Predykcja zespołu metabolicznego

Jaskuła Z., Kozakiewicz K., Miklaszewska N.

2025-05-31

Spis treści

1 Wstęp

Zespół metaboliczny jest istotnym wyzwaniem dla zdrowia publicznego i epidemiologii. Został po raz pierwszy opisany w 1981 roku, a jego definicja była wielokrotnie aktualizowana [1]. Obecna definicja, zaproponowana w artykule *„Zespół metaboliczny – nowa definicja i postępowanie w praktyce"* [2], opracowanym przez polskie towarzystwa medyczne, brzmi następująco:

***definicja*:** Rozpoznanie zespołu metabolicznego wymaga obecności otyłości (najczęściej brzusznej) oraz co najmniej dwóch z trzech poniższych czynników: podwyższonego ciśnienia tętniczego, nieprawidłowego metabolizmu glukozy, podwyższonego stężenia cholesterolu frakcji nie-HDL (aterogenna dyslipidemia).

Nowa definicja jest uproszczoną wersją wcześniejszych, które obejmowały szerszy zakres kryteriów diagnostycznych, m.in. występowanie cukrzycy typu drugiego, insulinooporności, mikroalbuminurii, czy podwyższonego stężenia trójglicerydów w surowicy [1]. Niezależnie od przyjętej definicji, niezmienne pozostaje stanowisko, że współwystępowanie wymienionych zaburzeń metabolicznych znacząco zwiększa ryzyko rozwoju chorób sercowo-naczyniowych – w stopniu większym, niż każdy z tych czynników rozpatrywany oddzielnie.

Obecnie mówi się o zespole metabolicznym jako o epidemii o zasięgu globalnym [3]. Główne przyczyny jego rosnącej częstości to upowszechnienie siedzącego trybu życia oraz nadmierne spożycie wysokokalorycznych pokarmów, co przyczynia się do wzrostu otyłości – jednego z kluczowych komponentów tego zespołu. W Stanach Zjednoczonych rozpowszechnienie zespołu metabolicznego w populacji dorosłych szacuje się na około 22%, przy czym w grupie pacjentów z cukrzycą typu 2 odsetek ten przekracza 80%. Co istotne, zespół coraz częściej diagnozowany jest również u dzieci i młodzieży, co stanowi poważne wyzwanie dla systemów ochrony zdrowia. Zjawisko to nie ogranicza się jedynie do krajów wysoko rozwiniętych – regiony takie jak Azja, Ameryka Łacińska czy Bliski Wschód odnotowują znaczący wzrost zachorowań, potwierdzając uniwersalny charakter tego zagrożenia zdrowotnego [3].

Z tego względu kluczowe jest zrozumienie mechanizmów leżących u podstaw zespołu metabolicznego oraz doskonalenie metod jego wczesnej diagnostyki, co może ułatwić skuteczną prewencję powikłań, w tym chorób sercowo-naczyniowych. Ze względu na rozbieżności w definicjach zespołu metabolicznego [1], opracowanie jednolitego narzędzia analitycznego, które umożliwi wczesną, trafną i zautomatyzowaną identyfikację pacjentów zagrożonych jego wystąpieniem na podstawie danych klinicznych, jest wyjątkowo trudne. Celem niniejszego projektu jest stworzenie modelu predykcyjnego, który, bazując na danych klinicznych pacjentów, umożliwi identyfikację osób spełniających kryteria rozpoznania zespołu metabolicznego.

Podział prac w zespole był następujący:
1. Jaskuła Z. – rozdziały 1, 2 i 5,
2. Kozakiewicz K. – rozdziały 3, 4 i 5,
3. Miklaszewska N. – rozdziały 3 i 4.

2 Przygotowanie i analiza zbioru danych

W projekcie wykorzystano dane pochodzące z publicznie dostępnego zbioru opublikowanego na platformie Kaggle. Zawiera on zarówno informacje demograficzne, jak i kliniczne dotyczące pacjentów, sklasyfikowanych ze względu na obecność lub brak zespołu metabolicznego. Zbiór zawiera numer identyfikacyjny pacjentów, 9 zmiennych ilościowych oraz 5 zmiennych jakościowych. Umożliwia analizę zależności pomiędzy cechami jednostki a występowaniem schorzenia, co czyni go odpowiednim do zastosowań predykcyjnych. Szczegółowy opis zmiennych przedstawiono w Tablicy 1:

Tablica 1: Opis zmiennych zawartych w zbiorze danych

Zmienna	Opis
seqn	numer identyfikacyjny
Age	wiek badanego (w latach)
Sex	płeć badanego (kobieta/mężczyzna)
Marital	stan cywilny jednostki
Income	dochód
Race	rasa
WaistCirc	pomiar obwodu talii (w cm)
BMI	wskaźnik masy ciała
Albuminuria	ilość albumin (białek) w moczu (wartości: 0, 1, 2)
UrAlbCr	stosunek albuminy do kreatyniny w moczu
UricAcid	poziom kwasu moczowego we krwi (w mg/dl)
BloodGlucose	poziom glukozy we krwi (w mg/dl)
HDL	poziom HDL ("dobrego" cholesterolu) (w mg/dl)
Triglycerides	poziom trójglicerydów we krwi (w mg/dl)
MetabolicSyndrome	obecność zespołu metabolicznego (0 = brak, 1 = obecność)

Źródło: opracowanie własne na podstawie zbioru danych

2.1 Statystyki opisowe

Zmiennymi demograficznymi uwzględnionymi w zbiorze są: wiek (`Age`), płeć (`Sex`), stan cywilny (`Marital`), dochód (`Income`) oraz rasa (`Race`). Część kliniczna obejmuje pomiary antropometryczne i wyniki badań laboratoryjnych istotnych w kontekście diagnostyki zespołu metabolicznego.

Zbiór danych obejmuje 2401 osób (1211 kobiet i 1190 mężczyzn) w wieku od 20 do 80 lat (średnia 48.69 lat). Większość badanych jest w średnim wieku. Wśród uczestników dominują osoby zamężne/żonate (ok. 50%), a rozkład dochodów jest zróżnicowany – od 300 do 9000 jednostek (średnio 4005). Dane obejmują też informacje o przynależności rasowej; największe

Tablica 2: Statystyki zbiorcze

Variable	N	Średnia	Odch. Std.	Min	Q1	Mediana	Q3	Max
Age	2401	48.69	17.63	20	34	48	63	80
Sex	2401							
... Female	1211	50.44%						
... Male	1190	49.56%						
Marital	2401							
...	208	8.66%						
... Divorced	242	10.08%						
... Married	1192	49.65%						
... Separated	95	3.96%						
... Single	498	20.74%						
... Widowed	166	6.91%						
Income	2284	4005	2954	300	1600	2500	6200	9000
Race	2401							
... Asian	349	14.54%						
... Black	548	22.82%						
... Hispanic	257	10.7%						
... MexAmerican	253	10.54%						
... Other	61	2.54%						
... White	933	38.86%						
WaistCirc	2316	98.31	16.25	56.2	86.68	97	107.6	176
BMI	2375	28.7	6.662	13.4	24	27.7	32.1	68.7
Albuminuria	2401							
... 0	2089	87.01%						
... 1	254	10.58%						
... 2	58	2.42%						
UrAlbCr	2401	43.63	258.3	1.4	4.45	7.07	13.69	5928
UricAcid	2401	5.489	1.439	1.8	4.5	5.4	6.4	11.3
BloodGlucose	2401	108.2	34.82	39	92	99	110	382
HDL	2401	53.37	15.19	14	43	51	62	156
Triglycerides	2401	128.1	95.32	26	75	103	150	1562
MetabolicSyndrome	2401							
... 0	1579	65.76%						
... 1	822	34.24%						

Źródło: opracowanie własne na podstawie zbioru danych

grupy stanowią osoby białe (933 osób, co stanowi ok. 39% badanych) i czarnoskóre (548, ok. 23%).

Średni obwód talii wynosi 98.31 cm, przy czym wartości skrajne sięgają od 56.2 cm do 176 cm. Średnia wartość BMI to 28.7, co klasyfikuje przeciętnego badanego jako osobę z nadwagą [4]. 25% badanych ma BMI powyżej 30, co jest klasyfikowane jako otyłość. W zestawie są zarówno osoby z niedowagą (BMI < 17), jak i ze skrajną otyłością (BMI > 40).

Albuminuria to zjawisko, gdy organizm wydala z organizmu białka wraz z moczem. Jest to naturalne, jednak wysokie stężenie albumin w moczu może być objawem nieprawidłowości w pracy nerek, spowodowanej np. cukrzycą lub nieskutecznie leczonym nadciśnieniem tętniczym [5]. W zbiorze stężenie białek jest indykatorem o wartościach: 0 - normalne stężenie albumin, 1 - podwyższony poziom, 2 - bardzo wysoki poziom albumin (białkomocz). Osoby z podwyższonym poziomem stanowią mniejszość. W przypadku zmiennej `UrAlbCr` opisującej stosunek albuminy do kreatyniny widać, że posiada ona bardzo duży przedział wartości. Mediana wynosi 7,07, ale występują wartości skrajnie wysokie (maksymalnie 5928), wskazując na obecność odstających obserwacji.

Poziom kwasu moczowego (`UricAcid`) we krwi wynosi średnio 5.49 mg/dl; około 75% badanych ma wyniki mieszczące się w normie, która wynosi między 3 a 7 mg/dl, w zależności od płci. Podwyższone ilości kwasu moczowego mogą być skutkiem niewłaściwej diety, otyłości lub cukrzycy. Wśród potencjalnych przyczyn wymieniany jest również zespół metaboliczny [6].

Poziom glukozy we krwi (`BloodGlucose`) również pokazuje dużą zmienność – od bardzo niskich do bardzo wysokich wartości – przy medianie 99 mg/dl, co jest wartością prawidłową.

Średni poziom "dobrego" cholesterolu (`HDL`) wynosi 53.37 mg/dl. Około połowa uczestników ma nieprawidłowy poziom HDL w stosunku do zalecanych norm (zależnie od płci granica nieprawidłowej ilości HDL wynosi między 40-50 mg/dl [7]). Średnie stężenie trójglicerydów (`Triglycerides`) wynosi 128.1 mg/dl, co jest wartością znajdującą się w normie (150-199 mg/dl) [8], jednak w zbiorze są jednostki mające wyniki znacznie powyżej poziomu bezpiecznego.

Zmienna `MetabolicSyndrome` informuje, że około 34.24% badanych ma już rozpoznany zespół metaboliczny.

2.2 Analiza braków

W zbiorze danych występują braki, które należy zbadać, i w miarę możliwości zniwelować. Z analizy statystyk opisowych wynika, że zmienna `Marital` zawiera 208 obserwacji oznaczonych jako "...". Ponadto zmienne `Income`, `WaistCirc` oraz `BMI` również zawierają braki, co zostało uwidocznione w tabeli statystyk zbiorczych, w kolumnie *N*, przedstawiającej liczbę obserwacji dla każdej z tych zmiennych. Sprawdzono, ile braków występuje dla każdej zmiennej oraz jaki stanowią one procent wszystkich obserwacji.

Najwięcej brakujących danych występuje w zmiennej `Marital` – stanowią one 8.66% wszystkich obserwacji. Jest to odsetek przekraczający powszechnie akceptowany próg 5%

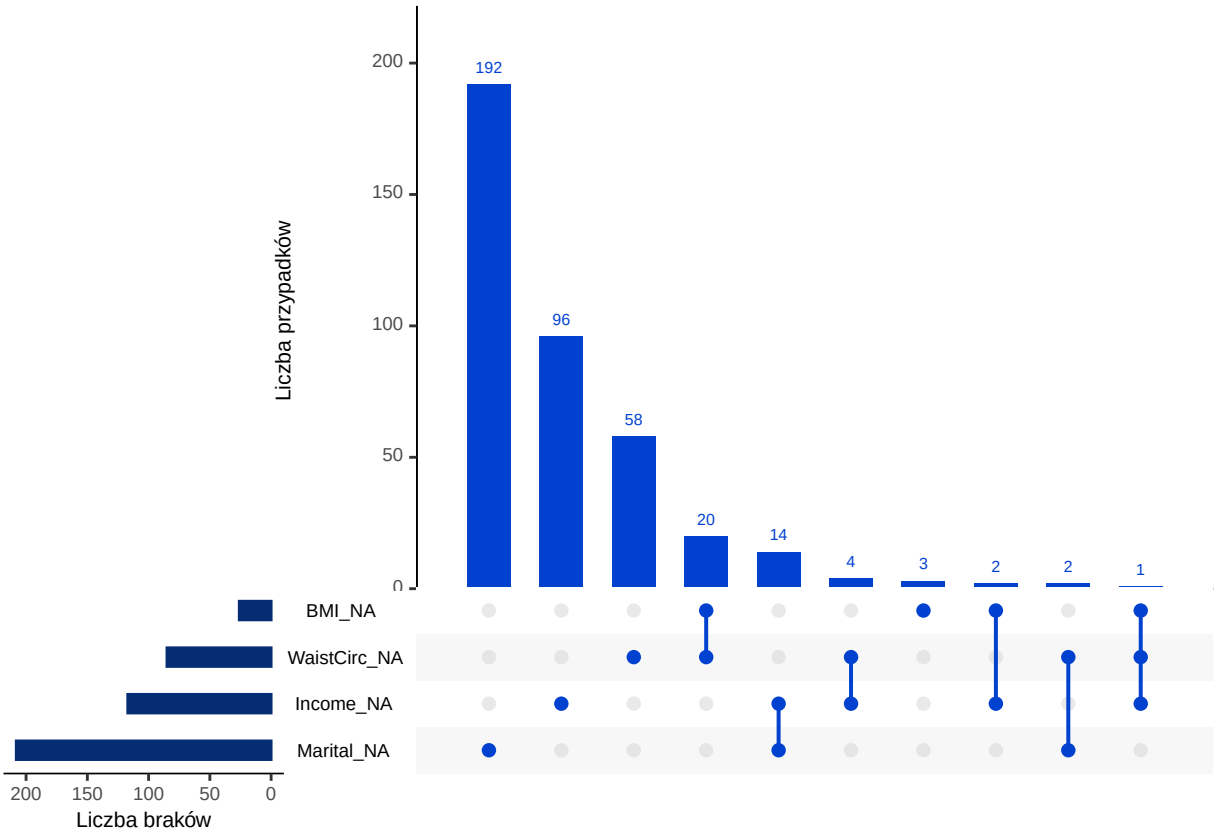
Tablica 3: Podsumowanie brakujących danych dla wybranych zmiennych

Zmienna	Liczba braków	Udział procentowy braków
Marital	208	8.66
Income	117	4.87
WaistCirc	85	3.54
BMI	26	1.08

Źródło: opracowanie własne na podstawie zbioru danych

braków. W pozostałych przypadkach braki stanowią mniej niż 5% wszystkich obserwacji. Dodatkowo, przeprowadzono analizę współwystępowania braków, uwzględniając zależności między brakującymi danymi w różnych zmiennych. Jej wyniki przedstawia Wykres 1.

