

Justyna Marcinkowska

Rozprawa na stopień doktora nauk medycznych

„Metody statystyczne i  
eksploracji danych (data mining)  
w ocenie występowania omdleń  
w grupie częstoskurczu  
z wąskim zespołem QRS  
(AVNRT i AVRT).”

Promotorzy:

**prof. dr hab. Jerzy Moczko** (Katedra i Zakład Informatyki i  
Statystyki)

oraz

**dr hab. med. Krzysztof Błaszyk** (I Klinika i Katedra Kardiologii)

Katedra i Zakład Informatyki i Statystyki

Uniwersytetu Medycznego im. Karola Marcinkowskiego w  
Poznaniu

Poznań, 2015

*Pragnę podziękować wszystkim, bez których  
niniejsza praca nie mogłaby powstać.*

*Przede wszystkim moim promotorom:*

*prof. dr hab. Jerzemu Moczko dziękuję za  
cenne wskazówki w czasie doktoratu,*

*dr hab. med. Krzysztofowi Błaszykowi  
dziękuję za udostępnienie danych do analizy  
oraz za wszelką pomoc, jakiej udzielił mi w  
czasie dotychczasowej współpracy, za  
wsparcie i dobrą radę.*

*Pragnę podziękować również koleżankom i  
kolegom z pracy: zwłaszcza Basi, która  
wydobyła mnie z odmętów chaosu i zawsze  
służyła dyskusją by wyjaśnić moje  
wątpliwości.*

*Dziękuję za niezłomną wiarę we mnie  
okazywaną mi przez moich rodziców.*

*I wreszcie mojemu mężowi za wsparcie w  
trudnych chwilach oraz wiarę w to, że się uda.*

*Ale najgoręcej dziękuję moim dzieciom: za  
cierpliwość.*

*" Nie odniosłem porażki. Po prostu odkryłem 10 000 błędnych rozwiązań. "*  
*[Thomas Edison]*

---

# Spis treści

---

<b>Spis oznaczeń i symboli</b> .....	<b>11</b>
<b>1 Wstęp</b> .....	<b>16</b>
<b>1.1 Omdlenie</b> .....	<b>16</b>
1.1.1 Omdlenia odruchowe (neurogenne).....	17
1.1.2 Omdlenia wtórne do hipotonii ortostatycznej.....	17
1.1.3 Omdlenia sercowe (sercowo-naczyniowe).....	17
<b>1.2 Częstoskurcz</b> .....	<b>20</b>
<b>1.3 Analizy danych wspierające diagnozę</b> .....	<b>22</b>
<b>2 Cel pracy</b> .....	<b>24</b>
<b>3 Materiał i metody</b> .....	<b>25</b>
<b>3.1 Struktura pracy</b> .....	<b>25</b>
<b>3.2 Materiał</b> .....	<b>25</b>
3.2.1 Wykonywane badania - dane kliniczne.....	25
3.2.1.1 Badanie elektrokardiograficzne (EKG).....	25
3.2.1.2 Badanie echokardiograficzne (ECHO).....	26
3.2.1.3 EKG oraz ECHO podczas badania elektrofizjologicznego (EP) i ablacji.....	26
3.2.2 Analizowana próba badawcza.....	28
3.2.3 Ograniczenia zakresu rozważań.....	29
3.2.4 Potencjalne czynniki ryzyka dla epizodów omdleń.....	30
3.2.4.1 Wiek.....	30
3.2.4.2 Czynniki charakteryzujące odnotowane w przeszłości epizody omdleń.....	31
3.2.4.3 Czynniki opisujące aktywność autonomicznego układu nerwowego.....	32
3.2.4.4 Częstoskurcz.....	33
3.2.4.5 Pozostałe potencjalne czynniki ryzyka dla epizodów omdleń....	35
<b>3.3 Metody</b> .....	<b>35</b>
<b>3.4 Opis metod</b> .....	<b>38</b>
3.4.1 Zastosowanie wstępnej analizy danych.....	38
3.4.1.1 Wybór predyktorów.....	38
3.4.1.2 Wybór reprezentantów – tworzenie wiązki zmiennych.....	39
3.4.1.3 Dyskretyzacja zmiennych.....	39

3.4.2	Regresja liniowa i logistyczna .....	40
3.4.2.1	Jakość modelu .....	44
3.4.2.2	Wybór zmiennych niezależnych .....	44
3.4.2.3	Modele z interakcjami.....	45
3.4.3	Metody redukcji wymiaru .....	45
3.4.3.1	Kryterium wartości własnej i wykresu osypiskowego .....	47
3.4.3.2	Kryterium części wariancji wyjaśnionej przez składowe główne .....	47
3.4.3.3	Idea analizy składowych głównych.....	48
3.4.4	Testy dokładne (exact) .....	48
3.4.5	Metoda Monte Carlo .....	49
3.4.6	Eksploracja danych (data mining).....	50
3.4.6.1	Modele punktacji (scoringowe).....	51
3.4.6.2	Zbiory przybliżone .....	54
3.4.6.3	Drzewa klasyfikacyjne.....	57
3.4.7	Sposoby porównywania modeli .....	61
3.4.8	Zmienne wkłające, interakcyjne i mediacyjne .....	65
3.4.9	Zastosowane oprogramowanie.....	67
<b>3.5</b>	<b>Zastosowana metodologia .....</b>	<b>68</b>
3.5.1	Spełnianie założeń.....	68
3.5.2	Wstępna charakterystyka badanej populacji.....	68
3.5.3	Czynniki powiązane z częstością HR częstoskurczu.....	69
3.5.4	Czynniki powiązane z występowaniem epizodów omdleń .....	70
3.5.5	Porównanie modeli .....	72
<b>4</b>	<b>Wyniki.....</b>	<b>73</b>
<b>4.1</b>	<b>Wstępna charakterystyka badanej populacji.....</b>	<b>73</b>
4.1.1	Porównanie populacji arytmii AVNRT i AVRT (zespół WPW).....	74
4.1.2	Przetwarzanie wstępne.....	77
4.1.2.1	Wybór predyktorów .....	77
4.1.2.2	Wybór reprezentantów .....	79
4.1.2.3	Dyskretyzacja zmiennych.....	80
<b>4.2</b>	<b>Czynniki powiązane z HR częstoskurczu.....</b>	<b>83</b>
4.2.1	Zależności proste.....	83
4.2.2	Wyniki regresji wielorakiej.....	87
<b>4.3</b>	<b>Czynniki ryzyka omdleń .....</b>	<b>88</b>
4.3.1	Statystyki jednowymiarowe:.....	88
4.3.1.1	Czasokres od pierwszego udokumentowanego częstoskurczu .....	91
4.3.1.2	Dodatkowe choroby współistniejące .....	92
4.3.1.3	Czas trwania kołatania .....	94
4.3.1.4	Pora dnia występowania częstoskurczu.....	95

4.3.2	Modele regresji logistycznej.....	96
4.3.2.1	Modele logistyczne po zastosowaniu wstępnej analizy danych	98
4.3.2.2	Zastosowanie składowych głównych.....	105
4.3.3	Modele punktacji (scoringowe) .....	109
4.3.3.1	Drzewa losowe .....	109
4.3.3.2	Ocena powstałych modeli punktacji .....	111
4.3.4	Zastosowanie zbiorów przybliżonych.....	114
4.3.4.1	Generowanie reguł decyzyjnych .....	114
4.3.4.2	Problem niezrównoważenia danych .....	118
4.3.4.3	Klasyfikacja systemu informacyjnego.....	120
4.3.5	Drzewa interakcyjne typu C&RT/CHAID .....	120
4.3.6	Zastosowanie testów dokładnych (exact) i Monte Carlo .....	121
4.4	<b>Porównanie modeli .....</b>	<b>124</b>
5	<b><i>Omówienie wyników i dyskusja .....</i></b>	<b>130</b>
5.1	<b>Istniejący stan wiedzy w zakresie tematu badań .....</b>	<b>131</b>
5.2	<b>Czynniki powiązane z częstotliwością HR częstoskurczu .....</b>	<b>132</b>
5.3	<b>Czynniki powiązane z występowaniem epizodów omdleń .....</b>	<b>134</b>
5.4	<b>Porównanie zastosowanych metod analizy danych .....</b>	<b>137</b>
5.5	<b>Podsumowanie.....</b>	<b>147</b>
6	<b><i>Wnioski.....</i></b>	<b>149</b>
7	<b><i>Streszczenie.....</i></b>	<b>150</b>
8	<b><i>Abstract.....</i></b>	<b>151</b>
	<b><i>Literatura.....</i></b>	<b>152</b>
	<b><i>Załącznik nr 1 .....</i></b>	<b>159</b>
	<b><i>Załącznik nr 2 .....</i></b>	<b>207</b>

