednakże brak odniesienia uzyskanych wyników badań do metod opartych na podejściu częstościowym nie stanowi ograniczenia zastosowania bayesowskich modeli hierarchicznych w takich dyscyplinach naukowych jak m.in. farmakokinetyka [65,66], chemia analityczna [67] czy transkryptomika [68]. Zastosowanie modelowania probabilistycznego w metabolomice oraz transkryptomice wydaje się być szczególnie atrakcyjne w świetle fizjologicznych czynników wpływających na mierzoną wielkość, wnioskowania na temat użyteczności potencjalnych wskaźników stanów patofizjologicznych oraz szacowania indywidualnego prawdopodobieństwa choroby w świetle diagnozy i leczenia.

V. WNIOSKI

1. Metoda LASSO w połączeniu z regresją krokową stanowi efektywne narzędzie do selekcji zmiennych z wielowymiarowej przestrzeni danych, tworząc model o dobrych zdolnościach predykcyjnych.

2. Bayesowskie modele hierarchiczne są efektywnym narzędziem do modelowania danych pochodzących z badań obserwacyjnych uwzględniając błąd pomiarowy, dostępne źródła zmienności, informację *a priori*, zapewniając efektywną estymację niepewności wokół parametrów i przewidywań.

3. Bayesowskie modelowanie hierarchiczne stanowi efektywne narzędzie do oceny różnic w poziomach potencjalnych wskaźników patofizjologicznych między pacjentami a zdrowymi ochotnikami dzięki czemu możliwe jest wnioskowanie na temat ich użyteczności w diagnostyce i przewidywaniu choroby.

4. Bayesowskie modelowanie hierarchiczne stanowi efektywne narzędzie do oceny wpływu kowariant na obserwowane stężenia stosunku szybkości wydalania nukleozydów do kreatyniny oraz umożliwiają estymację indywidualnego prawdopodobieństwa choroby, co może mieć praktyczne zastosowanie w diagnostyce i leczeniu indywidualnego pacjenta.

5. Bayesowskie modelowanie hierarchiczne stanowi efektywne narzędzie do modelowania zmian intensywności metabolitów „w czasie” w niecelowanej analizie metabolomicznej, umożliwiając ocenę dynamiki zmian w profilu metabolomicznym setek sygnałów.

VI. BIBLIOGRAFIA

1. Cuperlovic-Culf M., Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling, Metabolites, 2018, 8(1); doi:10.3390/metabo8010004.
2. Rencher AC., Bruce Schaalje G., Linear models in statistics, ISBN 978-0-471-75498-5, Wiley & Sons, Inc., Hoboken, New Jersey.
3. Tikhonov AN., Solution of incorrectly formulated problems and the regularization method, Soviet Mathematical Doklady, 1963, n.4, 1035-1038.
4. Gierasimczyk J., Algorytm spadku gradientu w uczeniu maszynowym, http://www.mif.pg.gda.pl/homepages/kdz/BIGDATA/machine%20learning.pdf
5. Kawalec A., Wajszczyk B., Biuletyn Wat, Vol.LV, Nr. 1, 2006.
6. Kubus M., Propozycja modyfikacji metody złagodzonego LASSO, Taksonomia 22 ISSN 1899-3192, Klasyfikacja i analiza danych – teoria i zastosowania.
7. Krężołek D., Model regresji grzbietowej i jego wykorzystanie do oceny ryzyka inwestycyjnego – przypadek rynku metali, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach ISSN 2083-8611Nr 288, 2016.
8. Conway D., White J.M., Uczenie maszynowe dla programistów, ISBN: 978-83-246-9816-5, 2015 Helion S.A. 2012.
9. Cole SR., Chu H., Greenland S., Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer, American Journal of Epidemiology, 2014, 179(2); 252–260.
10. Hutchinson RA., Valente JJ., Emerson SC., Betts MG., Dietterich TG., Penalized likelihood methods improve parameter estimates in occupancy models, Methods in Ecology and Evolution, 2015, 6, 949–959.
11. Bhattacharya A., Pati D., Pillai N., Dunson DB., Dirichlet-Laplace priors for optimal shrinkage, https://arxiv.org/pdf/1401.5398.pdf
12. De Mol C., De Vito E., Rosasco L., Elastic-net regularization in learning theory, Journal of Complexity, 2009; 25(2); 201-230.
13. Brown H., Prescott R. (1999) Applied mixed models in medicine (Wiley, Chichester, UK)