Wykres 1: Rozkład braków oraz ich współwystępowania



Źródło: opracowanie własne na podstawie zbioru danych

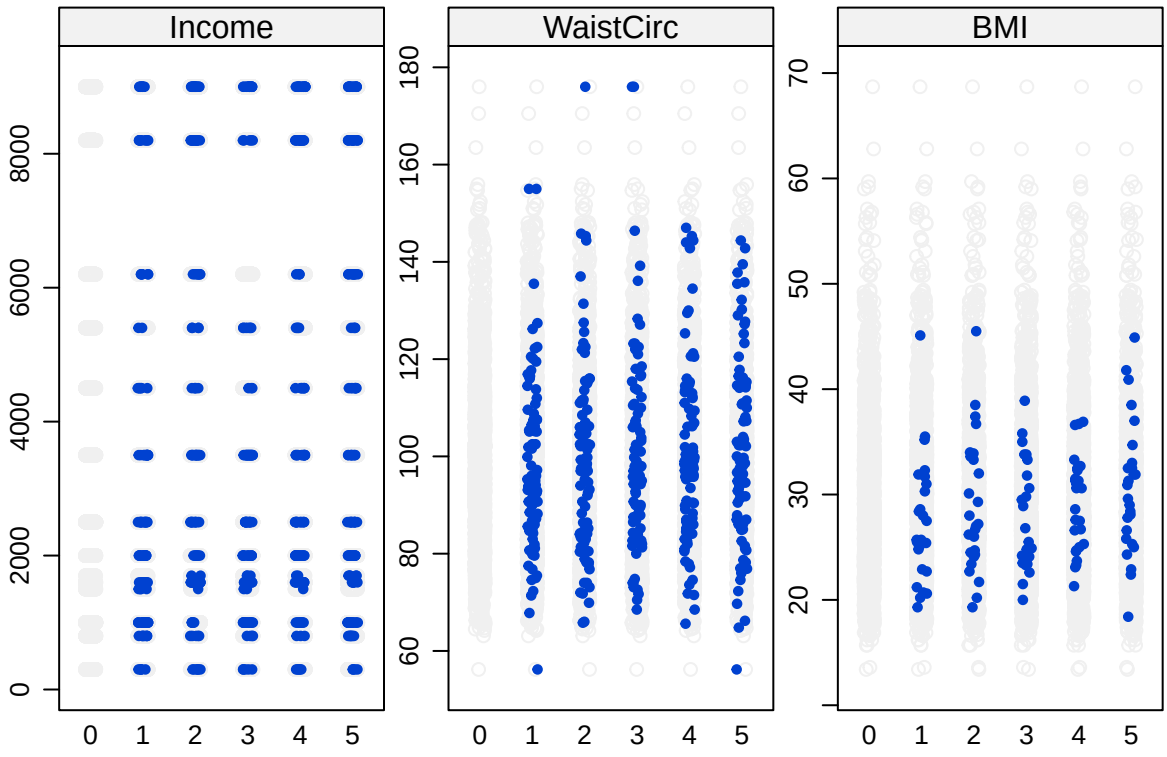
Najczęściej występującymi brakami są pojedyncze braki w jednej z trzech zmiennych: stanie cywilnym, dochodach lub obwodzie talii. W zbiorze danych odnotowano 20 przypadków brakujących wartości jednocześnie dla zmiennych `BMI` i `WaistCirc`, a także 14 przypadków braków dotyczących jednocześnie zmiennych `Income` i `Marital`.

W celu uzupełnienia brakujących danych zastosowano metodę imputacji wielokrotnej z wykorzystaniem procedury *MICE*. Metoda ta polega na wygenerowaniu kilku uzupełnionych wersji zbioru danych, w których wartości brakujące są zastępowane prognozowanymi wartościami opartymi na pozostałych zmiennych. Wygenerowanie kilku różnych zestawów pozwala na uwzględnienie niepewności związanej z brakami [9].

Dla zmiennych ilościowych zastosowano technikę *PMM* (ang. *Predictive Mean Matching*). Polega ona na znalezieniu obserwacji z brakującą wartością, a następnie dopasowaniu jej do jednej z obserwacji nieposiadającej braków, która wykazuje najbardziej zbliżone wartości innych zmiennych objaśniających. Wartość brakująca zostaje zastąpiona wartością z dopasowanego rekordu, co pozwala zachować zgodność z rozkładem empirycznym zmiennej. Dla zmiennej jakościowej `Marital` użyto techniki *polyreg*, odpowiedniej dla zmiennych kategorycznych, nieuporządkowanych, o więcej niż dwóch wariantach [10].

W procesie tworzenia zestawów imputowanych danych uwzględniono zmienną `Age`, która nie zawierała braków, aby poprawić predykcje brakujących wartości. Następnie porównano rozkłady zestawów imputowanych danych na wykresach (Wykres 2) oraz porównano statystyki zestawów: ich średnie, mediany i odchylenia standardowe.

Wykres 2: Rozkłady zestawów z imputowanymi danymi



Źródło: opracowanie własne na podstawie zbioru danych

Na podstawie analizy statystyk imputowanych zestawów danych (Tablica 4) stwierdzono, że zestaw 2 najtrafniej odwzorowuje rozkłady zmiennych oryginalnych. W przypadku

Tablica 4: Statystyki imputacji dla Income, WaistCirc i BMI

Zmienna	Zestaw	Średnia	Mediana	Odchylenie
Income	Oryginalne	4005.25	2500.00	2954.03
Income	Zestaw 1	3324.79	2500.00	2673.43
Income	Zestaw 2	4187.18	3500.00	2999.67
Income	Zestaw 3	3652.99	2500.00	2847.70
Income	Zestaw 4	4053.85	2500.00	2995.27
Income	Zestaw 5	3965.81	2500.00	2958.77
WaistCirc	Oryginalne	98.31	97.00	16.25
WaistCirc	Zestaw 1	98.19	95.30	17.57
WaistCirc	Zestaw 2	99.76	97.80	20.22
WaistCirc	Zestaw 3	100.46	97.80	20.33
WaistCirc	Zestaw 4	100.94	98.40	18.42
WaistCirc	Zestaw 5	101.17	99.00	19.25
BMI	Oryginalne	28.70	27.70	6.66
BMI	Zestaw 1	27.32	25.70	5.82
BMI	Zestaw 2	28.86	27.10	6.33
BMI	Zestaw 3	27.57	25.20	5.14
BMI	Zestaw 4	29.05	29.60	4.46
BMI	Zestaw 5	30.59	30.25	6.35

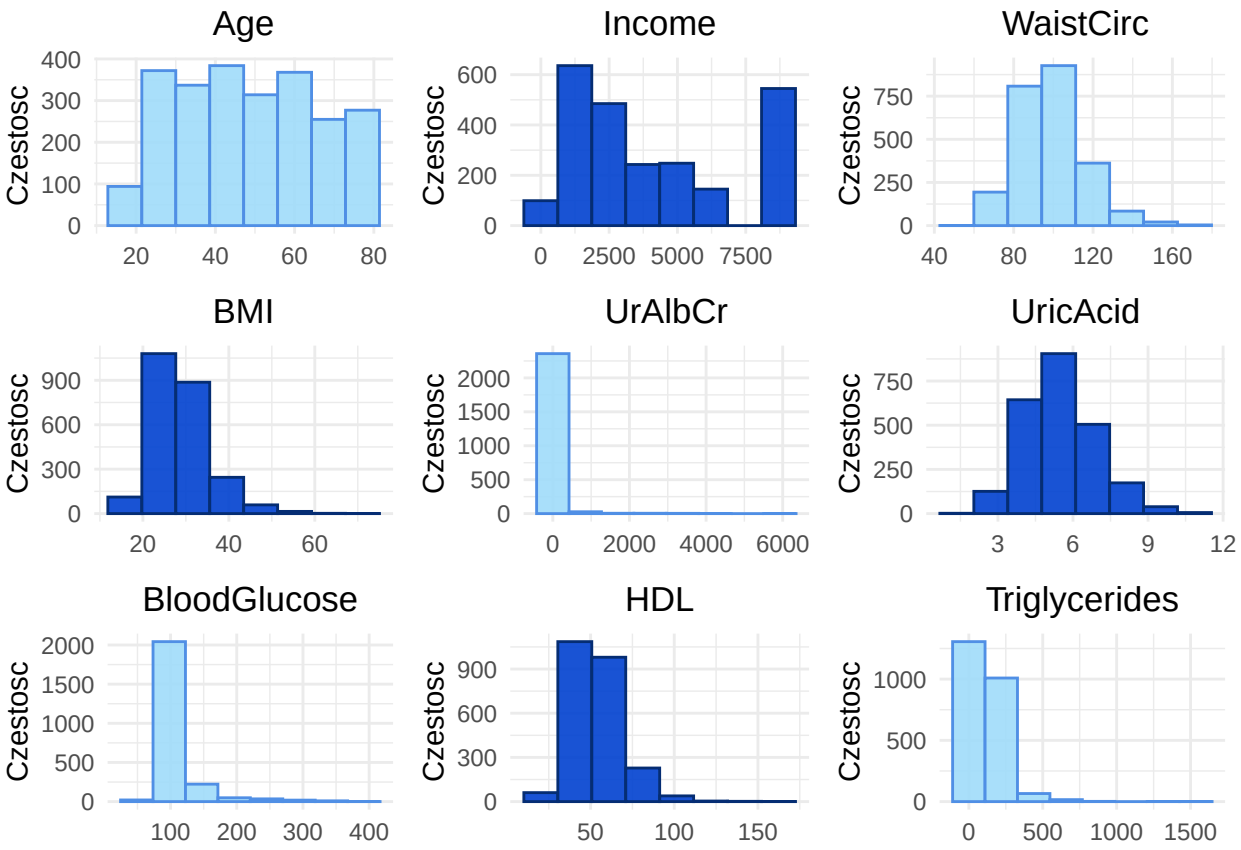
Źródło: opracowanie własne na podstawie zbioru danych

dochodu zestaw ten ma statystyki nieco bardziej odchylone od statystyk oryginalnych danych, ale zbliżone wartości średnich, median oraz odchyleń standardowych dla obwodu w talii oraz BMI uzasadniają wybór tego zestawu do dalszych działań.

2.3 Wizualizacje rozkładów

W celu lepszego zrozumienia struktury danych oraz identyfikacji potencjalnych różnic pomiędzy grupami pacjentów z rozpoznanym zespołem metabolicznym i bez niego, przeprowadzono wizualizację rozkładów zmiennych ilościowych. Dla każdej z tych zmiennych utworzono histogramy, które umożliwiają ocenę kształtu rozkładu oraz porównanie wartości pomiędzy dwiema grupami zdefiniowanymi przez zmienną `MetabolicSyndrome`.