# Spis ilustracji

---

Tabela 1. Tabela klasyfikacji na podstawie modelu. ....	61
Tabela 2. Wstępna charakterystyka badanej populacji osób o znanym HR częstoskurczu - zależność omdleń od płci w podgrupach arytmii. ....	73
Tabela 3. Wstępna charakterystyka badanej populacji osób o znanym HR częstoskurczu - zależność omdleń od płci dla obu arytmii łącznie. ....	73
Tabela 4. Wybrane różnice pomiędzy AVNRT i AVRT. ....	74
Tabela 5. Zmienne pozostawione w zbiorze danych po przeprowadzeniu rankingu predyktorów. ....	78
Tabela 6. Wybór reprezentantów z zastosowaniem analizy składowych głównych – ładunki zmiennych. ....	79
Tabela 7. Kategoryzowana zmienna: HR częstoskurczu standaryzowanego – charakterystyka przedziałów. ....	82
Tabela 8. Współczynniki korelacji Spearmana wieku i standaryzowanej wartości HR częstoskurczu w podgrupach płci i rodzaju arytmii. ....	83
Tabela 9. Wyniki porównania (p testu Manna-Whitneya) (1) wieku oraz (2) częstotliwości HR częstoskurczu pomiędzy dwoma podgrupami: płci, współwystępowania dodatkowych, jawnych objawów innych chorób, naprzemienności zespołu QRS oraz arytmii. ....	85
Tabela 10. Porównywane populacje, u których nie wystąpiły istotnie niższe wartości HR częstoskurczu w grupie starszej. ....	85
Tabela 11. Wybrane modele regresji wielorakiej dla HR częstoskurczu zawierające wiek i dodatkowe zmienne. ....	87
Tabela 12. Podsumowanie czynników ryzyka różnicujących grupy z omdleniami i bez omdleń w zależności od płci. ....	90
Tabela 13. Częstotliwość HR częstoskurczu po leku wobec epizodów omdleń dla obu arytmii razem. ....	92
Tabela 14. Należna wartość częstoskurczu w 4 podgrupach; w zależności od rodzaju arytmii i wystąpienia epizodów omdleń. ....	94
Tabela 15. Mediana czasu kołatania w zależności od rodzaju arytmii. ....	94
Tabela 16. Wybrane modele logistyczne opisujące wpływ wieku i częstotliwości HR częstoskurczu na omdlenia w zależności od czasookresu od pierwszego udokumentowanego częstoskurczu. ....	96
Tabela 17. Wybrane modele logistyczne opisujące wpływ pojawiających się stanów przedomdleniowych i częstotliwości HR częstoskurczu na omdlenia w zależności od płci oraz współwystępowania innych jawnych chorób. ....	97
Tabela 18. Zmienne użyte do analizy regresji logistycznej. ....	99
Tabela 19. Badanie współliniowości zmiennych wybranych do analizy. ....	99
Tabela 20. Interakcje istotne statystycznie dla zmiennych wybranych w procesie wstępnej analizy danych. ....	100

Tabela 21. Współczynniki regresji wybranych modeli .....	102
Tabela 22. Wartości własne czynników dla analizowanych grup zmiennych.	105
Tabela 23. Współczynniki modelu logistycznego wykorzystywanego w modelu punktacji.....	111
Tabela 24. Podsumowanie wyników analiz przeprowadzonych przy użyciu metodyki opartej na teorii zbiorów przybliżonych w programie ROSE2 (po automatycznym uzupełnieniu braków danych).....	115
Tabela 25. Przykłady reguł decyzyjnych dla klasy omdleń N, program ROSE2 .....	115
Tabela 26. Przykłady reguł decyzyjnych dla klasy omdleń T, program ROSE2 .....	116
Tabela 27. Przykłady reguł decyzyjnych dla danych po dyskretyzacji, dla klasy omdleń N, program RSES2.....	117
Tabela 28. Przykłady reguł decyzyjnych dla danych po dyskretyzacji, dla klasy omdleń T, program RSES2 .....	118
Tabela 29. Przykłady reguł decyzyjnych dla danych po dyskretyzacji i SMOTE, program RSES2 .....	119
Tabela 30. Podsumowanie wyników analiz przeprowadzonych przy użyciu metodyki opartej na teorii zbiorów przybliżonych dla bazy powiększonej algorytmem SMOTE.....	119
Tabela 31. Wyniki oceny skuteczności klasyfikatorów dla reguł decyzyjnych.	120
Tabela 32. Porównanie częstości występowania omdleń w grupie narażonej i nienarażonej na omawiane ryzyko. ....	121
Tabela 33. Porównanie wyników testów asymptotycznych i dokładnych porównujących cechy jakościowe. ....	122
Tabela 34. Porównanie grupy z i bez epizodów omdleń (test Manna-Whitneya) za pomocą testów asymptotycznych, dokładnych oraz symulacji Monte Carlo. ....	122
Tabela 35. Porównanie wyników wybranych modeli logistycznych dla omdleń, wyznaczonych za pomocą metod asymptotycznych i symulacji Monte Carlo.	123
Tabela 36. Definicje porównywanych, wybranych modeli logistycznych. ....	124
Tabela 37. Porównanie oceny modelu logistycznego (A) jako prognostyka omdleń za pomocą wieku i HR częstoskurczu dla modelu z i bez krosvalidacji. ....	125
Tabela 38. Ocena reguł decyzyjne utworzonych różnymi metodami eksploracji danych (data-minig). ....	128
Tabela 39. Przykładowy system informacyjny (SI) .....	207



Rycina 1. Podstawy patofizjologiczne klasyfikacji omdleń .	16
Rycina 2. Częstość występowania epizodów omdleń w populacji z podziałem na płeć .	19
Rycina 3. Najczęstsze rodzaje częstoskurczu z wąskim zespołem QRS; częstoskurcz typu AVNRT i AVRT.	21
Rycina 4. Postać funkcji logistycznej	42
Rycina 5. Przykładowe drzewo klasyfikacyjne dla zmiennej decyzyjnej „Omdlenie” przyjmującej dwa stany: TAK (T) lub NIE (N)	58
Rycina 6. Krzywa ROC	64
Rycina 7. Moderacja. Zmienna niezależna B jest moderatorem, tzn. wpływa na zależność zasadniczą A i C.	66
Rycina 8. Mediacja. Zmienna niezależna M jest mediatorem, tzn. pośredniczy w zależności zasadniczej A i C.	66
Rycina 9. Pora dnia pojawiania się częstoskurczów typu AVNRT i AVRT.	75
Rycina 10. Wstępna analiza zmiennych.	77
Rycina 11. Wartości WoE dla zmiennej kategorii BMI przed dyskretyzacją.	80
Rycina 12. Wartości WoE dla zmiennej kategorii BMI po dyskretyzacji.	81
Rycina 13. Wartości WoE dla zmiennej HR częstoskurczu standaryzowanego.	82
Rycina 14. Wykres rozrzutu częstotliwości HR częstoskurczu wobec wieku mężczyzn z grupy AVNRT.	83
Rycina 15. Mediana wieku i HR częstoskurczu w podgrupach płci, skategoryzowana względem rodzaju arytmii.	84
Rycina 16. Standaryzowana wartość częstotliwości HR częstoskurczu - zmienne uwikłane.	87
Rycina 17. Zależność HR częstoskurczu od wieku pacjenta w grupie z „krótszym” czasokresem częstoskurczu dla grupy z omdleniami i bez omdleń.	91
Rycina 18. Epizody omdleń wobec współwystępujących dodatkowych objawów innej choroby.	92
Rycina 19. Liczności, w 4 podgrupach, wg należącego częstoskurczu, w zależności od występowania (T) lub nie (N) epizodów omdleń.	93
Rycina 20. Ilorazy szans zmiennych LP i PK z modelu logistycznego P, określające brak wpływu na występowanie epizodów omdleń.	101
Rycina 21. Prawdopodobieństwo wystąpienia omdlenia w podgrupach z wyższymi i niższymi wartościami LP i PK.	103
Rycina 22. Prawdopodobieństwo wystąpienia omdlenia w podgrupach z wyższymi i niższymi wartościami TSH w różnych grupach wiekowych.	104
Rycina 23. Omawiane grupy zmiennych i liczba czynników wybranych za pomocą analizy składowych głównych.	106
Rycina 24. Drzewo klasyfikacyjne utworzone dzięki analizie losowy las dla zmiennych skategoryzowanych.	110

Rycina 25. Drzewo klasyfikacyjne utworzone dzięki analizie losowy las dla danych przed dyskretyzacją.....	111
Rycina 26. Wartości WoE dla zmiennej scoring.....	112
Rycina 27. Wykres czułości i swoistości dla punktacji wyliczonej za pomocą modelu scoringowego S.....	113
Rycina 28. Krzywa ROC dla punktacji wyliczonej za pomocą modelu punktacji (scoringowego) S.....	113
Rycina 29. Zastosowane w pracy kroki analizy danych metodą zbiorów przybliżonych.....	114
Rycina 30. Kroki postępowania z danymi niezrównoważonymi – nadlosowanie.....	118
Rycina 31. Przykładowe drzewo interaktywne typu C&RT.....	120
Rycina 32. Porównanie krzywych ROC dla modelu A z i bez walidacji.....	125
Rycina 33. Czułość i swoistość istotnych statystycznie modeli logistycznych.....	126
Rycina 34. F - miary modeli logistycznych.....	126
Rycina 35. Istotne statystycznie AUC - pola pod krzywą ROC dla wybranych modeli logistycznych.....	127
Rycina 36. G-miary istotnych modeli logistycznych.....	128
Rycina 37. Dolne i górne przybliżenie zbioru - interpretacja graficzna.....	208