14. Frątczak E., Analiza danych wzdłużnych - wybrane zagadnienia, "Statystyka - zastosowania biznesowe i społeczne", Warszawa, Wydawnictwo Wyższej Szkoły Menedżerskiej w Warszawie, 2014, s. 81-133.
15. Steenbergen MR., Jones BS., Modeling Multilevel Data Structures American Journal of Political Science, 2002; 46(1); 218-237.
16. Goldstein H., Multilevel Statistical Models, ISBN:9780470748657, John Wiley & Sons, Ltd. Wiley Series in Probability and Statistics 2010.
17. West BT., Welch KB., Galecki AT., Linear Mixed Models: A Practical Guide Using Statistical Software, Chapman & Hall/CRC Taylor & Francis Group 2007.
18. Raudenbush SW., Bryk AS., Hierarchical Linear Models: Applications and Data Analysis Methods(Advanced Quantitative Techniques in the Social Sciences) 2nd Edition. Sage Publications 2002, ISBN: 978-0761919049.
19. Raudenbush SW., A Crossed Random Effects Model for Unbalanced Data With Applications in Cross-Sectional and Longitudinal Research Journal of Educational and Behavioral Statistics, 1993, 18(4); 321-349.
20. Galbraith S., Daniel JA., Vissel B., A Study of Clustered Data and Approaches to Its Analysis The Journal of Neuroscience, 2010; 30(32);10601–10608.
21. Searle SR., Gruber MHJ., Linear Models 2nd Edition, 2016, ISBN: 978-1-118-95283-2.
22. Radkiewicz P., Zieliński MW., Hierarchiczne modele liniowe. Co nam dają i kiedy warto je stosować. Psychologia Społeczna, 2010 tom 5 2–3 (14) 217–233, ISSN 1896-1800.
23. Murphy V., Dunne A., Mixed Effects versus Fixed Effects Modelling of Binary Data with Inter-subject Variability Journal of Pharmacokinetics and Pharmacodynamics, 2005, 32(2).
24. Littell RC., Milliken GA., Stroup WW., Wolfinger RD., Schabenberger O., SAS for Mixed Models, Second Edition, A Review of: "SAS Institute Inc., Cary, NC, ISBN 1-59047-500-3, 2006.
25. Szymecki B., Uogólnione nieliniowe modele mieszane z wykorzystaniem języka SAS, http://www.mif.pg.gda.pl/homepages/kdz/StatystykaII/B.Szymecki.pdf
26. Littell RC., Pendergast J., Natarajan R. Modelling covariance structure in the analysis of repeated measures data. Stat Med. 2000,15;19(13):1793-819.

27. McElreath R., Statistical Rethinking, A Bayesian Course with Examples in R and Stan, ISBN: 978-1482253443, Chapman and Hall/CRC, 2015.
28. Dansirikul C., Morris RG., Tett SE., Duffull SB., A Bayesian approach for population pharmacokinetic modelling of sirolimus, Br. J Clin. Pharmacol, 2006; 62(4): 420–434.
29. Bonate PL., Howard DR., Pharmacokinetics in Drug Development: Advances and Applications Vol 3, ISBN 978-1-4419-7936-0, 2011.
30. Congon PD., Applied Bayesian hierarchical methods. Boca Raton: ISBN 9781584887201, Chapman & Hall–CRC Press, 2010.
31. Kruschke J.K., (2011) Doing Bayesian data analysis. ISBN: 9780124058880, Academic Press 2015.
32. Gelman A., Carlin, JB., Stern, HS., Rubin, DB. 2004, Bayesian data analysis, ISBN 9781439840955, Chapman & Hall–CRC Press 2013.
33. Spiegelhalter D.J., Abrams K.R., Myles J.P. 2004, Bayesian Approaches to clinical trial and health-care evaluation. ISBN: 9780471499756, Chichester: Wiley 2004.
34. Gelman A., Prior distributions for variance parameters in hierarchical models. Bayesian analysis, 2006, 1(3), 515-533.
35. Gelman A., Krantz D.H., Lin C., Price P.N., Analysis of Local Decisions Using Hierarchical Modeling, Applied to Home Radon Measurement and Remediation, Statist. Sci. 1999, 14(3); 305-337.
36. Glickman M.E., van Dyk D.A., Basic Bayesian Methods, Methods in Molecular Biology, Vol. 404: Topics in Biostatistics Edited by: W. T. Ambrosius Humana Press Inc., Totowa, NJ.
37. Ntzoufras I., Bayesian Modeling Using WinBUGS Wiley Series in Computational Statistics, ISBN: 978-0-470-14114-4, 2009.
38. Robert C.P., Casella G., Introducing Monte Carlo Methods with R, Springer-Verlag New York, ISBN 978-1-4419-1575-7, 2010.
39. Kruschke J., Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, ISBN: 0124059163, Academic Press, 2014.
40. Gelman A., Hill J., Data Analysis Using Regression and Multilevel/Hierarchical Models 1st Edition, ISBN-13: 978-05216868912009, 2006.