Wykres 3: Rozkłady zmiennych ilościowych

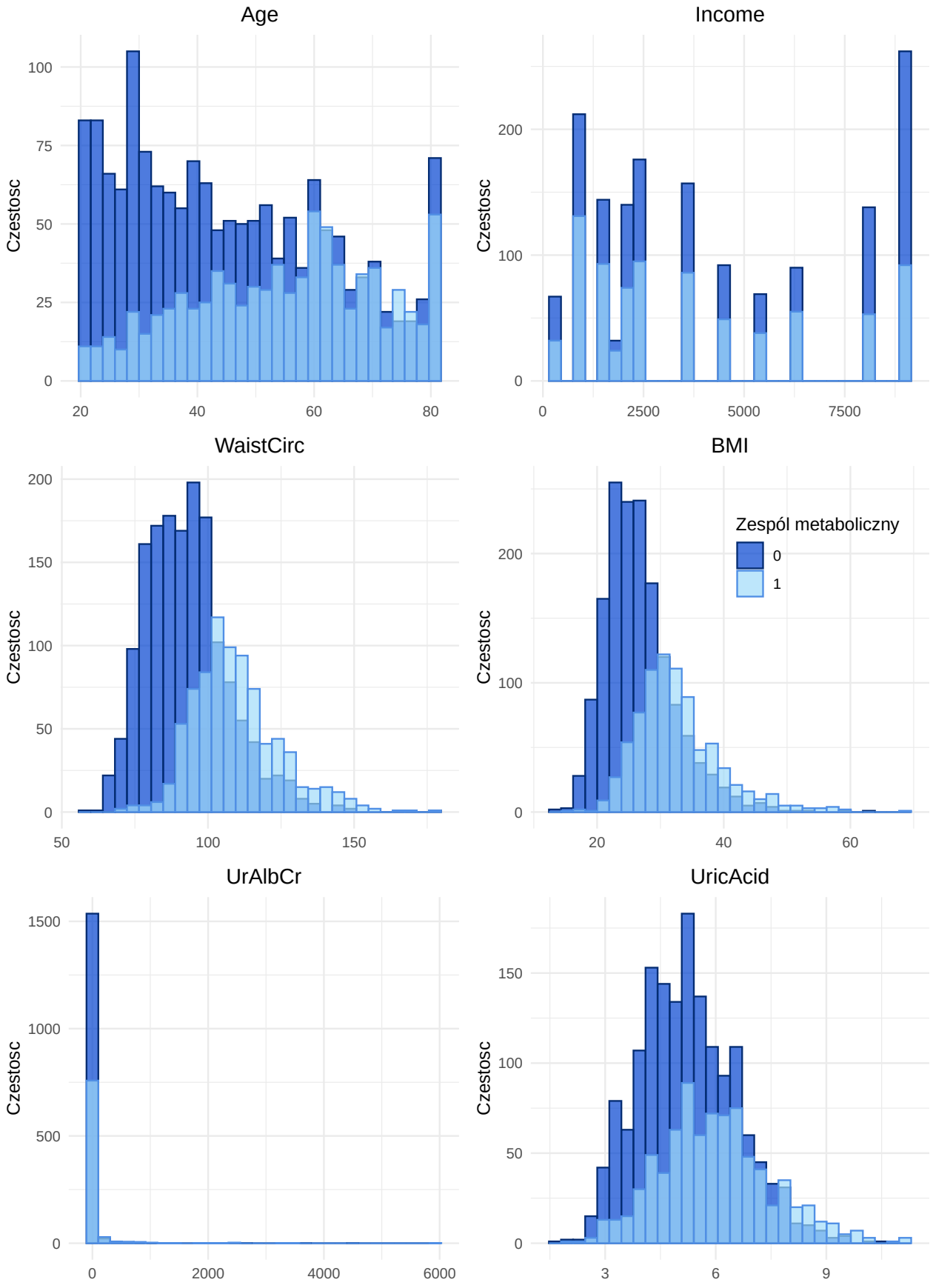


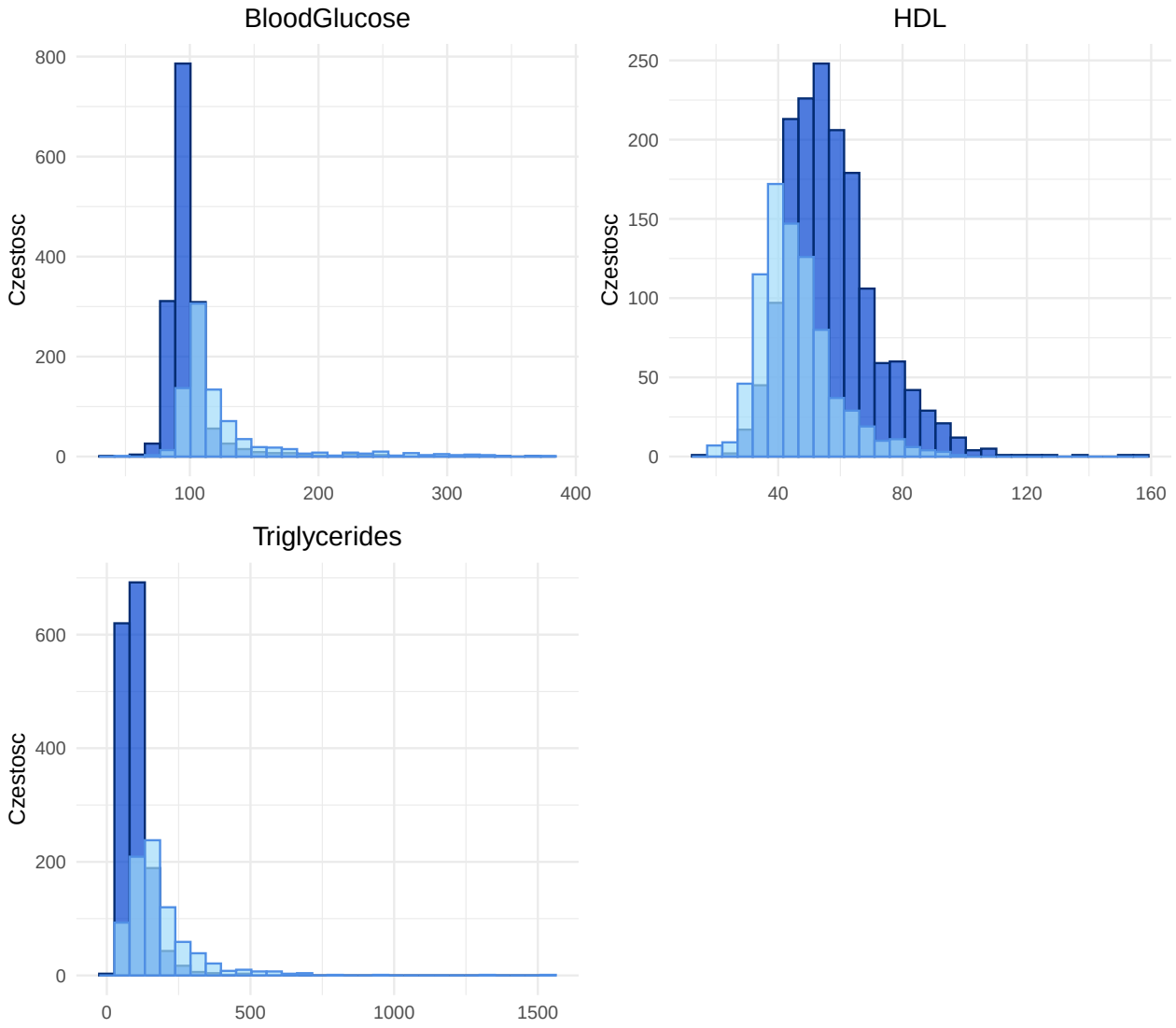
Źródło: opracowanie własne na podstawie zbioru danych

Histogramy zmiennych `BMI`, `UrAlbCr`, `BloodGlucose`, `HDL` oraz `Triglycerides` pokazują, że zmienne te zawierają wartości odstające, które mogą negatywnie wpływać na predykcje modelu. Zmienna `WaistCirc` również nie ma rozkładu normalnego, jednak jej zbiór wartości jest najbardziej "spójny" i nie zawiera punktów odstających o bardzo dużych wartościach. Rozkład wieku wskazuje, że w zbiorze znajduje się mniej więcej po tyle samo pacjentów (między 300 a 400 osób) w każdej grupie wiekowej, z wyjątkiem osób w wieku 20 lat, których jest kilkukrotnie mniej.

Po stworzeniu oddzielnych wykresów częstości (Wykres 4) dla osób z zespołem metabolicznym oraz bez niego, zauważalna jest znaczna różnica w rozkładach niektórych zmiennych.

Wykres 4: Rozkłady zmiennych ilościowych z uwzględnieniem informacji o posiadaniu lub nieposiadaniu zespołu metabolicznego


Age
Income
WaistCirc
BMI
UrAlbCr
UricAcid
Czestosc
Zespól metaboliczny
0
1



Źródło: opracowanie własne na podstawie zbioru danych

Analizując histogram wieku osób zdrowych, można stwierdzić, że większość z nich to osoby poniżej 45. roku życia, a rozkład wieku w tej grupie przyjmuje charakter prawoskośny. Z kolei histogram osób z zespołem metabolicznym wykazuje bardziej wyrównany rozkład, z widoczną koncentracją przypadków w starszych grupach wiekowych, co może sugerować zwiększone ryzyko wystąpienia zespołu metabolicznego w późniejszym wieku. W przypadku zmiennych takich jak obwód talii oraz wskaźnik BMI, również dostrzega się przesunięcie wykresu częstości w prawo dla osób chorych. Podobny trend występuje w przypadku poziomu kwasu moczowego we krwi oraz glukozy. Natomiast rozkład zmiennej `HDL` wśród osób z zespołem metabolicznym charakteryzuje się niższymi wartościami HDL w porównaniu do osób zdrowych. Z kolei dochody nie wykazują istotnych różnic pomiędzy obiema grupami.

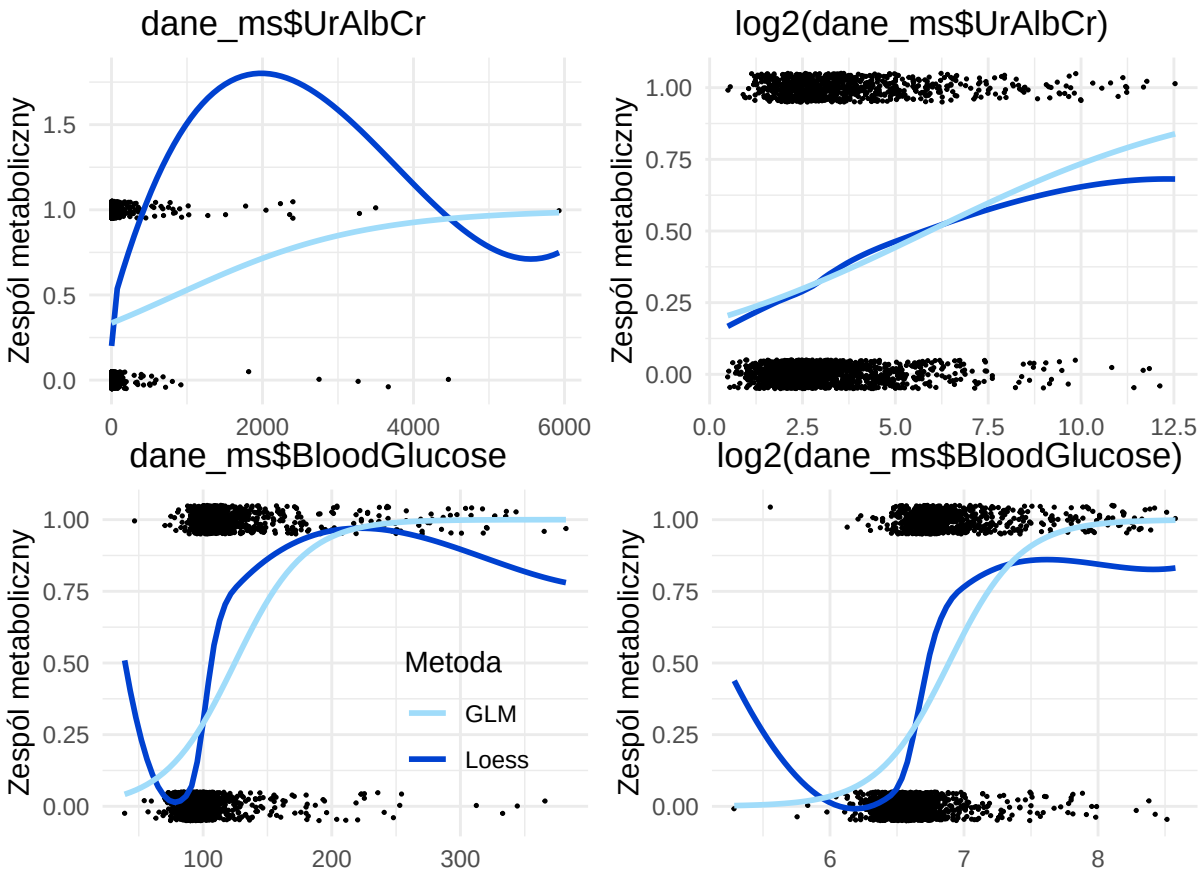
Zmienne `UrAlbCr`, `BloodGlucose` i `Tryglicerides` mają bardzo asymetryczne rozkłady (prawoskośne) z ekstremalnie wysokimi wartościami maksymalnymi. To oznacza obecność wartości odstających i dużego rozrzutu, który może zaburzać dopasowanie modeli.

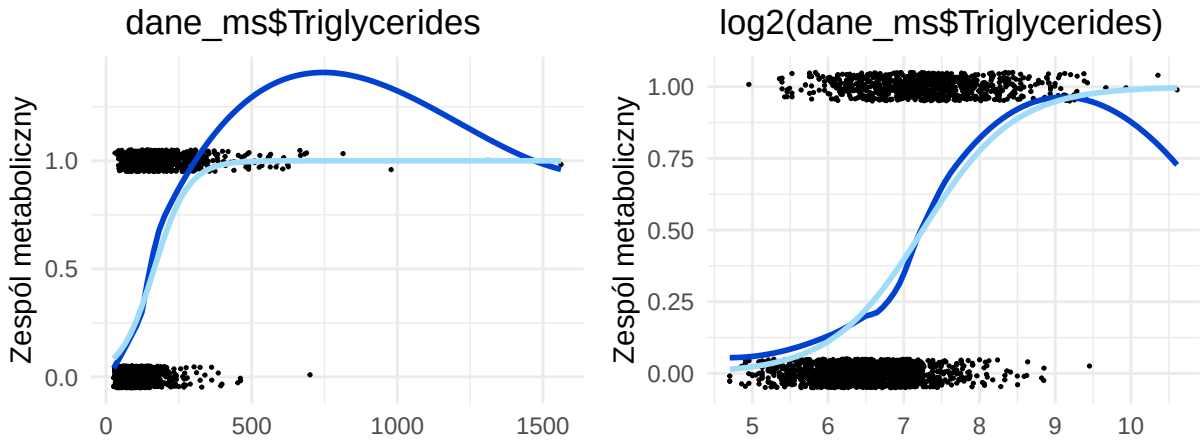
W celu dokładniejszego zbadania tego zjawiska wykonano wykresy punktowe z liniami:

- *loess* (ciemnoniebieska linia), która pokazuje nieliniowy związek zmiennej z obecnością zespołu metabolicznego,
- *glm* (jasnoniebieska linia), która pokazuje kształt dopasowania modelu logistycznego opartego na jednej, badanej zmiennej.

Porównując te dwie linie można ocenić, czy regresja logistyczna dobrze odwzorowuje zależność, czy może warto zastosować np. transformację logarytmiczną. Jeśli krzywa *loess* znacznie odbiega od *glm*, to znaczy, że regresja logistyczna może niedokładnie odwzorowywać związek i logarytmowanie może poprawić liniowość w modelu.

Wykres 5: Porównanie LOESS i GLM dla zmiennych asymetrycznych





Źródło: opracowanie własne na podstawie zbioru danych

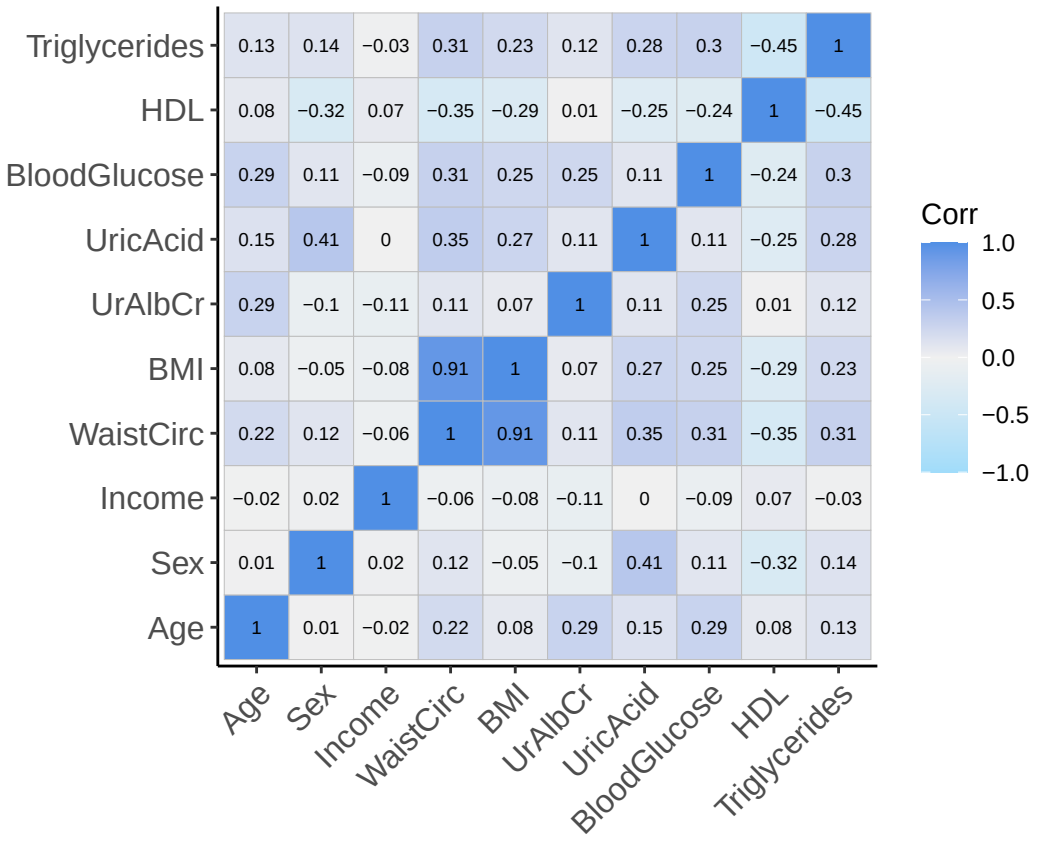
Dla wszystkich trzech zmiennych w oryginalnej (nielogarytmicznej) skali zauważalne są istotne rozbieżności między krzywą *loess* a dopasowaniem *glm*. Po transformacji logarytmicznej, przebieg krzywych staje się bardziej zbliżony. Szczególnie widoczne jest to dla zmiennych `UrAlbCr` i `Triglycerides`, gdzie zastosowanie przekształcenia skutkuje wyraźną poprawą dopasowania modelu logistycznego – linia regresji przyjmuje kształt zgodny z teoretyczną krzywą logistyczną. Dla `BloodGlucose` efekt ten jest również zauważalny, choć nieco mniej wyraźny. Powyższa analiza potwierdza, że zmienne o silnej asymetrii powinny zostać poddane transformacji przed włączeniem do modelu, aby poprawić liniowość zależności i stabilność parametrów regresji.