---

## Spis oznaczeń i symboli

---

<b>AF</b>	migotanie przedsionków (ang. atrial fibrillation)
<b>Ao</b>	określenie wymiaru aorty: opuszka aorty (mm)
<b>AP</b>	dotatkowy szlak przewodzenia impulsów między przedsionkami i komorami (ang. accessory pathway)
<b>AUC</b>	pole pod wykresem krzywej ROC (ang. area under the curve)
<b>AV</b>	przedsionkowo-komorowy (ang. atrioventricular)
<b>AVNRT</b>	nawrotny częstoskurcz węzłowy (ang. atrioventricular nodal reentry tachycardia)
<b>AVRT</b>	nawrotny częstoskurcz przedsionkowo-komorowy (ang. atrioventricular reentry tachycardia)
<b>bad rate</b>	wskaźnik zdarzeń niepożądanych, stosowany jako opis reguły decyzyjnej, oznaczający stosunek liczby „złych” przypadków (tu: tracących przytomność) do liczby wszystkich w danym przedziale/podziale.
<b>badanie EP</b>	badanie elektrofizjologiczne serca (ang. electrophysiological); EPS (ang. electrophysiological study)
<b>bagging</b>	agregacja wyników sekwencji modeli uzyskanych dla różnych zbiorów przez wielokrotne losowanie ze zwracaniem z analizowanej próby (ang. bagging, inaczej ang. bootstrap aggregating)
<b>BMI</b>	wskaźnik masy ciała (body mass index);
<b>CHAID</b>	typ drzew interakcyjnych (ang. Chi-squared automatic interaction detector - automatyczny detektor interakcji za pomocą Chi-kwadrat)

<b>CL</b>	długości cyklu, np. częstoskurczu, rytmu zatokowego (w ms); (ang. cycle length); odstęp R-R
<b>czasokres częstoskurczu</b>	czasokres od pierwszego udokumentowanego epizodu częstoskurczu (w latach)
<b>częstoskurcz z wąskim zespołem QRS</b>	określenie zapisu elektrokardiograficznego częstoskurczu o czasie trwania zespołu QRS < 120 ms (ang. a narrow QRS complex tachycardia); wskazuje na anatomiczną lokalizację pętli częstoskurczu w obszarze nadkomorowym oraz odzwierciedla fizjologiczną, szybką aktywację komórek poprzez prawidłowy układ His-Purkinje.
<b>dane niezbalansowane</b>	to takie dane, w których liczebność jednej klasy jest znacznie większa od liczebności pozostałych klas; wyniki pozytywne i negatywne nie są w przybliżeniu równoliczne (ang. imbalanced)
<b>dolno-boczna ściana LK</b>	określenie wymiaru grubości dolno-bocznej ściany lewej komory (mm) w badaniu echokardiograficznym (Echo); starsza nazywa: „tylna” ściana lewej komory serca;
<b>EP: CI-Wen-ante</b>	oznaczenie punktu Wenckebacha przewodzenia zstępującego CL (długości cyklu) przed ablacją
<b>EP: CI-Wen- retro</b>	oznaczenie punktu Wenckebacha przewodzenia wstecznego CL (długości cyklu) przed ablacją
<b>EP: HR-Wenck-ante</b>	oznaczenie punktu Wenckebacha przewodzenia zstępującego HR przed ablacją
<b>EP: HR-Wenck-retro</b>	oznaczenie punktu Wenckebacha przewodzenia wstecznego HR przed ablacją
<b>EP: po-CI-Wen-ante</b>	oznaczenie punktu Wenckebacha przewodzenia zstępującego CL (długości cyklu) po ablacji
<b>EP: po-CI-Wen-retro</b>	oznaczenie punktu Wenckebacha przewodzenia wstecznego CL (długości cyklu) po ablacji

<b>EP: po-ERP-Avna</b>	oznaczenie refrakcji przewodzenia zstępującego do komór (refrakcja AVn a) po ablacji
<b>EP: po-ERP-Avnr</b>	oznaczenie refrakcji węzła p-k w kierunku wstecznym (AVn r) po ablacji
<b>ERP</b>	efektywny okres refrakcji (ang. effective refractory period)
<b>FT3</b>	wolna trijodotyronina (ang. free T3) – hormon tarczycy
<b>FT4</b>	wolna tyroksyna; (ang. free T4) – hormon tarczycy
<b>GLM</b>	uogólniony model liniowy (ang. general linear model)
<b>HR</b>	częstotliwość rytmu serca (ang. heart rate)
<b>i.v.</b>	Infuzja dożylna (łac. in venam, i.v.) - sposób podawania płynów bezpośrednio do żył
<b>IV</b>	moc informacyjna zmiennej (ang. Information Value); miara umożliwiająca wykonanie rankingu predyktorów
<b>K+</b>	jony potasu
<b>LEM2</b>	algorytm generowania reguł decyzyjnych (ang. learning from Examples Module, version 2)
<b>LOWESS</b>	nieparametryczna metoda lokalnego wygładzania wielomianami niskiego stopnia; ważona regresja lokalnie wielomianowa (ang. locally weighted scatterplot smoothing); dla każdego punktu i punktów z jego otoczenia wykonywana jest regresja wielomianowa.
<b>LK</b>	lewa komora serca; określenie wymiaru LK (mm) w badaniu echokardiograficznym (Echo)
<b>LP</b>	lewy przedsionek serca; określenie wymiaru LP (mm) w badaniu echokardiograficznym (Echo)

<b>LR</b>	regresja logistyczna (ang. logistic regression)
<b>LR test</b>	test ilorazu wiarygodności (ang. likelihood ratio)
<b>MC</b>	metoda Monte Carlo
<b>Na+</b>	jony sodu
<b>OR</b>	iloraz szans (ang. odds ratio)
<b>overfitting</b>	nadmierne dopasowanie, przeuczenie, przetrenowanie modelu
<b>PCA</b>	analiza składowych głównych (ang. Principal Component Analysis), jedna z metod analizy czynnikowej (ang. Factor Analysis)
<b>PK</b>	prawa komora serca; określenie wymiaru PK (mm) w badaniu echokardiograficznym (Echo)
<b>PMK</b>	przegroda międzykomorowa serca; określenie wymiaru PMK (mm) w badaniu echokardiograficznym (Echo)
<b>p-k</b>	przedsionkowo-komorowy (węzeł), zamiennie: A-V lub AVn (łac. nodus atrioventricularis), węzeł przedsionkowo-komorowy, węzeł AV; jest to część układu bodźco-przewodzącego położona pomiędzy przedsionkami i komorami serca, która przewodzi fizjologicznie bodźce elektryczne z przedsionków do komór.
<b>RF</b>	fale radiowe (ang. radiofrequency); (energia) fal radiowych stosowana w przezskórnej ablacji RF
<b>ROC</b>	narzędzie do oceny poprawności klasyfikatora poprzez łączny opis jego czułości i swoistości (ang. receiver operating characteristic)
<b>RS</b>	zbiory przybliżone (ang. rough sets)

<b>SCD</b>	nagły zgon sercowy (ang. sudden cardiac death)
<b>SE</b>	błąd standardowy estymacji (ang. standard error)
<b>SI</b>	system informacyjny (ang. information system); struktura wiedzy pozwalająca na przetwarzanie, za pomocą procedur i modeli, informacji wejściowych w wyjściowe
<b>SMOTE</b>	metoda wstępnego przetwarzania danych (ang. synthetic minority oversampling technique); algorytm dogenerujący nowe przykłady klasy mniejszościowej pomiędzy przykładami pierwotnymi
<b>SVT</b>	supraventricular tachycardia (częstoskurcz „nadkomorowy”)
<b>szum informacyjny</b>	nadmiar informacji utrudniający wyodrębnienie informacji prawdziwych i istotnych
<b>TSH</b>	tyreotropina, hormon tyreotropowy; (ang. Thyroid-stimulating hormone) - hormon przysadki mózgowej
<b>węzeł AV</b>	węzeł przedsionkowo-komorowy (p-k)
<b>WoE</b>	waga dowodu (ang. Weight of Evidence)
<b>WPW</b>	zespół Wolffa-Parkinsona-White'a (zespół WPW, ang. Wolff-Parkinson-White syndrome)