41. Bergstrand M., Hooker A., Wallin J., Karlsson M., Prediction-Corrected Visual Predictive Checks for Diagnosing Nonlinear Mixed-Effects Models.Aaps Journal, 2011, 13(2): 143-151.
42. Wołodźko T., Żółtak M., Wprowadzenie do symulacyjnej estymacji Bayesowskiej, https://dokupdf.com/download/wprowadzenie-do-symulacyjnej-estymacji-bayesowskiej-_5a01b2b4d64ab2b9bd65e8c2_pdf
43. Gelman A., Understanding posterior p-values. Electronic Journal of Statistics ISSN: 1935-7524.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.310.145&rep=rep1&type=pdf
44. Spiegelhalter D.J., Best N.G., Carlin B.P., van der Linde A., Bayesian measures of model complexity and fit. J. R. Statist. Soc. B, 2002, 64(4), 583–639.
45. Plummer M. Penalized loss functions for Bayesian model comparison. Biostatistics, 2008, 9(3), 523-539.
46. Stan Development Team. 2017. The Stan Core Library, Version 2.16.0. http://mc-stan.org
47. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
48. Waszczuk-Jankowska M, Markuszewski MJ, Markuszewski M, Kaliszan R. Comparison of RP-HPLC columns used for determination of nucleoside metabolic patterns in urine of cancer patients. Bioanalysis, 2012; 4(10),1185-94.
49. Bocian S., Nowaczyk A., Buszewski B., A new alkyl-phosphatestationary bonded phases for liquid chromatographic separation of biologically active compounds, Anal. Bioanal. Chem. 404 (2012) 731–740.
50. Bocian S., Paca M., Buszewski B., Characterization of new N.O-dialkylphosphoramidate-bonded stationary phases for reversed-phase HPLC – retention and selectivity, Analyst 128 (2013) 5221–5229.
51. Kalivas JH., Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, J. Chemometrics, 26 (2012) 218-230.

52. Liu, X., Wong, H., Scearce-Levie, K., Watts R.J., Coraggio, M., Shin, J.G., et al. Mechanistic Pharmacokinetic-Pharmacodynamic Modeling of BACE1 Inhibition in Monkeys: Development of a Predictive Model for Amyloid Precursor Protein Processing. Drug Metabolism and Disposition, 2013, 41(7), 1319–1328.
53. Bonate, P.L., Howard, D.R. Pharmacokinetics in Drug Development Clinical Study Design and Analysis (Vol. 1): American Association of Pharmaceutical Scientists, 2005.
54. Gelman A., Bois F., Jiang J. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. Journal of American Statistical Association, 1996, 91, 1400–1412.
55. Gunasekara I., Richardson F., Carter K., Blakely K., Fixed effects analysis of repeated measures data. International Journal of Epidemiology, 2014, 43, 264–269.
56. Lindon JC., Nicholson JN., Holmes E., The handbook of metabonomics and metabolomics. ISBN: 9780444528414, Elsevier, 2007.
57. Kirovski G., Stevens A.P., Czech B., Dettmer K., Weiss T.S., Wild P., et al. Down-Regulation of Methylthioadenosine Phosphorylase (MTAP) Induces Progression of Hepatocellular Carcinoma via Accumulation of 5'-Deoxy-5'-Methylthioadenosine (MTA). The American Journal of Pathology, 2011, 178(3), 1145–1152.
58. Li, M., Marin-Muller, C., Bharadwaj, U., Chow, K.H., Yao, Q., Chen, C. MicroRNAs: Control and Loss of Control in Human Physiology and Disease. World J Surg. 2009, 33, 667–684.
59. Polak AG., Mroczka J., Regularyzacja identyfikacji obiektów złożonych opisanych modelami nieliniowymi, PAK Vol 53, 2007.
60. Ogutu JO., Piepho H.P., Regularized group regression methods for genomic prediction: Bridge. MCP. SCAD. group bridge. group lasso. sparse group lasso. Group MCP and group SCAD, BMC Proc. 2014, 8, S7.
61. Gemlan A. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics. 2006, 48.
62. Lachos VH., Castro, LM., Dey, DK. Bayesian inference in nonlinear mixed-effects models using normal independent distributions. Computational Statistics & Data Analysis, 2013, 64, 237–252.

63. Gelman A., Carlin, J., Stern H., Dunson D., Vehtari, A., Rubin, D. Bayesian Data Analysis, Third Edition, 2013. ISBN 9781439840955, Chapman and Hall/CRC, 2013.
64. Bujak R., Daghir-Wojtkowiak E., Kaliszan R., Markuszewski MJ. (2016). PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-targeted Metabolomics Data. Frontiers in Molecular Bioscience, 3, 35.
65. Basu C., Ahmed MA., Kartha RV., Brundage RC., Raymond GV., Cloyd JC., Carlin BP., A hierarchical Bayesian approach for combining pharmacokinetic/pharmacodynamic modeling and Phase IIa trial design in orphan drugs: Treating adrenoleukodystrophy with Lorenzo's oil. J Biopharm Stat. 2016; 26(6):1025-1039.
66. Wiczling P., Bartkowska-Śniatkowska A., Szerkus O., Siluk D., Rosada-Kurasińska J., Warzybok J., Borsuk A., Kaliszan R., Grześkowiak E., Bienert A. The pharmacokinetics of dexmedetomidine during long-term infusion in critically ill pediatric patients. A Bayesian approach with informative priors. J Pharmacokinet Pharmacodyn. 2016;43(3):315-24.
67. Wiczling P., Kaliszan R., How Much Can We Learn from a Single Chromatographic Experiment? A Bayesian Perspective. Anal Chem. 2016, 5, 997–1002.
68. Spies D., Ciaudo C.. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. Computational and Structural Biotechnology Journal, 2013, 24(13), 469–77.

VII. OŚWIADCZENIA WSPÓŁAUTORÓW