2.4 Analiza korelacji i zależności między zmiennymi

W celu lepszego zrozumienia zależności pomiędzy zmiennymi w zbiorze danych przeprowadzono analizę korelacji. Pozwala ona ocenić siłę i kierunek powiązań zarówno pomiędzy cechami ilościowymi, jak i jakościowymi, a także między zmiennymi różnego typu. Identyfikacja silnych korelacji może wskazywać na współzależności, które warto uwzględnić przy budowie modeli predykcyjnych lub w dalszej interpretacji wyników.

Analizę korelacji rozpoczęto od stworzenia macierzy korelacji między zmiennymi ilościowymi (z wykorzystaniem korelacji Pearsona). W zbiorze znajduje się jedna zmienna zero-jedynkowa – płeć. W przypadku takiej zmiennej dychotomicznej można zastosować współczynnik korelacji punktowo-dwuseryjnej, który mierzy siłę i kierunek związku pomiędzy zmienną ciągłą a zmienną binarną. Współczynnik ten jest równoważny współczynnikowi korelacji Pearsona, dlatego zmienną płeć uwzględniono w macierzy korelacji.

Wykres 6: Korelacje między zmiennymi ilościowymi

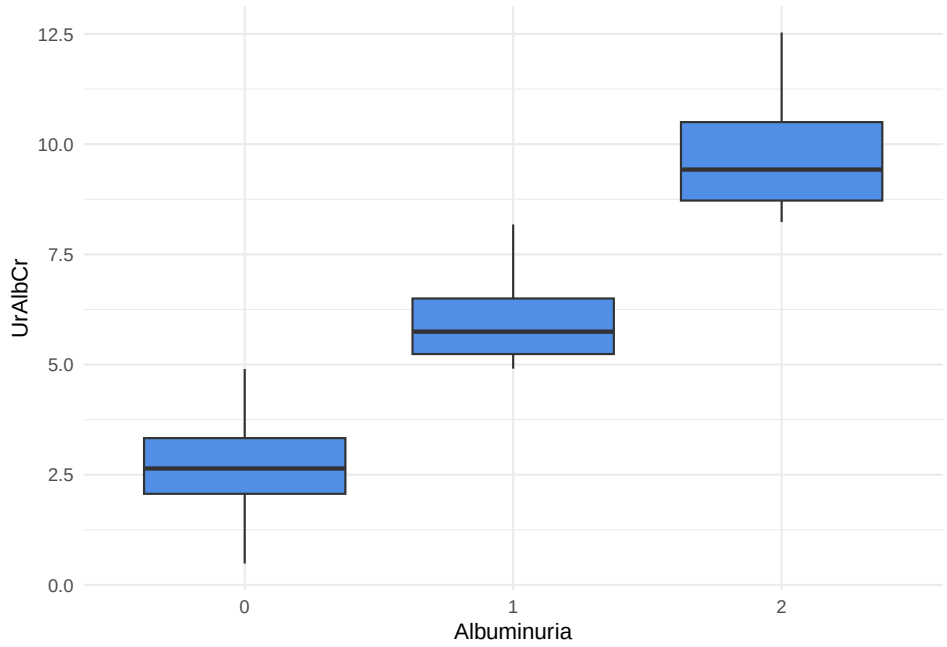


Źródło: opracowanie własne na podstawie zbioru danych

Jak przedstawiono na Wykresie 6, między analizowanymi zmiennymi ilościowymi zaobserwowano słabe korelacje, z wyjątkiem pary `BMI` i `WaistCirc`. Ze względu na możliwość wystąpienia współliniowości, obwód w talii został wykluczony z dalszych analiz. Uznano, że wskaźnik `BMI` niesie więcej informacji, ponieważ uwzględnia zarówno masę ciała, jak i wzrost pacjenta.

W analizie zmiennych zdecydowano się na odrzucenie zmiennej `Albuminuria` na rzecz zmiennej `UrAlbCr`, która określa stosunek albumin do kreatyniny w moczu i dostarcza bardziej szczegółowych informacji dotyczących stanu zdrowia pacjenta. Podjęta decyzja została potwierdzona poprzez jednoczynnikową analizę wariancji (*ANOVA*), zastosowaną w celu oceny, czy wartości zmiennych ilościowych różnią się istotnie pomiędzy kategoriami zmiennej `Albuminuria`. Wyniki testu wskazały na statystycznie istotny wpływ zmiennej na zróżnicowanie wartości zmiennych ilościowych, co zilustrowano na Wykresie 7.

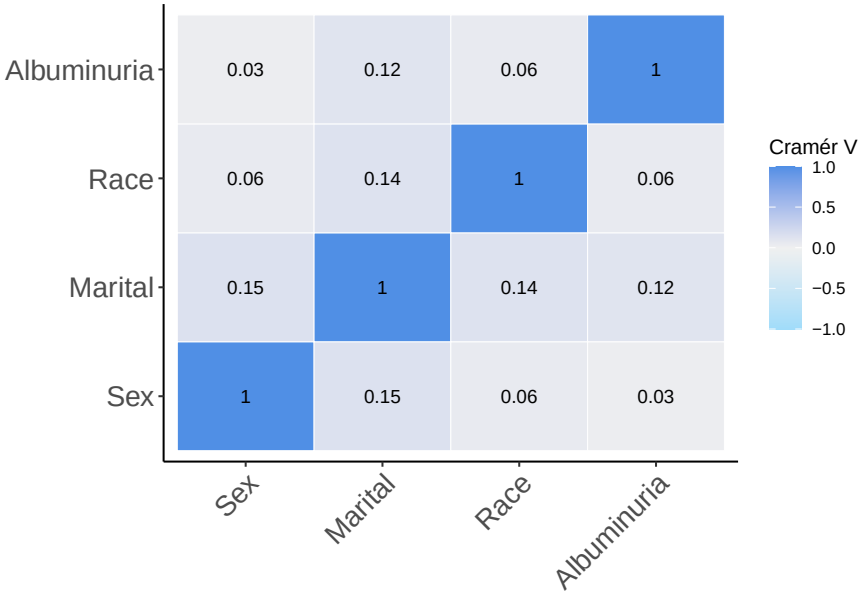
Wykres 7: Rozkład zmiennej UrAlbCr względem Albuminurii



Źródło: opracowanie własne na podstawie zbioru danych

Dla par zmiennych jakościowych obliczono współczynnik kontyngencji V Craméra, który służy do pomiaru siły powiązania. Przyjęto, że powiązanie jest silne, gdy $V > 0.5$. W analizowanym zbiorze danych nie stwierdzono wartości przekraczających przyjęty próg.

Wykres 8: Korelacje między zmiennymi jakościowymi



Źródło: opracowanie własne na podstawie zbioru danych

3 Budowa modeli

W celu otrzymania modelu z jak najlepszą zdolnością predykcyjną, przed przystąpieniem do jego budowy, oryginalne dane podzielone zostały na zbiór uczący i zbiór testowy. Modele budowane będą na zbiorze uczącym, który stanowi 70% całego zbioru danych, natomiast testowane będą na zbiorze testowym, który stanowi 30% oryginalnego zbioru. Taki podział pozwoli uniknąć przeuczenia modelu i możliwe będzie sprawdzenie jak radzi on sobie z nowymi danymi. W Tablicy 5 znajdują się proporcje wariantów zmiennej zależnej `MetabolicSyndrome` w pełnych danych i w danych uczących i testowych. Proporcje "odpowiedzi" zostały zachowane, zatem obydwa podzbiory dobrze reprezentują oryginalny zbiór danych.

Tablica 5: Proporcje w zbiorach danych

	0	1
dane	0.658	0.342
zbiór uczący	0.660	0.340
zbiór testowy	0.653	0.347

Źródło: opracowanie własne na podstawie zbioru danych

Zmienną objaśnianą jest zmienna dychotomiczna `MetabolicSyndrome`. Zmiennymi objaśniającymi będą natomiast wszystkie wymienione w poprzednim rodziale zmienne z wyjątkiem dwóch. Przy analizie korelacji zmienne `BMI` i `WaistCirc` oraz `Albuminuria` i `UrAlbCr` wykazały między sobą zbyt dużą korelację, w wyniku czego zmienne `WaistCirc` i `Albuminuria` wykluczone zostały z dalszych analiz. Do budowy modeli, jako zmienne objaśniające, wykorzystane zostaną zatem 3 zmienne jakościowe (`Sex`, `Marital`, `Race`) oraz 8 zmiennych ilościowych (`Age`, `Income`, `BMI`, `UrAlbCr`, `UricAcid`, `BloodGlucose`, `HDL`, `Triglycerides`).

3.1 Estymacja modelu dwumianowego logitowego

Estymacja modeli dwumianowych logitowych jest podstawową metodą analizy zależności między dychotomiczną zmienną endogeniczną (przyjmującą dwie wartości – występowanie lub brak zespołu metabolicznego), a zestawem predyktorów. Model logitowy szacuje prawdopodobieństwo zajścia danego zdarzenia przy założeniu, że logarytm ilorazu szans (logit) jest liniową funkcją zmiennych objaśniających, co matematycznie można zapisać w postaci:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Analizę rozpoczęto od estymacji modelu logitowego uwzględniającego wszystkie 11 wcześniej zidentyfikowanych zmiennych objaśniających. Model ten stanowi punkt odniesienia dla dalszej analizy, w tym oceny istotności poszczególnych predyktorów oraz ewentualnej ich selekcji.

W celu oceny ogólnej istotności predyktorów zastosowano test ilorazu wiarygodności, porównujący model zerowy (zawierający jedynie wyraz wolny) z modelem pełnym obejmującym wszystkie zmienne objaśniające. Hipoteza zerowa dla testu zakłada, że żaden z predyktorów nie wpływa na zmienną zależną, co oznacza brak związku między nimi. Test ten pozwala sprawdzić, czy dodanie zmiennych istotnie poprawia dopasowanie modelu do danych. Dla modelu *logit0* uzyskano p-wartość statystyki χ^2 mniejszą niż przyjęty poziom istotności $\alpha = 0.05$. W związku z tym hipotezę zerową odrzucono na rzecz hipotezy alternatywnej, co wskazuje, że przynajmniej jedna z uwzględnionych zmiennych ma istotny wpływ na zmienną zależną.

Tablica 6: Podsumowanie modelu logit0

	Ocena parametru modelu	Błąd std.	Statystyka testowa	p-value
(Intercept)	-34.276	2.494	-13.743	0.000
Age	0.041	0.006	7.143	0.000
SexMale	-0.808	0.186	-4.348	0.000
MaritalMarried	0.165	0.248	0.666	0.505
MaritalSeparated	0.722	0.413	1.748	0.081
MaritalSingle	0.523	0.289	1.811	0.070
MaritalWidowed	-0.772	0.367	-2.105	0.035
RaceBlack	0.159	0.301	0.527	0.598
RaceHispanic	0.203	0.319	0.636	0.525
RaceMexAmerican	0.256	0.333	0.770	0.441
RaceOther	-0.707	0.577	-1.226	0.220
RaceWhite	0.154	0.262	0.587	0.557
Income	0.000	0.000	1.069	0.285
BMI	0.137	0.014	9.577	0.000
UrAlbCr	0.077	0.047	1.644	0.100
UricAcid	0.121	0.063	1.915	0.055
BloodGlucose	2.709	0.311	8.700	0.000
HDL	-0.048	0.007	-6.564	0.000
Triglycerides	1.561	0.135	11.605	0.000

Test ilorazu wiarygodności (p-value): 0 | Wynik VIF: niewspółliniowe | AIC: 1185.126
Źródło: opracowanie własne na podstawie zbioru danych

Wskaźnik inflacji wariancji (VIF, ang. *Variance Inflation Factor*) jest miarą używaną do oceny poziomu współliniowości pomiędzy zmiennymi objaśniającymi w modelu regresji. Współczynnik ten informuje, w jakim stopniu wariancja oszacowania parametru regresji dla zmiennej X_j jest zawyżona z powodu jej współliniowości z pozostałymi zmiennymi objaśniającymi. Matematycznie, VIF dla zmiennej X_j definiuje się jako:

$$VIF_j = \frac{1}{1 - R_j^2}$$

gdzie R_j^2 to współczynnik determinacji uzyskany z regresji zmiennej X_j na pozostałe zmienne objaśniające X_i, $i \neq j$. Mówiąc inaczej, R_j to współczynnik korelacji liniowej wielorakiej między X_j a pozostałymi zmiennymi w modelu.

Wskaźnik VIF zawsze przyjmuje wartości większe lub równe 1. Wartość $VIF = 1$ oznacza brak współliniowości – zmienna X_j nie jest liniowo związana z żadną inną zmienną objaśniającą w modelu, co jest sytuacją idealną. Za umowną granicę akceptowalnej współliniowości przyjęto wartość na poziomie 2.5. Dla oszacowanego modelu logitowego, wartości GVIF (*Generalized Variance Inflation Factor*) wskazują na brak istotnej współliniowości pomiędzy zmiennymi objaśniającymi. Zmienne kategoryczne (np. `Marital`, `Race` i `Sex`) miały wyższe surowe wartości GVIF, ale po przeskalowaniu, ich wartości także były bardzo niskie (ok. 1.04–1.25), co oznacza, że nie wnoszą istotnej współliniowości. Wszystkie wartości mieściły się znacząco poniżej umownej granicy, zatem zmienne egzogeniczne są relatywnie niezależne od siebie.

Na podstawie podsumowania modelu `logit0`, zawartego w Tablicy 6, można zauważyć, że nie wszystkie zmienne są istotne statystycznie w kontekście budowy modelu. P-wartości dla statystyki z przekraczające przyjęty poziom istotności $\alpha = 0.1$ dotyczą zmiennych `Race`, `Income` i `UrAlbCr`, co wskazuje na ich nieistotność. Obniżoną istotność statystyczną wykazują również zmienne `Marital` oraz `UricAcid`. W związku z tym, oszacowany zostanie nowy model logitowy, uwzględniający wszystkie wcześniej rozpatrywane predyktory z wyjątkiem zmiennych jednoznacznie nieistotnych (`Race`, `Income`). Uwzględniona zostanie natomiast zmienna `UrAlbCr`, z uwagi na fakt, że p-wartość dla tej zmiennej jest bliska najwyższemu z rozważanych poziomów istotności ($\alpha = 0.1$). Dodatkowo w przeciwieństwie do dwóch wymienionych zmiennych nieistotnych, zawiera ona informacje kliniczne, które w kontekście identyfikacji występowania zespołu metabolicznego mogą mieć większe znaczenie.

Dla modelu `logit1` test ilorazu wiarygodności także pozwala na odrzucenie hipotezy zerowej, mówiącej o nieistotności wszystkich zmiennych w modelu. Wartości wskaźnika VIF dla drugiego modelu logitowego wskazują, że nie występuje istotny problem współliniowości między zmiennymi objaśniającymi. Najwyższe wartości dotyczą zmiennych `Age` (1.60) i `Sex` (1.54), ale nawet te wartości pozostają znacznie poniżej umownej granicy 2.5, przy której można by rozważać potencjalne problemy ze współliniowością. Wszystkie przeskalowane wartości GVIF mieszczą się w przedziale od około 1.04 do 1.27, co oznacza jedynie niewielką korelację pomiędzy predyktorami. W związku z tym uznano, że predyktory nie są ze sobą nadmiernie skorelowane.