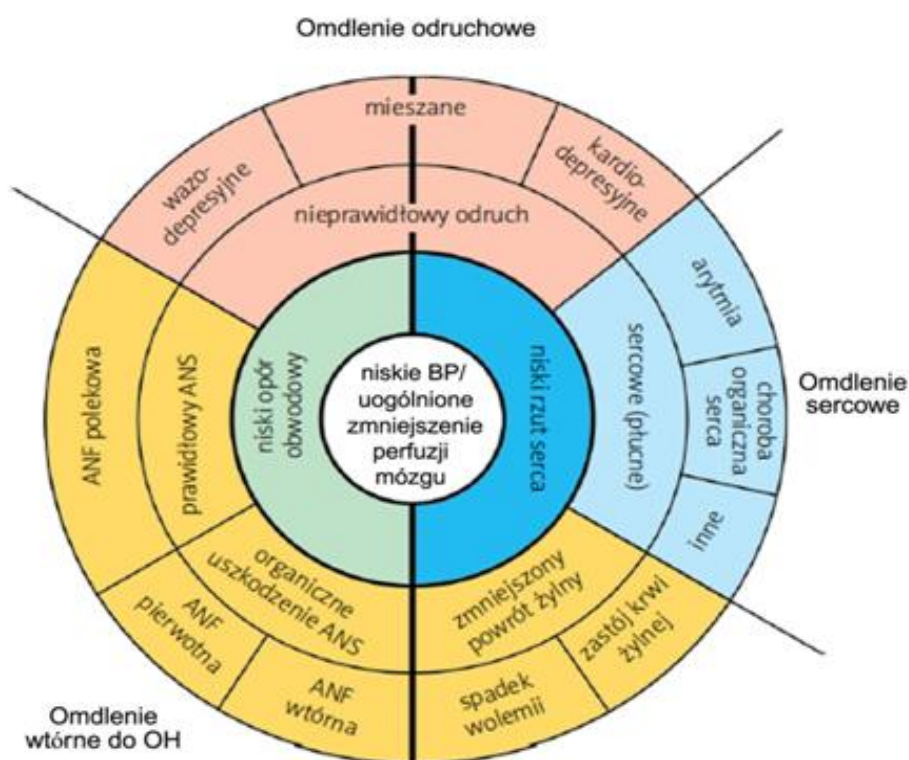
# 1 Wstęp

## 1.1 Omdlenie

Omdlenie (ang. syncope) jest objawem klinicznym określanym jako krótkotrwała, nagła i całkowicie odwracalna utrata przytomności, spowodowana przejściowym niedokrwieniem mózgu. [1] Charakteryzuje się szybkim rozwojem objawów: od gwałtownego początku, aż do samoistnego i całkowitego odzyskania przytomności. Omdlenia mogą być poprzedzone specyficznymi objawami, takimi jak np.: zaburzenia widzenia, potliwość, mdłości czy uczucie ciężkości kończyn dolnych [2]. Takie objawy, przy braku ich progresji, mogą świadczyć również o wystąpieniu stanu przedomdleniowego, czyli „niepełnego omdlenia”.

Mechanizmem warunkującym omdlenia jest przejściowe, uogólnione zmniejszenie perfuzji mózgu. Aby wystąpiła utrata przytomności zazwyczaj wystarczy spadek dopływu tlenu do mózgu o zaledwie 20%, przerwanie mózgowego przepływu krwi przez 6-8 sekund lub istotne klinicznie obniżenie ciśnienia skurczowego krwi do 60 mmHg. [3]

Podstawy patofizjologiczne klasyfikacji omdleń wyróżniają trzy główne mechanizmy [1] (Rycina 1): omdlenia odruchowe (neurogenne), wtórne do hipotonii ortostaticznej oraz sercowe (sercowo-naczyniowe).



Rycina 1. Podstawy patofizjologiczne klasyfikacji omdleń wg [3].

ANF – niewydolność autonomiczna, ANS – autonomiczny układ nerwowy, BP – ciśnienie tętnicze, OH – hipotonia ortostaticzna



### **1.1.1 Omdlenia odruchowe (neurogenne).**

Określenie „omdlenie odruchowe” tradycyjnie odnosi się do heterogenicznej grupy sytuacji klinicznych, w których odruchy sercowo-naczyniowe okresowo reagują na bodziec w sposób niewłaściwy. Prowadzi to do wazodylatacji (rozkurczu mięśni gładkich naczyń krwionośnych z poszerzeniem światła naczyń) i/lub bradykardii (zwolnienia rytmu), a tym samym do istotnego spadku ciśnienia tętniczego oraz uogólnionego zmniejszenia perfuzji mózgu.

### **1.1.2 Omdlenia wtórne do hipotonii ortostatycznej**

Przyczyną omdleń w mechanizmie hipotonii ortostatycznej może być

- pierwotna niewydolność autonomiczna (należy tu: izolowana niewydolność ortostatyczna, zanik wieloukładowy, choroba Parkinsona),
- hipotonia ortostatyczna polekowa (powodowana przez leki moczopędne, leki wazodylatacyjne, leki przeciwdepresyjne, a także alkohol)
- oraz hipotonia powodowana utratą płynów obserwowana w bieguncie, krwotoku czy wymiotach.

### **1.1.3 Omdlenia sercowe (sercowo-naczyniowe).**

Odrębną grupę stanowią omdlenia sercowe (sercowo-naczyniowe) powodowane:

- a. bradykardią (nadmiernym zwolnieniem pracy serca; poniżej 60 uderzeń na minutę),
- b. częstoskurczem (nadmiernym przyspieszeniem rytmu serca o charakterze napadowym; powyżej 100 uderzeń na minutę)
- c. bradykardią lub częstoskurczem wywołanym zażywaniem leków,
- d. chorobą organiczną serca jak np.: ostre niedokrwienie mięśnia serca /zawał serca, choroba zastawkowa, kardiomiopatia przerostowa, tamponada serca, guzy serca, wrodzona anomalie tętnic wieńcowych serca, ale także innymi chorobami (tętniak rozwarstwiający aorty, zator tętnicy płucnej).

Omdlenia w 30% powodowane są przyczynami kardiologicznymi. Wśród nich najczęstsze są zaburzenia rytmu i przewodzenia, czyli arytmie. Prowadzą one do niestabilności hemodynamicznej, a ta powoduje spadek rzutu serca, a w konsekwencji przepływu krwi przez mózg. Omdlenie lub stan przedomdleniowy występuje najczęściej we wczesnej fazie częstoskurczu, w okresie przed

uruchomieniem fizjologicznych mechanizmów wyrównawczych (mechanizmów kompensacyjnych). [3] Najczęściej świadomość powraca szybko, jeszcze przed ustąpieniem częstoskurczu. Jeżeli częstoskurcz powoduje niestabilność hemodynamiczną, pacjent pozostaje nieprzytomny i następuje zatrzymanie krążenia.

Istnieją też inne zaburzenia świadomości, które nie są omdleniami. Można wymienić tu zaburzenia metaboliczne np. hipoglikemia (zbyt niski poziom glukozy we krwi), padaczkę (patologiczne wyładowania bioelektryczne w mózgu z napadami drgawek) oraz przemijający napad niedokrwienia mózgu w obszarze unaczynienia tętnic kręgowo-podstawnych (zaburzenia krążenia podstawno-kręgowego). Występują również zaburzenia przypominające omdlenie, ale przebiegające bez utraty przytomności. Jest to np. katapleksja (chwilowy bezruch), napady padania (chwilowa utrata napięcia mięśniowego) lub zaburzenia psychogenne np. u osób z nerwicą histeryczną [3].

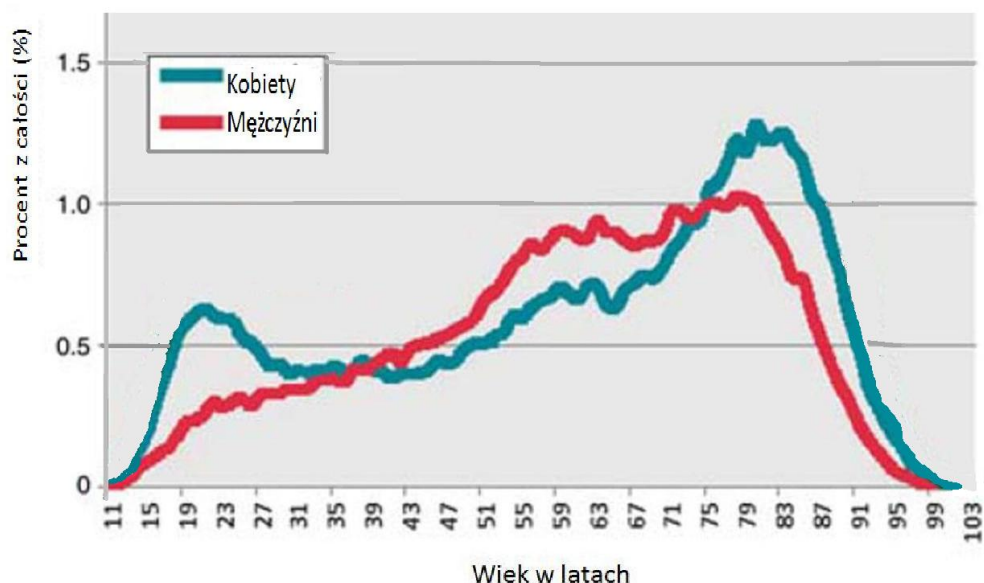
Omdlenia mogą być związane z zawansowanym blokiem przedsionkowo-komorowym (p-k), gdzie nastąpiło istotne zwolnienie czynności serca. Także częstotliwość rytmu serca (nie tylko bardzo szybka czynność serca, ale i nadmiernie wolna) oraz rodzaj arytmii (nadkomorowa lub komorowa), to zwyczajowo najważniejsze czynniki sprawcze omdlenia sercowego. Bardzo istotne jest ułożenie ciała (wyjściowa pozycja ciała) na początku epizodu arytmii oraz współmierność naczyniowych mechanizmów kompensacyjnych. Ponadto istotny jest stan lewej komory serca (brak wcześniejszego uszkodzenia lewej komory) oraz odruchy neurogenne z baroreceptorów niwelujące arytmie.

Omdlenia kardiogenne często występują w czasie wysiłku fizycznego lub bezpośrednio po nim. Całkowita utrata napięcia mięśniowego występuje nagle i często dochodzi do uszkodzenia ciała wskutek nagłego upadku.

Jednocześnie badania wykazują, że roczna śmiertelność osób z omdleniami kardiogennymi jest większa niż u osób z omdleniami o innej lub nieustalonej etiologii [4] [5]. Konsekwencją tego faktu jest konieczność jak najtrafniejszej identyfikacji tej podgrupy chorych w celu zastosowania odpowiedniego i jak najszybszego leczenia arytmii.

Omdlenia stanowią jedynie 3-5% powodów porad ambulatoryjnych oraz 1-3% przyczyn wszystkich hospitalizacji [6]. W populacji ogólnej zdarzają się jednak znacznie częściej, natomiast tylko niewielki odsetek chorych trafia do placówek ochrony zdrowia. Pierwszy epizod omdlenia występuje przeważnie między 10 a 30 rokiem życia, a szczyt występowania omdleń to ok. 15-ty rok życia [7]. Dotyczy to aż 47% kobiet oraz 31% mężczyzn, jednak w tej grupie wiekowej najczęściej są to omdlenia odruchowe [8]. Drugi szczyt występowania omdleń obserwuje się u osób starszych (po 65-tym roku życia). Wykazano, iż

częstość występowania omdleń znacznie wzrasta po ukończeniu 70-go roku życia [9] [10] (Rycina 2)



Rycina 2. Częstość występowania epizodów omdleń w populacji z podziałem na płeć (wg [11] – w modyfikacji własnej).

Omdlenia są istotnym objawem klinicznym, mogącym powodować urazy mechaniczne w czasie upadku (np. złamanie ręki, wstrząśnienie mózgu czy uraz głowy). Mogą być także sygnałem poważnej choroby. Nawracające omdlenia związane są ponadto z utratą pewności siebie, stanami depresyjnymi, strachem przed upadkiem, co (nie tylko w podeszłym wieku) wiąże się z koniecznością objęcia chorego opieką instytucjonalną. Stąd potrzeba wyjaśnienia przyczyn takich epizodów jak i adekwatny wybór postępowania leczniczego, którego głównym celem jest poprawa przeżycia, a także ograniczenie urazów fizycznych i zapobieganie nawrotom.

U pacjentów z omdleniami wtórnymi do organicznej choroby serca, w tym wrodzonych wad serca lub chorób dotyczących krążenia płucnego, celem jest nie tylko zapobieganie nawrotom omdleń, ale także leczenie pierwotnej choroby i zmniejszenie ryzyka poważnych zdarzeń sercowo-naczyniowych lub nagłego zgonu sercowego (SCD, ang. sudden cardiac death). [1]

## 1.2 Częstoskurcz

Częstoskurcz (ang. tachycardia) to zbyt szybkie lub nieregularne tętno – zazwyczaj jest to co najmniej 100 uderzeń na minutę. Przy częstotliwości skurczów w zakresie 150 – 230 uderzeń na minutę (a nawet 350 – 400 uderzeń), serce nie zapewnia organizmowi wystarczającej podaży krwi nasyconej tlenem, gdyż faza rozkurczowa jest za krótka, dla prawidłowego wypełnienia krwią komór serca. [12]

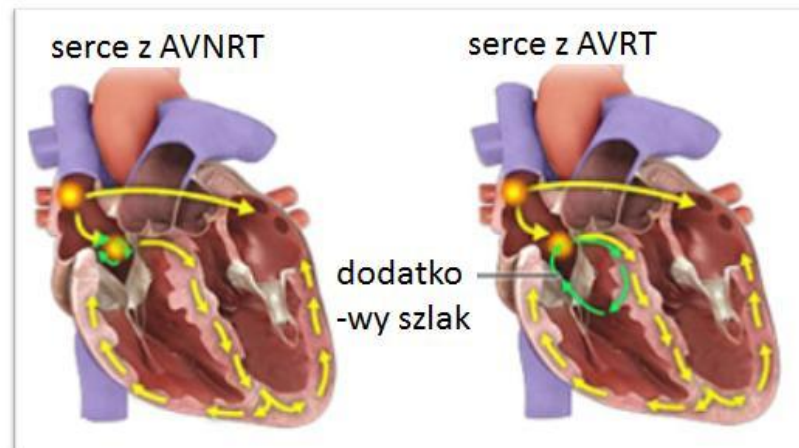
Częstoskurcz, który odzwierciedla aktywację komór poprzez fizjologiczny układ Hisa-Purkiniego, z szerokością zespołu QRS (w EKG) o czasie trwania < 120 ms, zwany jest także, choć nie zawsze poprawnie, częstoskurczem „nadkomorowym” (SVT, ang. supraventricular tachycardia). Anatomiczne umiejscowienie ogniska częstoskurczu lub krążącego obwodu pobudzenia (pętli reentry), mieści się w obrębie przedsionków serca, węzła zatokowego, łączy p-k, pęczka Hisa, albo stanowi dowolną kombinację wymienionych obszarów. Poprawna i zalecana nazwa (wg Europejskiego Towarzystwa Kardiologicznego oraz amerykańskiego towarzystwa Heart Rhythm Society) to częstoskurcz z wąskim zespołem QRS, gdy w zapisie EKG szerokość zespołu QRS < 120 ms (ang. a narrow QRS complex tachycardia).

Najczęstsze formy częstoskurczu z wąskim zespołem QRS:

- **Nawrotny częstoskurcz węzłowy** (AVNRT – AV nodal reentrant tachycardia). Najprostsza teoria tłumacząca umiejscowienie tej arytmii zakłada obecność dwóch odmiennych elektrofizjologicznie struktur w obrębie węzła p-k: drogi wolnej i szybkiej. Tworzą one pętlę częstoskurczu typu AVNRT.
- **Nawrotny częstoskurcz przedsionkowo-komorowy** (AVRT – AV reentrant tachycardia). Częstoskurcz typu AVRT jest arytmia w grupie objawowych chorych z dodatkowym szlakiem p-k. Ramię zstępujące AVRT stanowi najczęściej fizjologiczny układ Hisa-Purkiniego, natomiast ramię wsteczne tworzy dodatkowe połączenie przedsionkowo-komorowe.
- **Częstoskurcz przedsionkowy** (AT – atrial tachycardia). Częstoskurcz przedsionkowy występuje w mechanizmie automatyzmu patologicznego lub zjawiska pobudzenia krążącego. Obecność izolowanego ogniska AT w prawym lub lewym przedsionku sugeruje automatyzm patologiczny, w odróżnieniu od częstoskurczu typu reentry, gdzie pętla AT obejmuje większy obszar prawego i/lub lewego przedsionka.

Zjawisko krążącego pobudzenia (ang. reentry) jest przyczyną większości istotnych klinicznie zaburzeń rytmu serca. Ten mechanizm warunkuje powstanie wielu różnorodnych arytmii obejmujących migotanie przedsionków, trzepotanie przedsionków, nawrotny częstoskurcz węzłowy, częstoskurcz przedsionkowo-

komorowy, a także pozawałowy częstoskurcz komorowy i migotanie komór. Zgodnie z definicją [13], częstoskurcz to częstotliwość rytmu serca powyżej 100/min. Jednak częstotliwość rytmu w częstoskurczu zależy w dużym stopniu od szybkości przewodzenia przez węzeł przedsionkowo-komorowy w kierunku zstępującym, a także w kierunku wstecznym przez węzeł p-k, bądź wstecznie, poprzez dodatkowe połączenie p-k. Najczęściej występuje częstoskurcz z częstotliwością rytmu 150 – 190/min, ale nierzadko osiąga rytm 200 – 230/min. [14]



Rycina 3. Najczęstsze rodzaje częstoskurczu z wąskim zespołem QRS; częstoskurcz typu AVNRT i AVRT. (wg [15])

Nawrotny częstoskurcz węzłowy oraz nawrotny częstoskurcz przedsionkowo-komorowy przedstawiono na Rycinie 3.

Objawy związane z arytmia mogą być różnorodne, zależą od częstotliwości rytmu serca podczas częstoskurczu, efektywnego ciśnienia tętniczego krwi oraz perfuzji tkanek. Istotnym czynnikiem generowania oraz nasilenia objawów klinicznych są także współistniejące choroby. Najczęstsze objawy kliniczne częstoskurczu to: kołatania serca, omdlenia i stany przedomdleniowe. Ponadto występują: zawroty głowy, zaburzenia dyspeptyczne, dyskomfort i/lub ból w klatce piersiowej oraz odczucie „krótkiego” oddechu.