Podsumowanie modelu *logit1*, przedstawione w Tablicy 7, nadal wskazuje zmienną `UrAlbCr` jako zmienną nieistotną. Dodatkowo można zauważyć, że warianty zmiennej `Marital` wykazują różną istotność statystyczną. W szczególności p-wartość dla wariantu `Marital: Married` jest wyższa od najwyższego poziomu istotności (równego 0.1), wskazując na jego nieistotność. Inne warianty istotne są na poziomach 0.05 i 0.1, co jest wynikiem gorszym niż dla pozostałych zmiennych. Podobnie jak w przypadku poprzedniego modelu, zmienną nieuwzględnioną w kolejnym modelu będzie zmienna demograficzna (`Marital`). Pomimo nieistotności statystycznej zmiennej `UrAlbCr`, będzie ona w dalszym ciągu brana pod uwagę przy estymacji kolejnego modelu, ponownie, ze względu na wnoszenie do niego informacji klinicznych.

Tablica 7: Podsumowanie modelu logit1

	Ocena parametru modelu	Błąd std.	Statystyka testowa	p-value
(Intercept)	-33.557	2.434	-13.788	0.000
Age	0.041	0.006	7.290	0.000
SexMale	-0.800	0.184	-4.338	0.000
MaritalMarried	0.202	0.240	0.842	0.400
MaritalSeparated	0.743	0.407	1.826	0.068
MaritalSingle	0.480	0.286	1.678	0.093
MaritalWidowed	-0.778	0.365	-2.128	0.033
BMI	0.138	0.014	10.144	0.000
UrAlbCr	0.074	0.046	1.618	0.106
UricAcid	0.120	0.062	1.935	0.053
BloodGlucose	2.657	0.307	8.663	0.000
HDL	-0.048	0.007	-6.596	0.000
Triglycerides	1.544	0.131	11.793	0.000

Test ilorazu wiarygodności (p-value): 0 | Wynik VIF: niewspółliniowe | AIC: 1177.611

Źródło: opracowanie własne na podstawie zbioru danych

Tablica 8: Podsumowanie modelu logit2

	Ocena parametru modelu	Błąd std.	Statystyka testowa	p-value
(Intercept)	-32.885	2.416	-13.609	0.000
Age	0.033	0.005	6.600	0.000
SexMale	-0.725	0.180	-4.029	0.000
BMI	0.138	0.014	10.208	0.000
UrAlbCr	0.082	0.045	1.802	0.071
UricAcid	0.108	0.062	1.763	0.078
BloodGlucose	2.646	0.307	8.623	0.000
HDL	-0.046	0.007	-6.426	0.000
Triglycerides	1.534	0.130	11.819	0.000

Test ilorazu wiarygodności (p-value): 0 | Wynik VIF: niewspółliniowe | AIC: 1184.911

Źródło: opracowanie własne na podstawie zbioru danych

Wyestymowano nowy model `logit2` zawierający zmienne wiek i płeć oraz istotne zmienne kliniczne (z uwzględnieniem `UrAlbCr`). Ponieważ p-wartość dla testu wiarygodności jest znacznie mniejsza niż przyjęty poziom istotności $\alpha = 0.05$, odrzucono hipotezę zerową o braku wpływu zmiennych objaśniających na zmienną zależną. Oznacza to, że ujęte w modelu predyktory istotnie poprawiają jego dopasowanie do danych.

W obecnym modelu wszystkie zmienne są numeryczne lub binarne, dlatego każda ma tylko

jeden parametr i klasyczny VIF = GVIF, a przeskalowanie nie jest potrzebne. Wskaźnik VIF również nie wskazuje na istotną współliniowość. Wszystkie wartości mieszczą się w przedziale od 1.10 do 1.50, co oznacza bardzo niski poziom współzależności pomiędzy predyktorami. Najwyższe wartości można zaobserwować dla zmiennych `Sex` (1.50) i `HDL` (1.34), ale nadal są one dalekie od granicy uznawanej za problematyczną (2.5).

Model `logit2`, którego podsumowanie zostało przedstawione w TABLICY 8, dostarcza czytelnych i spójnych wyników estymacji. Większość zmiennych okazała się statystycznie istotna na poziomie $\alpha = 0.001$. P-wartość dla zmiennej `UricAcid` znajduje się poniżej poziomu istotności $\alpha = 0.1$, natomiast zmienna `UrAlbCr` mimo wcześniejszej nieistotności, dla tego modelu okazała się być istotna, także na poziomie istotności 0.1. Wyniki tych zmiennych mogą sugerować ich potencjalną rolę, choć z mniejszą pewnością statystyczną. Znaki estymowanych współczynników są zgodne z oczekiwaniami – wyższe wartości BMI, glukozy, trójglicerydów i kwasu moczowego zwiększają ryzyko, natomiast wyższe wartości HDL zmniejszają je.

3.2 Estymacja modelu probitowego

Model probitowy to jedna z podstawowych metod analizy zmiennej zależnej typu zero-jedynkowego, podobnie jak model logitowy. W modelu tym zakłada się, że prawdopodobieństwo zajścia danego zdarzenia (wystąpienia zespołu metabolicznego) jest funkcją skumulowanego rozkładu normalnego. Oznacza to, że związek między zmienną zależną a zestawem predyktorów opisuje się jako:

$$p = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

gdzie Φ oznacza dystrybuantę standardowego rozkładu normalnego.

Analizę rozpoczęto od estymacji modelu probitowego z pełnym zestawem 11 zmiennych objaśniających. Model ten stanowi punkt wyjścia do dalszej analizy porównawczej z modelem logitowym oraz do ewentualnej selekcji istotnych predyktorów.

Test ilorazu wiarygodności dla pełnego modelu probitowego (z 11 zmiennymi objaśniającymi) wykazuje istotną poprawę dopasowania względem modelu zerowego (zawierającego tylko wyraz wolny). Dla statystyki χ^2, p-wartość jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, co pozwala jednoznacznie odrzucić hipotezę zerową. Oznacza to, że przynajmniej część zmiennych objaśniających istotnie wpływa na prawdopodobieństwo wystąpienia zespołu metabolicznego, a model probitowy z tym zestawem predyktorów jest statystycznie istotny.

Wskaźniki VIF (GVIF) dla modelu probitowego wskazują, że w modelu nie występuje problem współliniowości. Przeskalowane wartości mieszczą się w zakresie od 1.04 do 1.29, co oznacza niski poziom korelacji między zmiennymi objaśniającymi. Najwyższe wartości odnotowano dla zmiennych `Age` (1.29) i `Sex` (1.24), jednak są one znacznie poniżej wartości uznawanych za niepokojące.

Podsumowanie powyżej opisanego modelu (`probit0`) znajduje się w TABLICY 9. Podobnie jak w przypadku modelu `logit0`, zmienne `Race`, `Income` oraz `UrAlbCr` wydają się być

Tablica 9: Podsumowanie modelu probit0

	Ocena parametru modelu	Błąd std.	Statystyka testowa	p-value
(Intercept)	-17.995	1.286	-13.992	0.000
Age	0.024	0.003	7.452	0.000
SexMale	-0.417	0.103	-4.062	0.000
MaritalMarried	0.120	0.139	0.865	0.387
MaritalSeparated	0.449	0.231	1.944	0.052
MaritalSingle	0.336	0.161	2.087	0.037
MaritalWidowed	-0.340	0.203	-1.676	0.094
RaceBlack	0.081	0.166	0.486	0.627
RaceHispanic	0.084	0.178	0.470	0.638
RaceMexAmerican	0.109	0.184	0.592	0.554
RaceOther	-0.270	0.314	-0.860	0.390
RaceWhite	0.079	0.145	0.547	0.584
Income	0.000	0.000	1.110	0.267
BMI	0.078	0.008	9.933	0.000
UrAlbCr	0.037	0.026	1.436	0.151
UricAcid	0.069	0.035	1.974	0.048
BloodGlucose	1.333	0.165	8.103	0.000
HDL	-0.026	0.004	-6.449	0.000
Triglycerides	0.852	0.072	11.849	0.000

Test ilorazu wiarygodności (p-value): 0 | Wynik VIF: niewspółliniowe | AIC: 1200.439
Źródło: opracowanie własne na podstawie zbioru danych

nieistotne statystycznie, ponieważ ich p-wartości dla statystyki z są wyższe niż najwyższy analizowany poziom istotności ($\alpha = 0.1$). Co więcej w przypadku zmiennej `Marital` również zaobserwowano brak istotności jednej z jej kategorii (`Marital: Married`), co jest zgodne z wynikami wcześniejszych modeli logitowych. W poprzednich analizach usunięcie zmiennej `Marital` skutkowało pogorszeniem wartości AIC, dlatego na tym etapie zdecydowano się ograniczyć model jedynie o zmienne demograficzne, które okazały się być całkowicie nieistotne statystycznie (`Race`, `Income`). Dodatkowo, w przeciwieństwie do modeli logitowych, podjęto decyzję o nieuwzględnieniu zmiennej `UrAlbCr`, z uwagi na fakt, że zmienna ta stała się statystycznie istotna dopiero po usunięciu ze zmiennych egzogenicznych zmiennej `Marital` (która zostanie uwzględniona). Takie podejście pozwala na uproszczenie modelu bez znaczącej utraty jakości dopasowania, zachowując jednocześnie istotne informacyjnie zmienne i minimalizując ryzyko nadmiernego dopasowania.

Test porównujący model `probit1` z modelem zerowym (zawierającym jedynie wyraz wolny) wskazuje na bardzo silną istotność statystyczną modelu pełnego. Odrzucono hipotezę zerową, co oznacza, że wprowadzone zmienne w istotny sposób poprawiają dopasowanie modelu.

Wartości wskaźnika VIF dla wszystkich predyktorów w modelu `probit1` mieszczą się

Tablica 10: Podsumowanie modelu probit1

	Ocena parametru modelu	Błąd std.	Statystyka testowa	p-value
(Intercept)	-17.866	1.254	-14.243	0.000
Age	0.024	0.003	7.981	0.000
SexMale	-0.434	0.101	-4.292	0.000
MaritalMarried	0.138	0.134	1.028	0.304
MaritalSeparated	0.478	0.228	2.098	0.036
MaritalSingle	0.317	0.160	1.984	0.047
MaritalWidowed	-0.346	0.202	-1.708	0.088
BMI	0.078	0.007	10.493	0.000
UricAcid	0.074	0.034	2.166	0.030
BloodGlucose	1.348	0.160	8.409	0.000
HDL	-0.026	0.004	-6.483	0.000
Triglycerides	0.847	0.070	12.101	0.000

Test ilorazu wiarygodności (p-value): 0 | Wynik VIF: niewspółliniowe | AIC: 1191.484
Źródło: opracowanie własne na podstawie zbioru danych

w przedziale 1.09-1.53, co oznacza brak istotnego problemu współliniowości. Szczególnie niska wartość przeskalowanego GVIF dla zmiennej `Marital` (1.04) potwierdza, że mimo wielokategoryczności, zmienna ta nie wprowadza nadmiernej redundancji.

W modelu `probit1` uwzględniono osiem zmiennych objaśniających. Wyniki estymacji, przedstawione w Tablicy 10, wskazują, że wiek, płeć, BMI, poziom glukozy, HDL oraz trójglicerydy są istotnymi predyktorami prawdopodobieństwa wystąpienia zespołu metabolicznego ($p < 0.001$). `UricAcid` osiąga poziom istotności granicznej ($p \approx 0.03$), a więc również może mieć duży wpływ na zmienną zależną. W przypadku zmiennej `Marital`, istotności nadal nie wykazuje tylko jeden poziom `Marital:Married`. Pozostałe warianty są istotne statystycznie na poziomach 0.05 i 0.1.

3.3 Estymacja modelu logitowego z interakcjami

Model z interakcją to rozszerzenie klasycznego modelu regresji, w którym uwzględniony jest nie tylko bezpośredni wpływ zmiennych niezależnych na zmienną zależną, ale również ich współzależność. Interakcja oznacza, że efekt jednej zmiennej może zależeć od wartości drugiej zmiennej. Model z jedną interakcją między zmiennymi X_1 i X_2 można wyrazić wzorem:

$$g(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + ...$$

gdzie g to funkcja wiążąca (w poniższej analizie jest wykorzystywana funkcja logit dla regresji logistycznej). Współczynnik β_3 to współczynnik interakcji, który informuje, jak zmienia się efekt X_1 w zależności od X_2.

W celu utworzenia modelu z interakcjami dokonano przekształcenia trzech zmiennych liczbowych na zmienne dychotomiczne. Wskaźnik masy ciała (BMI) podzielono według klasyfikacji [4] na dwie kategorie: „otyłość" (BMI $\geq$ 30) oraz „brak otyłości" (BMI $<$ 30). Poziom cholesterolu HDL również poddano kategoryzacji na podstawie wartości progowych, różniących się dla kobiet i mężczyzn. Dla kobiet wartości HDL poniżej 45 mg/dl oznaczono jako „zły", a równe lub powyżej tej wartości jako „dobry". Dla mężczyzn próg wynosił 40 mg/dl [7]. Dodatkowo, stężenie trójglicerydów podzielono na dwie kategorie: „prawidłowe" oraz „podwyższone", na podstawie ustalonych norm laboratoryjnych [8]. Z uwagi na fakt, że zmienna ta została zlogarytmowana, normy graniczne również zostały przekształcone poprzez logarytmowanie, aby zachować spójność interpretacyjną. Za grupę referencyjną przyjęto osoby o dobrym poziomie cholesterolu HDL, trójglicerydach mieszczących się w ustalonych normach, które nie znajdują się w gronie otyłych (jest to grupa najmniejszego ryzyka w badanej populacji).

Tablica 11: Podsumowanie modelu int1

	Ocena parametru modelu	Błąd std.	Statytstyka testowa	p-value
(Intercept)	-60.707	8.766	-6.925	0
Age	0.642	0.137	4.678	0
BloodGlucose	8.423	1.313	6.415	0
HDLzły	1.823	0.193	9.462	0
Triglyceridespodwyższo	2.550	0.183	13.910	0
BMIotyłość	1.859	0.159	11.681	0
Age:BloodGlucose	-0.091	0.021	-4.423	0

Test ilorazu wiarygodności (p-value): 0 | AIC: 1149.905

Źródło: opracowanie własne na podstawie zbioru danych

Wyniki modelu regresji logistycznej przedstawione w Tablicy 11 wykazały, że wszystkie uwzględnione zmienne, w tym interakcja między wiekiem a poziomem glukozy we krwi, są istotne statystycznie ($p < 0.001$), co świadczy o ich silnym związku z prawdopodobieństwem wystąpienia zespołu metabolicznego. Istotność całego modelu została potwierdzona za pomocą testu ilorazu wiarygodności, który wykazał bardzo niską wartość p, co oznacza, że model z uwzględnionymi predyktorami znacząco lepiej dopasowuje się do danych niż model zerowy. Dodatkowo, wartość kryterium informacyjnego Akaike (AIC) wyniosła 1149.9 i była najniższa spośród wszystkich dotychczas estymowanych modeli. Niska wartość AIC wskazuje na lepsze dopasowanie modelu przy jednoczesnym uwzględnieniu jego złożoności, co jest sugestią, że jest to najbardziej optymalny model spośród rozważanych.

4 Ocena modeli i interpretacja modelu wynikowego

Wybór najlepszego modelu statystycznego jest kluczowym etapem analizy, ponieważ wpływa bezpośrednio na jakość wnioskowania oraz trafność prognoz. Ocena modeli pozwala określić, który z estymowanych wariantów najlepiej odzwierciedla zależności występujące w danych oraz zapewnia równowagę między dopasowaniem a prostotą modelu. Ocena modeli została przeprowadzona dwuetapowo. W pierwszej kolejności analizie poddane zostały miary dopasowania modeli. Następnie oceniona została jakość predykcji przy użyciu odpowiednich wskaźników klasyfikacyjnych. Celem tego etapu jest wybór modelu wynikowego, który najlepiej oddaje struktury obecne w danych i może być wiarygodnie stosowany w dalszej interpretacji.

4.1 Ocena dopasowania modeli

Na potrzeby porównania oszacowanych modeli regresji wykorzystano wybrane miary dopasowania: kryterium informacyjne AIC (Akaike Information Criterion), kryterium BIC (Bayesian Information Criterion), a także dwie miary pseudo-R^2: współczynnik McFaddena oraz współczynnik Cragga-Uhlera.

Kryteria AIC i BIC służą do porównywania jakości dopasowania modeli uwzględniających różną liczbę parametrów. Miara AIC uwzględnia zarówno dopasowanie modelu, jak i jego złożoność. Niższa wartość AIC wskazuje na lepszy kompromis między tymi dwoma aspektami. Kryterium BIC bardziej penalizuje złożone modele (preferuje modele prostsze przy zbliżonym dopasowaniu). Współczynnik pseudo-R^2 McFaddena opiera się na logarytmie funkcji wiarygodności i pełni analogiczną funkcję jak klasyczne R^2, przy czym wartości powyżej 0.2 uznawane są za satysfakcjonujące, a powyżej 0.4 za dobre. Z kolei współczynnik Cragga-Uhlera stanowi przeskalowaną wersję pseudo-R^2, mieszczącą się w przedziale od 0 do 1, co ułatwia jego interpretację. Wyniki ocen dopasowania wyestymowanych modeli przedstawione zostały w Tablicy 12.

Tablica 12: Wyniki ocen modeli

	Kryterium AIC	Kryterium BIC	R^2 McFaddena	R^2 Cragga-Uhlera
logit0	1185.126	1288.242	0.4679	0.6244
logit1	1177.611	1248.164	0.4658	0.6224
logit2	1184.911	1233.755	0.4587	0.6154
probit0	1200.439	1303.554	0.4608	0.6174
probit1	1191.484	1256.609	0.4584	0.6151
int1	1149.905	1187.895	0.4731	0.6294

Źródło: opracowanie własne na podstawie zbioru danych

Kierując się kryterium Akaike uznanym za najlepszy model został model z interakcją `int1` z najniższą wartością tego kryterium. Zwykłe modele logitowe wykazały mniejszą dobroć

dopasowania według kryterium AIC, przy czym model `logit1` miał spośród nich najlepszy wynik. Obydwa modele probitowe natomiast, wykazały gorsze dopasowanie według tego kryterium niż pozostałe modele, jednak model `probit1` wypadł lepiej w stosunku do modelu `probit0`.

Rozważając Bayesowskie kryterium Schwarza, najlepszym modelem również okazał się być model z interakcją (`int1`), który był także modelem z najmniejszą liczbą zmiennych egzogenicznych. Modele logitowe bez interakcji wykazały niewiele gorsze dopasowanie, przy czym model `logit2` był spośród nich najlepszy. Modele probitowe dla tego kryterium również wypadły gorzej niż pozostałe modele. Jedynym wyjątkiem jest model `probit1`, dla którego wartość BIC była minimalnie niższa niż dla modelu `logit0` (prawdopodobnie ze względu na mniejszą liczbę zmiennych niezależnych). W pozostałych przypadkach model ten jednak wykazał gorsze dopasowanie.

Według miar pseudo-R^2 najlepiej dopasowanym modelem również jest model `int1`. W pozostałych modelach, wraz ze zmniejszaniem ilości zmiennych w modelach można zaobserwować spadek wartości tych miar. Zarówno dla modeli logitowych jak i probitowych, najlepsze wyniki prezentują modele '`0`', czyli modele zawierające wszystkie zmienne brane pod uwagę do budowy modeli.

Kierując się kryteriami dopasowania oraz wartościami pseudo-R^2, przy jednoczesnym zachowaniu istotności zmiennych, za najlepszy model uznano model `int1`. Model ten zbudowany został przy użyciu najmniejszej liczby (statystycznie istotnych) zmiennych niezależnych. Zarówno według kryterium AIC, jak i BIC, które bierze pod uwagę liczbę zmiennych użytych do budowy modelu, był to model najlepszy. Obydwie miary pseudo-R^2, także wskazały ten model jako najlepiej dopasowany. Ponadto został on zbudowany przy użyciu najmniejszej liczby zmiennych niezależnych, a wszystkie z nich były statystycznie istotne na poziomie istotności $\alpha = 0.001$.

4.2 Ocena jakości predykcji

W celu weryfikacji zdolności predykcyjnych modeli, dane zostały uprzednio podzielone na zbiór uczący oraz zbiór testowy. Estymacja modeli została przeprowadzona na zbiorze uczącym, natomiast ocena trafności prognozowanych klasyfikacji zostanie przeprowadzona zarówno na zbiorze uczącym, jak i zbiorze testowym.

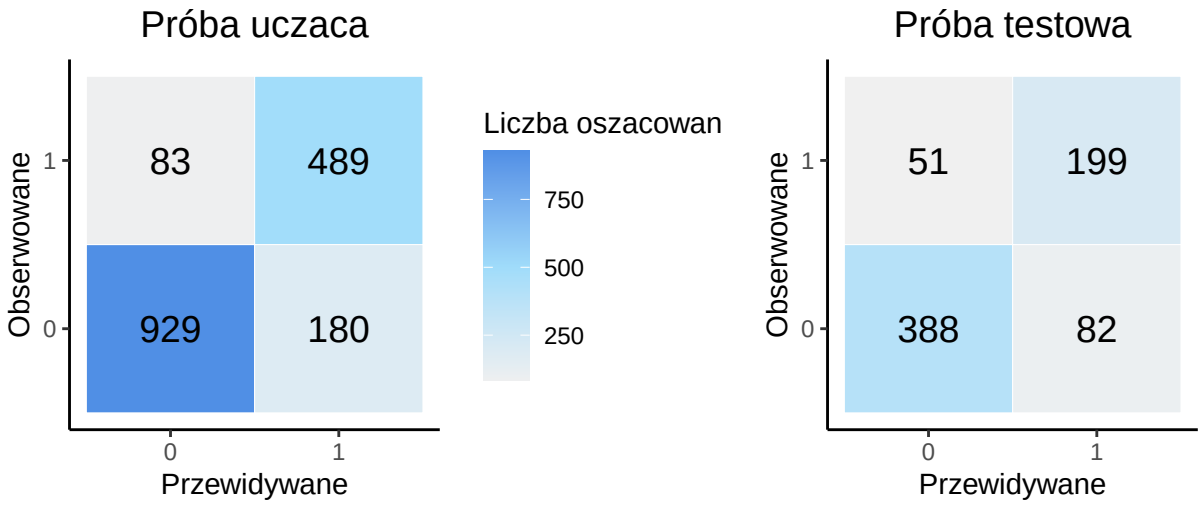
Punkt odcięcia (wartość progową) wybrano w następujący sposób:

$$p^* = \frac{n_{1\bullet}}{N}$$

gdzie $n_{1\bullet}$ to łączna liczba zaobserwowanych 1, a N to liczebność zbioru.

Tablica trafności przedstawia liczbę poprawnych i błędnych predykcji dokonanych przez model, zestawionych z rzeczywistymi klasami.

Wykres 9: Porównanie tablic trafności dla modelu logit1



Źródło: opracowanie własne na podstawie zbioru danych

Wyniki oceny jakości predykcji oparte na tablicy trafności (Wykres 9) dla wybranego punktu odcięcia p^* dla modelu `logit1` przedstawiono osobno dla zbioru uczącego oraz testowego (Tablica 13). W przypadku zbioru uczącego, model osiągnął dokładność klasyfikacji (ACC) na poziomie 84.35%, a błąd klasyfikacji (ER) wyniósł 15.65%. Czułość modelu (SENS), czyli zdolność do wykrywania przypadków pozytywnych (obecności zespołu metabolicznego) wśród wszystkich zaobserowanych pozytywnych przypadków, osiągnęła wartość 85.49%. Natomiast swoistość (SPEC), określająca skuteczność w identyfikacji przypadków negatywnych wśród wszystkich zaobserwowanych negatywnych przypadków, wyniosła 83.77%. Dodatnia zdolność predykcyjna (PPV), czyli zdolność do trafnego wykrywania przypadków pozytywnych wśród wszystkich prognozowanych przypadków pozytywnych wyniosła 73.09%. Ujemna zdolność predykcyjna (NPV), określająca zdolność do identyfikacji przypadków negatywnych wśród wszystkich prognozowanych, negatywnych przypadków, wyniosła 91.80%. Model radzi sobie zatem lepiej z przewidywaniem przypadków niewystąpienia zespołu metabolicznego niż z przewidywaniem jego wystąpienia (w zbiorze uczącym).

Tablica 13: Wyniki miar predykcyjnych dla modelu logit1

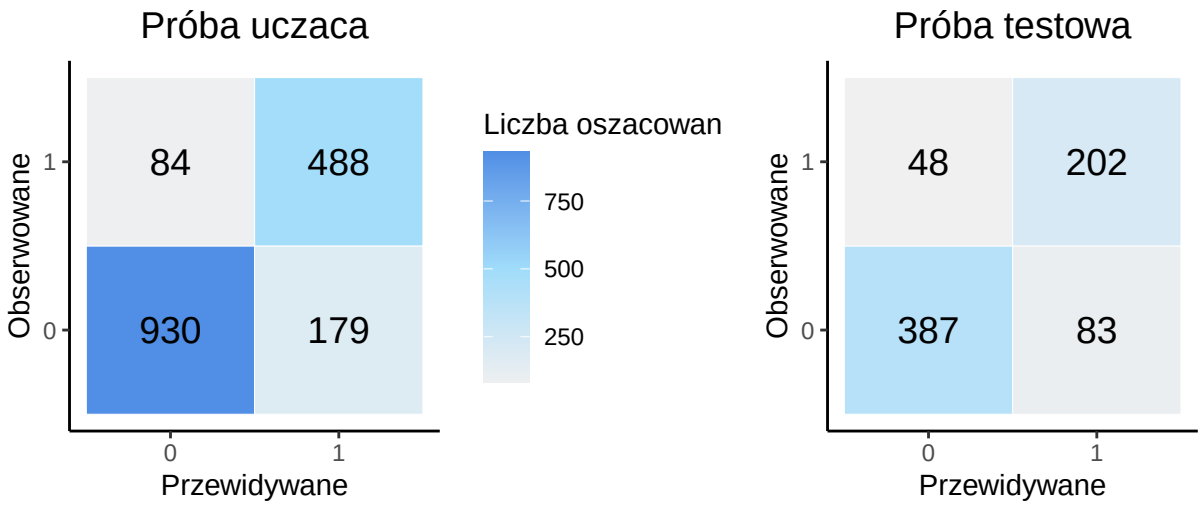
	ACC	ER	SENS	SPEC	PPV	NPV
logit1_uczacy	0.8435	0.1565	0.8549	0.8377	0.7309	0.9180
logit1_testowy	0.8153	0.1847	0.7960	0.8255	0.7082	0.8838

Źródło: opracowanie własne na podstawie zbioru danych

Dla zbioru testowego, model logitowy wykazał nieco niższą skuteczność, co jest typowe w kontekście oceny modelu na danych niezależnych. Dokładność klasyfikacji spadła do 81.53%, a błąd klasyfikacji wzrósł do 18.47%, co różni się od wyników dla zbioru uczącego o 2.82%. Czułość wyniosła 79.60%, co wskazuje na nieco słabszą zdolność wykrywania

przypadków pozytywnych w porównaniu do zbioru uczącego (różnica 5.89%). Swoistość pozostała na wysokim poziomie i wyniosła 82.55%. Zarówno dodatnia, jak i ujemna zdolność predykcyjna minimalnie się obniżyły i wynoszą kolejno 70.82% i 88.38%. Model prezentuje zatem dobrą jakość predykcji. Różnice w wynikach pomiędzy zbiorem uczącym a testowym są niewielkie, co sugeruje, że model generalizuje dobrze i nie jest przeuczony. Jedynym wyjątkiem jest dodatnia zdolność predykcyjna, która w stosunku do innych miar jest o wiele niższa i utrzymuje się na poziomie ok. 70%, co nie jest najlepszym wynikiem.

Wykres 10: Porównanie tablic trafności dla modelu logit2



Źródło: opracowanie własne na podstawie zbioru danych

Tablice trafności dla modelu `logit2` przedstawione zostały na Wykresie 10. Można zauważyć, że zdolność modelu do trafnego przewidywania przypadków zarówno występowania, jak i niewystępowania zespołu metabolicznego dla zbioru uczącego jest bardzo zbliżona do modelu `logit1`. Dokładność klasyfikacji (ACC) modelu `logit2` dla tego zbioru wyniosła 84.35%, natomiast błąd klasyfikacji (ER) – 15.65%, co jest takim samym wynikiem jak dla modelu `logit1` dla zbioru uczącego. Jego czułość (SENS) wyniosła natomiast 85.31%, a swoistość (SPEC) – 83.86%. Wyniki te także są bardzo podobne do wyników modelu `logit1` dla tego samego zbioru. Dodatnia zdolność predykcyjna (PPV) osiągnęła wartość 73.16%, a ujemna osiągnęła wartość wyższą, równą 91.72%. Wyniki tych miar także są zbliżone do poprzedniego modelu.