Częstoskurcz nawrotny w węźle przedsionkowo-komorowym (AVNRT) jest najczęstszą, a jednocześnie najbardziej złożoną postacią miarowego częstoskurczu z wąskim zespołem QRS. Częstoskurcz typu AVNRT dominuje w populacji osób dorosłych i częściej występuje u kobiet niż u mężczyzn, powstaje w wyniku zaburzenia szerzenia impulsu elektrycznego w obrębie serca, zarówno u zdrowych, jak i u osób ze strukturalnym uszkodzeniem serca. Początkowo lokalizowano go wyłącznie w obrębie węzła przedsionkowo-komorowego (p-k), jednak później wykazano [16], że dla podtrzymania częstoskurczu węzłowego potrzebny jest niewielki obszar przedsionka w bezpośrednim sąsiedztwie węzła p-k [17]. Występują liczne postaci tej arytmii, ich modele anatomiczne i patofizjologiczne różnią się pętlą częstoskurczu

obejmującą nieco inny obszar, w tym łącze przedsionkowo-komorowe, przyległe obszary przedsionka, a szczególnie tzw. wejścia przedsionkowe zróżnicowane co najmniej na wejście przednio-górne i tylny-dolne, rzadziej wejście od strony lewego przedsionka [18].

Diagnostyka nawrotnego częstoskurczu obejmuje dane z wywiadu pacjenta z potwierdzeniem objawów kołatania serca i/lub omdleń, analizę zapisu EKG w trakcie kołatania serca, rzadziej EKG wysiłkowego (jeśli w wywiadzie okoliczności wystąpienia arytmii obejmują wysiłek fizyczny). Ponadto przydatne jest 24-godzinne monitorowanie EKG metodą Holtera, a najlepiej wielodniowe monitorowanie EKG. Metoda inwazyjnego badania elektrofizjologicznego (EPS, ang. electrophysiological study), z wprowadzeniem elektrod diagnostycznych do jam serca, rzadko stosowana jest jedynie dla celów diagnostycznych,

Celem postępowania diagnostycznego jest:

- Określenie stabilności hemodynamicznej pacjenta (lub realnego zagrożenia życia) w czasie stwierdzanych objawów klinicznych,
- Ocena zapisu EKG z ustaleniem prawdopodobnego rozpoznania dotyczącego typu częstoskurczu (częstoskurcz komorowy vs. inne formy częstoskurczu).

W praktyce te dwa cele realizowane są równocześnie w czasie klinicznej oceny pacjenta z częstoskurczem.

Układ bodźco-przewodzący serca, zawierający węzeł zatokowy, węzeł p-k oraz układ Hisa-Purkiniego jest odpowiedzialny za tworzenie impulsu elektrycznego, jego przemieszczanie oraz koordynację elektrycznej aktywacji w celu zapewnienia efektywnego skurczu serca. W zapisie EKG zespół QRS reprezentuje depolaryzację elektryczną komórek serca. Mechaniczną odpowiedzią na ten bodziec elektryczny jest skurcz komórek. Fizjologiczną depolaryzację mięśnia serca odzwierciedla w zapisie EKG wąski zespół QRS trwający krócej niż 110 ms. Poszerzenie zespołu QRS powyżej 120 ms, najczęściej utożsamiane jest z podejrzeniem częstoskurczu komorowego, gdzie ognisko aktywacji częstoskurczu zlokalizowane jest w obrębie komórek serca.

### **1.3 Analizy danych wspierające diagnozę**

W praktyce klinicznej tzw. „kardiologiczne stany nagłe”, obejmujące między innymi utraty przytomności, w tym omdlenia, ale także stany przedomdleniowe, stanowią często realne zagrożenie dla pacjenta. [19] To też niezmiernie ważna jest szybka diagnoza oceniająca faktyczny stan pacjenta i, co za tym idzie, wybór adekwatnego postępowania leczniczego. Wiąże się to z poszukiwaniem czynników ryzyka omdlenia, które może być wywołane między innymi

częstoskurczem z wąskim zespołem QRS. Wyselekcjonowanie czynników ryzyka omdlenia dla osób ze zdiagnozowaną lub podejrzaną arytmia może wzbogacić wiedzę dotyczącą mechanizmu omdlenia oraz przyczynić się do lepszego wyselekcjonowania potencjalnie zagrożonych pacjentów, a tym samym do poprawy ich przeżycia. [6] Niniejsza praca jest również próbą znalezienia odpowiedzi na pytanie o ewentualny związek częstotliwości HR częstoskurczu z występowaniem epizodów omdleń w grupie częstoskurczu z wąskim zespołem QRS.

W dziedzinie medycznych badań naukowych coraz częściej sięga się po narzędzia modelujące zależności pomiędzy czynnikami chorobotwórczymi. Rozwój metod statystycznych oraz technik służących eksploracji danych, oferuje narzędzia sprzyjające przeprowadzaniu coraz dokładniejszych analiz, a co za tym idzie, sprzyjające rozwojowi nauk medycznych. Wśród powszechnie stosowanych technik analitycznych znajdują się zarówno te bardziej klasyczne (np. metody regresji logistycznej - [20]), jak i te zaliczane do grupy metod określanych terminem uczenia maszynowego (ang. machine learning) – [21]. Ze względu na duży postęp technologiczny w dziedzinie sprzętu i oprogramowania komputerów oraz w dziedzinie zautomatyzowanej identyfikacji i gromadzenia danych, coraz częściej stosuje się metody eksploracji danych, inaczej „data mining” - czyli metody „drażenia” lub „zglobiania” danych przechowywanych w bazach lub tzw. hurtowniach danych [22]. Metody te stosowane są również w przypadku, gdy w badanych danych występuje więcej zmiennych (cech) niż można efektywnie i wydajnie uwzględnić podczas tworzenia klasycznych modeli statystycznych. W przypadku metod eksploracji danych ocena modelu nie dokonuje się na podstawie istotności statystycznej, lecz na podstawie poprawności na zbiorze testowym. Formułowane wnioski wypływają zawsze z analizy dostępnych zbiorów danych, a nie z „a priori” przyjmowanych abstrakcyjnych teorii. Uzyskiwane w ten sposób rezultaty analizy mogą być reprezentowane i przechowywane w różnych formach (np. drzewa decyzyjnego) [23]. Wybór konkretnych metod analizy danych związany jest z próbą jak najbardziej intuicyjnego i przejrzystego sposobu wspierania diagnozy oraz ze specyfiką analizowanych danych.

W pracy założono, że odpowiednio dobrane metody statystyczne i eksploracji danych mogą pomóc w wyjaśnieniu zjawiska omdleń w analizowanej populacji chorych z częstoskurczem z wąskim zespołem QRS poprzez identyfikację pacjentów o podwyższonym, czy też znacznie zmniejszonym ryzyku wystąpienia omdleń.

---

## 2 Cel pracy

---

Celem pracy jest statystyczna ocena wpływu częstotliwości rytmu serca podczas częstoskurczu oraz innych potencjalnych czynników ryzyka na występowanie epizodów omdleń w grupie chorych z wąskim zespołem QRS (AVNRT i AVRT). Dążono do utworzenia modelu, który mógłby wyodrębnić grupę najbardziej zagrożoną wystąpieniem omdlenia w analizowanej populacji.

### W szczególności celem pracy było:

1. Przedstawienie wad i zalet oraz porównanie wyników uzyskanych za pomocą różnych metod statystycznych i technik eksploracji danych użytych do rozwiązania omawianych problemów klinicznych.
2. Ocena czynników powiązanych z częstotliwością HR częstoskurczu w grupie chorych z wąskim zespołem QRS (AVNRT i AVRT).
3. Ocena czynników ryzyka omdlenia w analizowanej grupie chorych z wąskim zespołem QRS (AVNRT i AVRT).
4. Utworzenie modelu charakteryzującego osoby z epizodami omdleń, który mógłby wesprzeć proces diagnozy pacjentów o podwyższonym zagrożeniu wystąpieniem omdlenia sercowego.



---

## 3 Materiał i metody

---

### 3.1 Struktura pracy

Praca zawiera elementy klinicznej oceny rozpatrywanych czynników ryzyka omdlenia w analizowanej grupie badawczej oraz ocenę przydatności wybranych metod statystycznych i eksploracji danych wykorzystanych do poszukiwania tychże czynników.

Dla przejrzystości pracy, odstąpiono od zwyczajowego umieszczania wszystkich tabel zawierających wyniki analiz w tekście rozprawy i umieszczono je w Załączniku nr 1. Szczegółowe wyniki, ze względu na ich znaczną objętość, umieszczono na stronie Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu (<http://www.ump.edu.pl/index.php?department=70>).

### 3.2 Materiał

#### 3.2.1 Wykonywane badania - dane kliniczne

U każdego pacjenta, włączonego do badanej grupy w trakcie hospitalizacji na oddziale kardiologicznym I Kliniki Kardiologii Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu, zebrano wywiad, wykonano badanie przedmiotowe, rutynowe badania laboratoryjne, standardowy zapis EKG, a także badanie echokardiograficzne serca (ECHO). Zweryfikowano stosowane wcześniej leki antyarytmiczne (AA). Inwazyjne badanie EP i ablację RF wykonywano najczęściej w 2. dobie hospitalizacji bez uprzedniego stosowania leków AA lub zalecano wcześniej odstawienie AA na okres obejmujący co najmniej 5 okresów półtrwania leku.

##### 3.2.1.1 Badanie elektrokardiograficzne (EKG)

U wszystkich pacjentów analizowano 12-odprowadzeniowy zapis elektrokardiograficzny (EKG) wykonany w dniu hospitalizacji lub dodatkowo w czasie arytmii. Analizowano także każdy dostarczony przez pacjenta EKG i/lub Holter-EKG z arytmia udokumentowaną w okresie przed hospitalizacją. EKG w czasie hospitalizacji wykonano z przesuwem 25 mm/s oraz cechą 10 mm/mV. Oceniono występowanie cech preekscytacji, arytmii (np. AVNRT, AVRT, AF), a także inne odchylenia od stanu prawidłowego.

W EKG oceniano: rytm wiodący, częstotliwość zespołów QRS, czas trwania QRS, średnią oś elektryczną zespołu QRS oraz jego morfologię. W EKG oceniano falę delta przy szerokości zespołu QRS  $\geq 120$  ms, skróceniu odstępu PQ  $< 120$  ms, wiodącym rytmem zatokowym. Za falę delta przyjęto początkowe 40 ms zespołu QRS przewidzianego z preekscytacją, w każdym

odprowadzeniu kończynowym oraz początkowe 60 ms w odprowadzeniach przedsercowych.

Dodatkowo podczas badania EP i ablacji RF wykonywano EKG (stosując komputerowy zestaw elektrofizjologiczny BARD Lab System™PRO EP Recording; C.R.Bard.Inc. USA) każdego zaobserwowanego /wyzwolonego częstoskurczu lub innej formy arytmii.

### ***3.2.1.2 Badanie echokardiograficzne (ECHO)***

U każdego pacjenta wykonano badanie echokardiograficzne serca w dniu poprzedzającym badanie EP i ablację RF oraz badanie kontrolne ECHO po zabiegu. W badaniu oceniano podstawowe parametry echokardiograficzne obejmujące wielkość jam serca, stan czynnościowy zastawek serca, kurczliwość i frakcję wyrzutową lewej komory serca oraz ewentualną obecność płynu w worku osierdziowym. [24][25]

### ***3.2.1.3 EKG oraz ECHO podczas badania elektrofizjologicznego (EP) i ablacji***

#### **3.2.1.3.1 Badanie EP**

Inwazyjne badanie elektrofizjologiczne wykonano stosując standardowe techniki. Badanie EP prawie zawsze wykonywano jako badanie poprzedzające ablację serca RF. Badanie EP (z ablacją RF) wykonywano w 2. dobie hospitalizacji. Każdorazowo uzyskano zgodę od pacjenta na wykonanie badania elektrofizjologicznego i ablacji RF. Żyłny dostęp naczyniowy wg techniki Seldingera wykonywano najczęściej z prawej pachwiny. Nakłucie żyły podobojczykowej prawej (lub szyjnej wewnętrznej) stosowano przy wprowadzaniu elektrody diagnostycznej do zatoki wieńcowej. Zapisy wewnątrzsercowe z dwubiegunowych lub czterobiegunowych elektrod diagnostycznych gromadzono cyfrowo (BARD Lab System™PRO EP Recording; C.R.Bard.Inc.; USA) równoległe z 12 –odprowadzeniowym zapisem EKG. Elektrody diagnostyczne 6F czterobiegunowe (z 2 lub 4 mm odstępem między biegunami) lokowano w górnej części przedsionka, obszarze pęczka Hisa oraz wierzchołka prawej komory. Dodatkowo gromadzono wewnątrzsercowe zapisy z wielopolowej elektrody diagnostycznej 6 F (dziesięciobiegunowej) wprowadzanej do zatoki wieńcowej. Sygnały zapisano z częstotliwością próbkowania 1000 Hz. Zastosowano typowe ustawienia filtrów dla wewnątrzsercowych, dwubiegunowych zapisów. Wszystkie pomiary dokonano przy przesuwie zapisu EKG 200 mm/s (BARD Lab System). Wewnątrzsercową stymulację prowadzono impulsem prostokątnym o szerokości 1 ms (Biotronik) przy dwukrotnej wartości potencjału progowego. Amplituda impulsu mieściła się w zakresie 5 – 10 mA. Przemieszczanie oraz lokalizację elektrody diagnostycznej potwierdzono stosując obraz RTG w

projekcji: skośnej lewej, skośnej prawej lub przednio-tylnej.

EKG 12-odprowadzeniowy, jak i zapisy sygnałów wewnątrzsercowych w dalszych etapach badania EP wykonywano w sposób umożliwiający analizowanie krzywej EKG (w dowolnym momencie) w zakresie przesuwu 25-200 mm/s. Wszystkie zapisy EKG, a także sygnały wewnątrzsercowe zapisywano w pamięci komputera. Wykonano pomiary podstawowych wartości elektrofizjologicznych w czasie rytmu zatokowego: częstotliwość rytmu serca (HR), odstęp R-R (jest to synonim długości cyklu rytmu zatokowego – CL), odstęp PQ lub P-delta, szerokość zespołu QRS, odstęp QT, w tym skorygowany odstęp QT (QTc) obliczony wg wzoru Bazetta. [26] [27] [28] Później z zapisu wewnątrzsercowego dodatkowo oceniano czas trwania odstępu AH i HV. Równocześnie analizowano krzywą EKG w celu oceny obecności preekscytacji.

W dalszym etapie badania EP wykonywano:

- a) Stymulację prawej komory z dodatkowym impulsem ze skracanym sprzężeniem – RRP (retrograde refractory period) np. PCL 600+S2; PCL 500+S2.

Cel: oznaczenie refrakcji węzła p-k w kierunku wstecznym (AVn r)

- b) Stopniowaną stymulację komór – IVP (incremental –HR; decremental – CL ventricular pacing) Cel: wyznaczenie punktu Wenckebacha przewodzenia wstecznego, ocena refrakcji AVn r, potwierdzenie przewodzenia przez AVn i/lub dodatkowe połączenie p-k (AP)

- c) Stymulację prawego przedsionka z dodatkowym impulsem ze skracanym sprzężeniem – ARP (antegrade refractory period) np. PCL 600+S2; PCL 500+S2.

Cel: oznaczenie refrakcji przewodzenia zstępującego do komór (refrakcja AVn a) lub ocena preekscytacji z oceną refrakcji AP a; dodatkowo ocena maksymalnej preekscytacji.

- d) Stopniowana stymulacja przedsionków – IAP (incremental –HR decremental – CL atrial pacing) Cel: wyznaczenie punktu Wenckebacha przewodzenia zstępującego, refrakcja AVn a, potwierdzenie przewodzenia przez AVn i/lub AP.

Niektórym pacjentom podczas badania elektrofizjologicznego podano leki (izoproterenol, adenozyne) w celu diagnostycznym.

Obecnie bardzo rzadko wykonuje się badanie EP jako procedurę samodzielną. Inwazyjne badanie elektrofizjologiczne serca najczęściej poprzedza zabieg ablacji RF, czyli wchodzi w zakres procedury usunięcia podłoża zaburzeń rytmu serca.

### 3.2.1.3.2 Ablacja RF

U pacjentów z omdleniami poprzedzonymi nagłym, krótkotrwałym epizodem uczucia kołatania serca, sugerującym SVT, wykonanie badania EP jest wskazane w celu potwierdzenia mechanizmu arytmii oraz wykonania terapeutycznego zabiegu ablacji przezskórnej.

Przezskórna ablacja prądem o wysokiej częstotliwości (RF, ang. radiofrequency) jest uznaną metodą leczenia częstoskurczów nawrotnych [29]. Istota zabiegu polega na zniszczeniu podłoża anatomicznego krążącego pobudzenia nawrotnego (ang. reentry). W przypadku częstoskurczu węzłowego (AVNRT, ang. atrioventricular nodal reentrant tachycardia) wykonuje się ablację drogi wolnej łączy przedsionkowo-komorowego (AV ang. atrio-ventricular), natomiast u chorych z zespołem preekscytacji i częstoskurczami przedsionkowo-komorowymi (AVRT, ang. atrioventricular reciprocating tachycardia) leczenie polega na ablacji dodatkowego szlaku przedsionkowo-komorowego. Dodatkowy szlak przewodzenia pomiędzy przedsionkami a komorami może niekiedy stanowić przyczyną nagłego zgonu. Ablacja RF natomiast można usunąć dodatkowy szlak p-k w niemal 100% przypadków. Pacjenci po skutecznej ablacji wracają do pełni zdrowia i nie wymagają dalszego leczenia.