W przypadku zbioru testowego, miary dla tego modelu (Tablica 14) także nieco się pogorszyły, jednak nadal pozostały wysokie. Model trafnie przewidział o jeden mniej przypadek niewystąpienia zespołu metabolicznego niż model `logit1` dla tego zbioru. Dobrze przewidział jednak wystąpienie zespołu metablicznego w większej liczbie przypadków niż poprzedni model. Dokładność jego klasyfikacji spadła do 81.81% (o 2.54% w stosunku do zbioru uczącego), natomiast błąd klasyfikacji wzrósł do 18.19%. Czułość modelu dla zbioru testowego wyniosła 80.8%, a zatem spadła o 4.51% w porównaniu z czułością dla

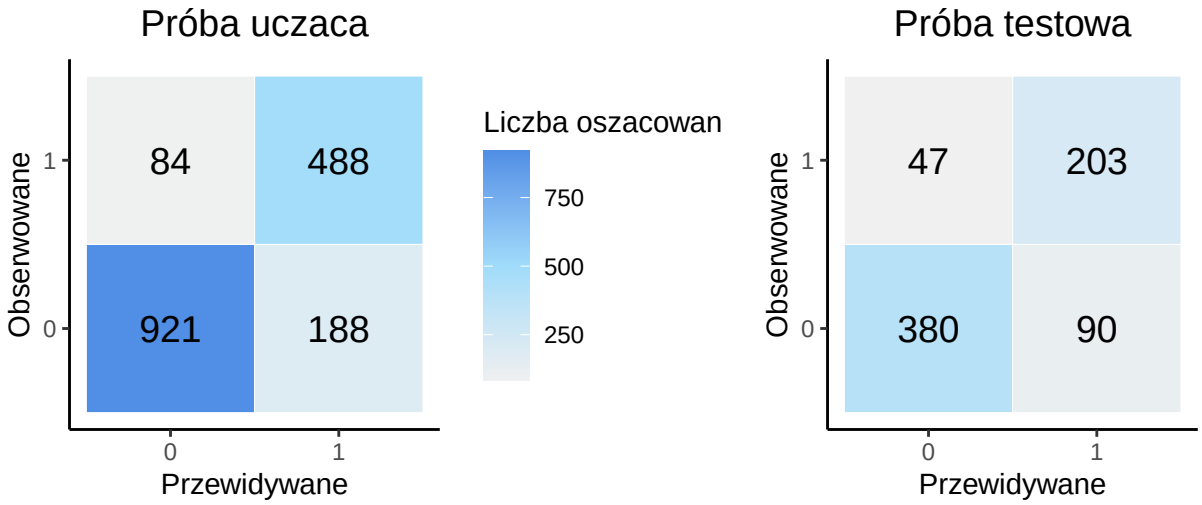
zbioru uczącego. Dla porównania, czułość dla modelu `logit1` spadła o 5.89% dla zbioru testowego w stosunku do uczącego. Swoistość również obniżyła się do 82.34%, czyli o 1.52% w stosunku do zbioru uczącego. W przypadku modelu `logit1` swoistość spadła o 1.22% dla zbioru testowego. Dodatnia i ujemna zdolność predykcyjna wyniosły kolejno 70.88% i 88.97%. Jest to także wynik gorszy niż dla zbioru uczącego (o 2.28% dla PPV i 2.75% dla NPV). Mimo spadków dla zbioru testowego, model nadal wykazuje bardzo dobrą jakość predykcji. Różnice między powyższymi miarami pomiędzy modelami `logit1` oraz `logit2` są minimalne, jednak w większości wyższe wyniki wykazuje model `logit2`. Podobnie jak w przypadku poprzedniego modelu, dodatnia zdolność predykcyjna także nie prezentuje się bardzo dobrze.

Tablica 14: Wyniki miar predykcyjnych dla modelu logit2

	ACC	ER	SENS	SPEC	PPV	NPV
logit2_uczacy	0.8435	0.1565	0.8531	0.8386	0.7316	0.9172
logit2_testowy	0.8181	0.1819	0.8080	0.8234	0.7088	0.8897

Źródło: opracowanie własne na podstawie zbioru danych

Wykres 11: Porównanie tablic trafności dla modelu probit1



Źródło: opracowanie własne na podstawie zbioru danych

W przypadku modelu `probit1` trafność przewidywań niewystępienia zespołu metabolicznego w zbiorze uczącym jest gorsza w porównaniu do modeli logitowych (Wykres 11). Model ten przewidział minimalnie mniej trafnych przypadków. Przewidział on także o 1 trafny przypadek wystąpienia zespołu metoblicznego mniej niż model `logit1`, ale tyle samo co model `logit2` (dla zbioru uczącego). Dla tego zbioru dokładność predykcji (ACC) wyniosła 83.82%, natomiast błąd klasyfikacji (ER) – 16.18% (Tablica 15). Wyniki różnią się o

0.53% w stosunku do modeli logitowych. Czułość (SENS) modelu wyniosła natomiast 85.31%, co jest wynikiem niższym o 0.18% niż dla modelu `logit1`, natomiast takim samym jak dla modelu `logit2`. Jego swoistość wyniosła 83.05%, a zatem jest niższa niż dla poprzednich modeli o 0.72% i 0.81%. Dodatnia i ujemna zdolność predykcyjna utrzymuje się na podobnym poziomie co modele logitowe, jednak wartości dla modelu `probit1` są minimalnie niższe. Mimo niższych wyników w porównaniu do modeli logitowych, model `probit1` nadal wykazuje dobrą jakość predykcji dla zbioru uczącego.

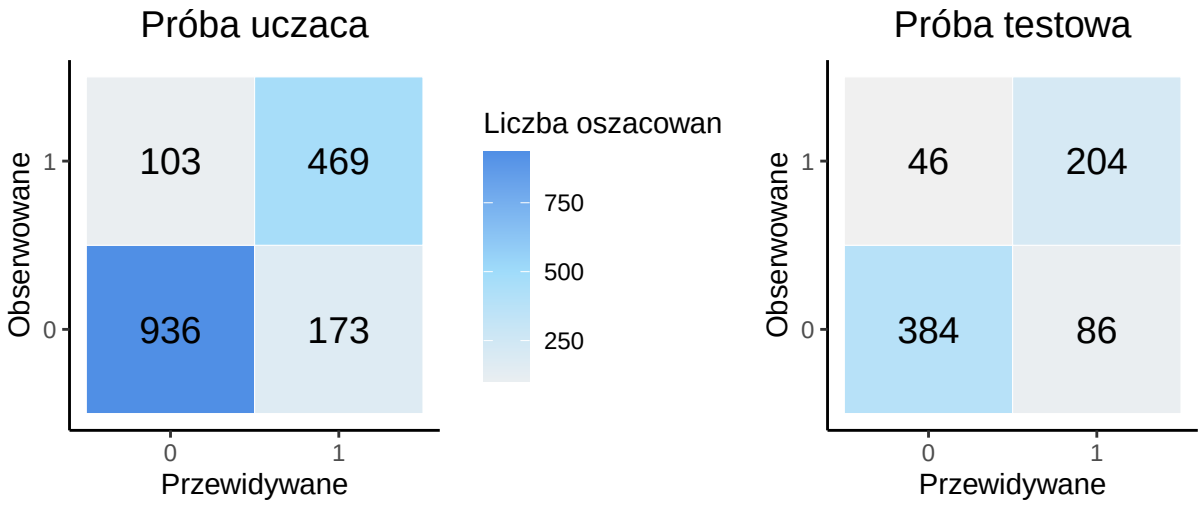
Dla zbioru testowego, model `probit1` przewidział mniej trafnych przypadków niewystąpienia zespołu metoblicznego niż obydwa modele logitowe dla zbioru testowego. Przewidział natomiast trafnie więcej przypadków pozytywnych niż modele `logit1` i `logit2`. Dokładność klasyfikacji dla tego zbioru wyniosła 80.97%, natomiast błąd wyniósł 19.03%. Wyniki te różnią się o 2.85% w stosunku do wartości tych miar dla zbioru uczącego. Czułość, czyli zdolność wykrywania przypadków pozytywnych wyniosła 81.2% (o 4.11% mniej niż dla zbioru uczącego), natomiast swoistość – 80.85% (2.2% mniej niż dla zbioru uczącego). Dodatnia zdolność predykcyjna tego modelu spadła do 69.28%, czyli o 2.91% w stosunku do PPV dla zbioru uczącego. Ujemna zdolność predykcyjna obniżyła się do 88.99% (o 2.65% w stosunku do NPV dla zbioru uczącego). Wartości miar są minimalnie niższe niż dla obydwu modeli logitowych dla tego samego zbioru, z wyjątkiem czułości i ujemnej zdolności predykcyjnej. Mimo tego, model nadal wykazuje dobrą jakość predykcji dla obydwu zbiorów.

Tablica 15: Wyniki miar predykcyjnych dla modelu probit1

	ACC	ER	SENS	SPEC	PPV	NPV
probit1_uczacy	0.8382	0.1618	0.8531	0.8305	0.7219	0.9164
probit1_testowy	0.8097	0.1903	0.8120	0.8085	0.6928	0.8899

Źródło: opracowanie własne na podstawie zbioru danych

Wykres 12: Porównanie tablic trafności dla modelu int1



Źródło: opracowanie własne na podstawie zbioru danych

Tablice trafności dla modelu z interakcją (`int1`) zaprezentowane zostały na Wykresie 12. Dla zbioru uczącego, model lepiej radził sobie z trafnym przewidywaniem przypadków negatywnych niż opisane wcześniej modele. Dobrze przewidział jednak najmniej przypadków pozytywnych spośród wszystkich modeli. Dokładność (ACC) modelu wyniosła 83.58%, natomiast błąd klasyfikacji (ER) równy był 16.42% (Tablica 16). W porównaniu do poprzednich modeli są to wyniki najgorsze. Jego czułość (SENS) wyniosła 81.99%, co także jest najniższym wynikiem. Swoistość (SPEC) wyniosła 84.4%, a dodatnia (PPV) i ujemna (NPV) zdolność predykcyjna, kolejno 73.05% i 90.09%. Wyniki te także w większości są wynikami gorszymi w porównaniu do innych modeli.

Dla zbioru testowego, model trafnie przewidział więcej przypadków pozytywnych niż którykolwiek z modeli, a jednocześnie wykrył mniej przypadków negatywnych. Osiągnął dokładność na poziomie 81.67%, a błąd klasyfikacji wyniósł 18.33%, co stanowi różnicę jedynie o 1.91% względem zbioru uczącego. Minimalnie lepszą dokładność niż ten model uzyskał tylko model `logit2`, co czyni int1 drugim najlepiej dopasowanym modelem pod względem ogólnej trafności klasyfikacji. Model `int1` wyróżniał się najwyższą czułością spośród wszystkich ocenianych modeli – osiągnęła ona poziom 81.6%. Wartość ta oznacza, że model skutecznie identyfikuje osoby chore, co jest kluczowe w analizie przypadków zespołu metabolicznego, gdzie priorytetem jest wychwycenie jak największej liczby przypadków. Wysoka czułość odbywa się kosztem nieco niższej swoistości, która wyniosła 81.7% – mniej niż w przypadku modeli bez interakcji, lecz nadal na bardzo dobrym poziomie. Miary te różnią się dla zbioru uczącego, kolejno o 0.39% i 0.27%. Dodatnia zdolność predykcyjna spadła do 70.34%, czyli o 2.71% w stosunku do zbioru uczącego, co oznacza, że większy odsetek przewidzianych jako chorzy faktycznie choruje, choć część wyników pozytywnych jest fałszywa. Ujemna zdolność predykcyjna wyniosła 89.30%, co oznacza spadek o 0.79% względem zbioru uczącego. Wszystkie miary wykazały niewielkie różnice między zbiorem uczącym a testowym, co świadczy o dobrej stabilności i zdolności generalizacji modelu `int1`.

Tablica 16: Wyniki miar predykcyjnych dla modelu int1

	ACC	ER	SENS	SPEC	PPV	NPV
int1_uczacy	0.8358	0.1642	0.8199	0.844	0.7305	0.9009
int1_testowy	0.8167	0.1833	0.8160	0.817	0.7034	0.8930

Źródło: opracowanie własne na podstawie zbioru danych

Podsumowując, model `int1` uzyskał drugą najwyższą dokładność wśród wszystkich porównywanych modeli oraz najwyższą czułość, co w kontekście diagnostyki ma szczególne znaczenie. Ze względu na istotność kliniczną nieprawidłowego rozpoznania osób chorych, bardziej pożądana jest nadmierna klasyfikacja przypadków jako pozytywnych niż ich pominięcie. Taki kompromis wiąże się z nieco niższą wartością PPV, jednak w analizach epidemiologicznych i diagnostycznych wykrycie możliwie jak największej liczby przypadków pozostaje kluczowe.

Krzywa ROC obrazuje zależność między czułością (Sensitivity lub inaczej True Positive Rate) a fałszywym wskaźnikiem pozytywności (False Positive Rate), który mierzy udział błędnie sklasyfikowanych przypadków negatywnych wśród wszystkich zaobserwowanych, negatywnych przypadków. Wykres ilustruje zatem skuteczność modelu dla każdej wartości progowej p^*. Wskaźnik AUC, czyli pole powierzchni pod krzywą ROC, mierzy zatem ogólną zdolność modelu do rozróżniania przypadków pozytywnych i negatywnych.

Tablica 17: Pole powierzchni pod krzywą ROC (AUC) dla zbioru uczącego

logit1_uczacy	0.9205
logit2_uczacy	0.9186
probit1_uczacy	0.9194
int1_uczacy	0.9222

Źródło: opracowanie własne na podstawie zbioru danych

Wartość wskaźnika AUC dla zbioru uczącego jest wysoka dla każdego z modeli (powyżej 0.9), co wskazuje na bardzo dobrą jakość predykcji (Tablica 17). Wskaźniki dla modeli `logit2` i `probit1` utrzymują się lekko poniżej poziomu 0.92, natomiast dla modeli `logit1` i modelu z interakcją (`int1`) są nieznacznie wyższe niż 0.92, co sugeruje lepszą zdolność rozróżniania przypadków pozytywnych i negatywnych.

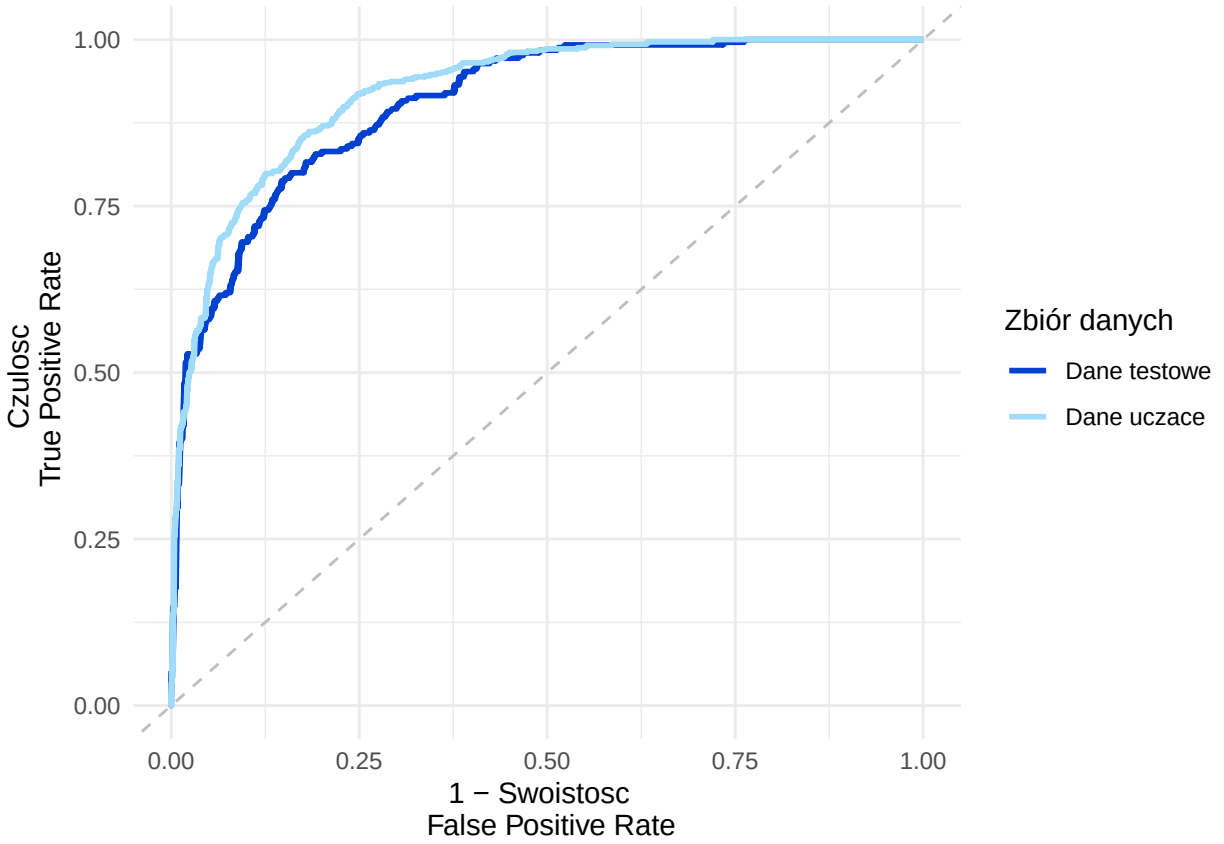
Tablica 18: Pole powierzchni pod krzywą ROC (AUC) dla zbioru testowego

logit1_testowy	0.8871
logit2_testowy	0.8911
probit1_testowy	0.8862
int1_testowy	0.9045

Źródło: opracowanie własne na podstawie zbioru danych

Dla zbioru testowego wartości AUC również utrzymują się na wysokim poziomie (Tablica 18), jednak model `int1` osiągnął najwyższy wynik spośród wszystkich – 0.9045. Pozostałe modele uzyskały wartości poniżej 0.9. Różnice te, choć niewielkie, wskazują na lepszą zdolność predykcyjną modelu z interakcją na nowym, nieuczonym zbiorze danych. Według kryterium AUC, model `int1` cechuje się najlepszym dopasowaniem do danych testowych oraz najwyższą jakością predykcji, co dodatkowo wzmacnia jego pozycję jako modelu wynikowego w dalszej analizie. W związku z tym zdecydowano się na przedstawienie jego krzywej ROC na Wykresie 13.

Wykres 13: Krzywa ROC dla modelu int1



Krzywa ROC dla tego modelu przebiega blisko lewego górnego rogu wykresu, co świadczy o bardzo dobrej zdolności klasyfikacyjnej modelu. Obszar pod krzywą (AUC = 0.9045) potwierdza wysoką jakość predykcji - model dobrze rozróżnia między przypadkami pozytywnymi (zespół metaboliczny) a negatywnymi (brak zespołu metabolicznego). Kształt krzywej wskazuje również na korzystny kompromis między czułością a swoistością, co jest szczególnie istotne w kontekście medycznym, gdzie priorytetem często jest identyfikacja jak największej liczby rzeczywiście chorych osób.

4.3 Interpretacja modelu wynikowego

We wcześniejszej części dokonano wyboru modelu logitowego z interakcją `int1` jako modelu wynikowego ze względu na jego dobre dopasowanie i zrównoważone właściwości predykcyjne. Model ten został oszacowany na podstawie danych uczących i zawiera 6 zmiennych objaśniających, w tym jedną interakcję między zmiennymi `Age` i `BloodGlucose`. Interakcja jest traktowana jako osobny składnik modelu, więc łącznie otrzymano 5 zmiennych głównych i 1 składnik interakcyjny. Jego postać, wyrażona równaniem logitowym, przedstawia się następująco:

$$logit(p) = ln(\frac{p}{1-p}) = -60.7 + 0.64 \cdot Age - 8.42 \cdot BloodGlucose + 1.82 \cdot HDL_{zły}$$
$$+2.55 \cdot Triglycerides_{Podwyższone} + 1.86 \cdot BMI_{Otyłość} - 0.09 \cdot Age \cdot BloodGlucose$$

Współczynniki modelu logitowego należy interpretować jako wpływ danej zmiennej na logarytm ilorazu szans (logit) wystąpienia zespołu metabolicznego. Dla zmiennych ilościowych, takich jak `Age` (wiek) i `BloodGlucose` (poziom glukozy we krwi), oznacza to zmianę logit(p) o daną wielkość przy jednostkowym wzroście danej zmiennej, przy pozostałych zmiennych na stałym poziomie. W przypadku zmiennych jakościowych – zakodowanych jako zmienne binarne – współczynniki określają różnicę w logit(p) względem grupy referencyjnej.

W modelu `int` oszacowano sześć składników: pięć efektów głównych oraz jeden składnik interakcyjny. Grupy referencyjne to osoby z prawidłowym poziomem HDL, prawidłowym poziomem trójglicerydów oraz bez otyłości (prawidłowe BMI), dlatego wartości dodatnie współczynników oznaczają wzrost ryzyka wystąpienia zespołu metabolicznego w stosunku do tych grup (Tablica 11).

Dla każdego współczynnika z modelu logitowego obliczono iloraz szans (OR) jako:

$$OR = e^{\beta_i}$$

Wyraz wolny `Intercept` nie ma praktycznej interpretacji – odpowiada teoretycznej wartości logitu (logarytmu ilorazu szans) dla osoby, która: ma 0 lat (grupa noworodków), poziom glukozy wynosi 0 mmol/l, ma prawidłowy poziom HDL i trójglicerydów oraz nie ma otyłości. Czyli interpretacja tego parametru dla osoby referencyjnej we wszystkich zmiennych ma charakter czysto matematyczny (nie praktyczny), gdyż jest to fizjologicznie nierealistyczne, a do tego dane nie obejmowały grupy noworodków. Przy założeniu stałego poziomu glukozy, każdemu dodatkowemu roku życia odpowiada wzrost szans wystąpienia zespołu metabolicznego o 90.06%. Jednak ze względu na obecność interakcji z poziomem glukozy, efekt ten obowiązuje wyłącznie dla poziomu glukozy równemu 0 mmol/l – co w praktyce nie występuje. Zatem interpretacja efektu wieku wymaga uwzględnienia poziomu glukozy. Dla niemowląt (osoby w wieku 0 lat – czysto teoretycznie), każdy dodatkowy mmol/l glukozy zwiększa szansę zachorowania ponad 45-krotnie. W praktyce interpretacja tego efektu musi uwzględniać interakcję z wiekiem.

Osoby z nieprawidłowym HDL mają ponad 6-krotnie wyższe szanse wystąpienia zespołu metabolicznego niż osoby z prawidłowym HDL, przy pozostałych zmiennych utrzymanych na stałym poziomie, ceteris paribus (osoby: w tym samym wieku, przy tym samym poziomie glukozy, tym samym poziomie trójglicerydów i takim samym BMI). Podwyższone trójglicerydy zwiększają szanse wystąpienia zespołu metabolicznego prawie 13-krotnie, ceteris paribus. Otyłość zwiększa szansę wystąpienia zespołu metabolicznego ponad 6-krotnie, ceteris paribus.

Interakcja między wiekiem a poziomem glukozy jest istotna i posiada ujemny współczynnik (-0.0907), co oznacza, że z każdym dodatkowym rokiem życia wpływ glukozy na logit(p) maleje. Iloraz ilorazu szans wynosi 0.913, a zatem każdy dodatkowy rok życia zmniejsza wpływ poziomu glukozy o około 8.7%. Innymi słowy, ryzyko związane z wysokim poziomem glukozy jest największe u młodych osób, a maleje z wiekiem.

5 Podsumowanie i wnioski

W niniejszym projekcie podjęto próbę opracowania modelu predykcyjnego umożliwiającego identyfikację osób spełniających kryteria rozpoznania zespołu metabolicznego na podstawie danych klinicznych i demograficznych. W analizie wykorzystano publiczny zbiór danych pochodzący z platformy Kaggle. Wstępna analiza wykazała istotne różnice w wybranych zmiennych pomiędzy grupą pacjentów chorych i zdrowych.

Stwierdzono obecność brakujących danych w kilku zmiennych, w szczególności `Marital`, `Income`, `WaistCirc` oraz `BMI`. W związku z tym, zastosowano metodę imputacji wielokrotnej *MICE*, wykorzystując techniki odpowiednie dla danych ilościowych i jakościowych. Na podstawie analizy statystyk zestawów imputowanych danych wybrano jeden z nich do dalszej analizy, ze względu na jego największe podobieństwo do rozkładów oryginalnych zmiennych.

W ramach eksploracyjnej analizy danych przeprowadzono wizualizację rozkładów zmiennych ilościowych w podziale na osoby z zespołem metabolicznym i bez niego. Zidentyfikowano obecność wartości odstających oraz istotne różnice w kształcie rozkładów zmiennych takich jak BMI, obwód talii, poziom glukozy, HDL, kwasu moczowego i trójglicerydów. W celu zbadania nieliniowych zależności między zmiennymi a występowaniem zespołu metabolicznego zastosowano wykresy z liniami *loess* oraz *glm*, a następnie dokonano transformacji logarytmicznej wybranych zmiennych celem poprawy dopasowania regresji logistycznej. Analiza potwierdziła, że przekształcenie logarytmiczne zmiennych o silnie asymetrycznym rozkładzie może znacząco poprawić dopasowanie modelu, co stanowi istotną przesłankę dla dalszych etapów modelowania predykcyjnego.

W przeprowadzonej analizie wykorzystano modele regresji logitowej i probitowej do identyfikacji istotnych czynników wpływających na prawdopodobieństwo wystąpienia zespołu metabolicznego. W każdym z modeli istotność statystyczna predyktorów została potwierdzona testem ilorazu wiarygodności. Selekcja zmiennych oparta została na analizie wartości p oraz uwzględnieniu znaczenia klinicznego wybranych zmiennych, dzięki czemu ostateczne modele łączą podejście statystyczne z merytorycznym. We wszystkich przypadkach poziom współliniowości oceniany za pomocą wskaźników VIF/GVIF mieścił się istotnie poniżej przyjętej granicy, co potwierdza niezależność predyktorów i stabilność estymacji parametrów.

Nastepnie przeprowadzono ocenę jakości dopasowania oraz zdolności predykcyjnych modeli regresji logitowej i probitowej, z naciskiem na wybór optymalnego modelu oraz interpretację jego parametrów. Do porównania modeli wykorzystano kryteria AIC, BIC oraz miary pseudo-R^2. Najlepszym modelem okazał się model z interakcją (`int1`), który charakteryzował się najlepszym kompromisem między dopasowaniem a złożonością oraz zawierał jedynie istotne zmienne. Ocena jakości predykcji przeprowadzona na zbiorach uczącym i testowym wykazała, że wybrane modele zachowują dobrą trafność klasyfikacji i stabilność predykcji, choć naturalnie wyniki na zbiorze testowym są nieco niższe. Modele logitowe generalnie przewyższały modele probitowe pod względem skuteczności klasyfikacji, zwłaszcza w kontekście dodatniej zdolności predykcyjnej, która pozostawała na niższym poziomie. Analiza krzywych ROC i wskaźnika AUC potwierdziła wysoką zdolność rozróżniania przypadków pozytywnych i negatywnych przez wszystkie modele, przy czym

model `int1` uzyskał nieznacznie lepsze wyniki.

Na podstawie uzyskanych wyników oraz oceny jakości modeli dokonano szczegółowej interpretacji współczynników najlepszego modelu – regresji logistycznej z interakcją (`int1`). Model ten uwzględnia zarówno 5 zmiennych głównych: `Age`, `BloodGlucose`, dychotomiczne zmienne `HDL`, `Triglycerides`, `BMI`, jak i interakcję pomiędzy wiekiem a poziomem glukozy. Pomimo braku możliwości bezpośredniej interpretacji wpływu tych dwóch zmiennych oddzielnie (ze względu na efekt interakcyjny), analiza współczynnika interakcji pozwoliła na sformułowanie istotnych wniosków klinicznych.

Stwierdzono, że wpływ poziomu glukozy na ryzyko wystąpienia zespołu metabolicznego maleje wraz z wiekiem – im osoba starsza, tym mniejsze znaczenie predykcyjne glukozy. Innymi słowy, podwyższona glikemia u młodych osób istotnie zwiększa ryzyko rozwoju zespołu metabolicznego, podczas gdy u osób starszych efekt ten jest słabszy. Oznacza to, że glukoza jest szczególnie niebezpiecznym czynnikiem ryzyka u młodszych osób, co ma istotne znaczenie z punktu widzenia profilaktyki.

Wyniki te wskazują na potrzebę ukierunkowania działań prewencyjnych na młodsze grupy wiekowe, u których podwyższony poziom glukozy może prowadzić do znacznie wyższego ryzyka rozwoju zespołu metabolicznego. Edukacja zdrowotna, odpowiednia dieta i aktywność fizyczna powinny być szczególnie promowane wśród młodszych dorosłych, zanim pojawią się nieodwracalne zmiany metaboliczne.

Przeprowadzona analiza umożliwiła nie tylko stworzenie skutecznego modelu predykcyjnego identyfikującego osoby spełniające kryteria rozpoznania zespołu metabolicznego, ale także przyczyniła się do lepszego zrozumienia zależności pomiędzy zmiennymi klinicznymi a ryzykiem jego wystąpienia. Ujawnienie istotnego efektu interakcji pomiędzy wiekiem a poziomem glukozy pozwala podkreślić, że ryzyko metaboliczne należy analizować w sposób zindywidualizowany – uwzględniając wiek pacjenta i jego profil metaboliczny. Wyniki wskazują, że obecnie stosowane kryteria rozpoznawania zespołu metabolicznego, choć przydatne diagnostycznie, mogą nie uwzględniać zmienności ryzyka w zależności od wieku i interakcji pomiędzy czynnikami ryzyka. Opracowany model nie tylko wspiera proces wczesnej diagnozy, ale również stanowi krok w kierunku bardziej precyzyjnych, spersonalizowanych narzędzi analitycznych w profilaktyce chorób cywilizacyjnych.

Bibliografia

1. Kalinowski P, Mianowana M. *Zespół Metaboliczny Cz I: Przegląd Kryteriów Rozpoznania Zespołu Metabolicznego = Metabolic Syndrome Part I: Overview Of Criteria Of Recognition Of Metabolic Syndrome*. Zenodo, 2016.

2. Dobrowolski P, Prejbisz A, Kuryłowicz A, Baska A, Burchardt P, Chlebus K i wsp. *Metabolic syndrome — a new definition and management guidelines*. Arterial Hypertension, 2022; 26(3): 99–121.

3. Saklayen MG. The Global Epidemic of the Metabolic Syndrome. Current Hypertension Reports, 2018; 20(2).

4. Szpital Medicover. *Wskaźnik BMI – stopnie otyłości*, 2018.

5. Poradnik Zdrowie. *Albuminuria – przyczyny, objawy, leczenie*, 2025.

6. Diagnostyka+. *Kwas moczowy – norma, badanie Jak sprawdzić jego poziom?*, 2025.

7. Medicare.pl. *Cholesterol – normy, rodzaje, dobry i zły cholesterol (HDL i LDL)*, 2025.

8. Diagnostyka+. *Wysokie trójglicerydy – co oznaczają i czym skutkują?*, 2025.

9. Marszałek M: *Nowoczesne metody imputacji braków danych – porównanie wybranych metod*. Wroclaw University of Economics; Business; 2023, s. 49–62.

10. van Buuren, Stef. *mice: Multivariate Imputation by Chained Equations in R – Reference Manual*, 2024.