### 3.2.2 Analizowana próba badawcza

Niniejsza praca rozpatruje problem omdlenia w grupie pacjentów kierowanych do Poradni Kardiologicznej i/lub hospitalizowanych na oddziale kardiologicznym. W tej grupie dominują omdlenia o podłożu kardiogennym, czyli spowodowane częstoskurczem. Kliniczny obraz epizodów omdleń bywa różnorodny: od łagodnego pogorszenia funkcji układu krążenia do istotnej niestabilności hemodynamicznej serca. Jeżeli kliniczne cechy omdlenia sugerują podejrzenie aktywnej choroby wieńcowej, istotnej arytmii serca, stanowi to pilne wskazanie do oceny stanu pacjenta i często wymaga hospitalizacji. [5]

W pracy przeanalizowano dane dotyczące pacjentów z częstoskurczem z wąskim zespołem QRS (AVNRT i AVRT). Badaniem została objęta grupa 517 osób, którym w latach 2000-2012 wykonano rutynowy zabieg ablacji RF z powodu częstoskurczu w Pracowni Elektrofizjologii Serca I Kliniki Kardiologii Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu. Podstawą procesu diagnostycznego były dane z wywiadu dotyczące omdleń w badanej grupie.

Analizie poddano dane kliniczne oraz wyniki badań elektrofizjologicznych serca. Przeanalizowano dwie grupy arytmii:

- nawrotny częstoskurcz węzłowy (atrioventricular nodal reentry tachycardia, AVNRT) (N=289) oraz
- nawrotny częstoskurcz przedsionkowo-komorowy (atrioventricular reentry tachycardia, AVRT) (N=228) z dodatkowym połączeniem przedsionkowo-komorowym.

Grupa AVRT obejmuje zespół Wolffa-Parkinsona-White'a (WPW). Dla wszystkich pacjentów zgromadzono dane zebrane z wywiadu oraz wyniki badania elektrofizjologicznego przeprowadzonego przed zabiegiem i bezpośrednio po nim.

Analizowany zbiór danych może być uznany za próbę reprezentatywną dla częstoskurczu z wąskim zespołem QRS (AVNRT i AVRT) z uwagi na wystarczającą licznosc próby (N=517) oraz opisaną częstość występowania omdleń dla tej grupy (ok. 20%) zgodną z opisywaną w literaturze [30] [31] [32].

### 3.2.3 Ograniczenia zakresu rozważań

W przyszłości, w diagnostyce omdleń, prawdopodobnie wzrośnie rola implantowanych urządzeń długotrwale monitorujących rytm serca, umożliwiających analizę zapisu EKG w momencie wystąpienia omdlenia i/lub utraty przytomności. Aktualnie dostępne są urządzenia do wielodniowego zapisu EKG, ale umożliwiające zapis tylko z dwóch kanałów EKG (w standardowym zapisie EKG jest 12 odprowadzeń). Ponadto istnieją już wszczepialne urządzenia do monitorowania EKG (ILR – insertable loop recorder, zwane też ICM – insertable cardiac monitor), jednokanałowe zapisy umożliwiające monitorowanie EKG nawet do 3 lat. Nie jest to tylko problem braku dostępności tych urządzeń (urządzenia nie są refundowane przez Narodowy Fundusz Zdrowia), ale także konieczność zastosowania odpowiedniego algorytmu wykrywającego (różnicującego) obecność zatokowego załamka P (także wstecznego P') w czasie częstoskurczu. Bowiem, tylko to umożliwia automatyczną dyskryminację częstoskurczu z wąskim lub szerokim zespołem QRS, a niekiedy różnicowanie zjawiska jest niewiarygodne lub nadal niemożliwe. Aktualnie nadal głównym źródłem informacji o występowaniu omdleń są dane z wywiadu od pacjenta. Taka analiza danych zbieranych głównie z wywiadu jest obciążona pewnym błędem związanym najczęściej z niekompletną dokumentacją, zawodnością pamięci pacjentów itp.

Analizowany zbiór danych nie posiadał wyników badań precyzujących odchylenia stanu neurologicznego, a także stanu psychiatrycznego pacjentów, które mogłyby być istotne w sprawie analizy przyczyn omdleń. [33] [1]

W niniejszej pracy przyjęto jednak założenie, iż analizowane przypadki omdleń to omdlenia kardiogenne, ze względu na zdiagnozowany napadowy częstoskurcz. Jest to założenie konieczne, gdyż pacjenci z omdleniami i niewydolnością serca charakteryzują się wysokim ryzykiem zgonu niezależnie od przyczyny omdleń [19]. Poza tym zalecenia z 2009 roku [3] precyzują zmiany kliniczne lub elektrokardiograficzne przynależne omdleniom arytmicznym, w tym kołatania serca poprzedzające omdlenie lub w trakcie omdlenia oraz preekscytację w zapisie EKG [34].

#### **3.2.4 Potencjalne czynniki ryzyka dla epizodów omdleń**

Czynniki związane z potencjalnym wzrostem ryzyka pojawienia się niepożądanego epizodu omdlenia mogą być różnorodne. [9] W niniejszej pracy, opisującej grupę chorych z arytmia z wąskim zespołem QRS (AVNRT i AVRT), poszukiwano czynników ryzyka omdleń pośród danych klinicznych oraz wyników badania elektrofizjologicznego serca. [4] W grupie z napadowym częstoskurczem rola wieku pacjenta jest szczególna. [11] Należy również zwrócić uwagę na charakterystykę odnotowanych w przeszłości epizodów omdleń: częstość omdleń, czasokres od pierwszego udokumentowanego częstoskurczu, czas trwania kołatań oraz występowanie stanów przedomdleniowych. [30] Ważnym elementem prawidłowej funkcji układu krążenia, w tym także rytmu serca, jest aktywność autonomicznego układu nerwowego. Modulujący wpływ układu współczulnego na czynność serca, jak i układu przywspółczulnego, zapewnia dobową zmienność rytmu serca (odmienną w czasie aktywności dziennej, jak i w czasie spoczynku w godzinach nocnych). Z uwagi na fakt, iż rytm serca modulowany jest oprócz napięcia układu nerwowego również czynnikami biochemicznymi (jest m.in. zależny od stężenia jonów wapnia, potasu, a w mniejszym stopniu sodu we krwi) i hormonalnymi (np. adrenaliny i tyroksyny), także te czynniki rozpatrywane były w niniejszej pracy jako potencjalne czynniki ryzyka dla epizodów omdleń.

##### **3.2.4.1 Wiek**

Rozpatrywane w niniejszej pracy grupy chorych z dwoma rodzajami arytmii (AVNRT i AVRT) zależą od procesu starzenia. [12] Ponieważ w młodym wieku rzadziej współistnieją także inne choroby, a u osób starszych omdlenia mogą występować przy wolniejszym częstoskurczu, w niniejszej pracy wiek był jednym z czynników branych pod uwagę przy tworzeniu modeli opisujących grupę omdleń.

### **3.2.4.2 Czynniki charakteryzujące odnotowane w przeszłości epizody omdleń**

#### **(1) Częstość występowania omdleń**

W analizowanej populacji trudno jest ustalić pełną częstość omdleń w ciągu życia z uwagi na niemierną oceną epizodów, do których dochodziło przed wieloma laty. Liczba epizodów omdleń w trakcie życia jest najsilniejszym czynnikiem prognostycznym nawrotów. Około 1/3 chorych doświadcza ponownego omdlenia w okresie 3 lat od obserwowanego zdarzenia [30].

#### **(2) Stany przedomdleniowe**

Objawy stanu przedomdleniowego mogą być nieswoiste (np. zawroty głowy) i często są takie same jak objawy poprzedzające utratę przytomności. U niektórych osób występują: mroczki przed oczami (zaburzenia widzenia), pocenie, zawroty głowy, mdłości czy uczucie ciężkości kończyn dolnych [2]. W niniejszej pracy podjęto próbę oceny, czy takie stany przedomdleniowe mogą być uznane za diagnostyczne w analizowanej grupie chorych z arytmia z wąskim zespołem QRS.

#### **(3) Czasokres od pierwszego udokumentowanego częstoskurczu**

Dla analizowanej grupy zgromadzono dane dotyczące wieku pacjenta (roku życia), gdy wystąpiło pierwsze udokumentowane kołatanie serca lub pierwszy udokumentowany epizod częstoskurczu. [3] Zmienna ta została następnie wykorzystana do wyliczenia tzw. czasokresu objawów kołatania i częstoskurczu. „Najdłuższy” czasokres częstoskurczu obejmował 66 lat. W pracy przyjęto okres 20 lat od pojawienia się kołatań jako granicę pomiędzy „dłuższym” a „krótszym” czasokresem częstoskurczu (podział własny).

#### **(4) Czas trwania kołatania**

W bazie EP zgromadzono dane dotyczące czasu trwania kołatania serca (w minutach bądź w godzinach), czyli objawu, który reprezentował trwający już częstoskurcz. Pacjentów różnicował czas trwania typowych (często występujących) epizodów oraz tzw. maksymalnego czasu trwania kołatania serca, czyli najdłuższego epizodu częstoskurczu, którego doświadczyli. Poszukiwano odpowiedzi na pytanie, czy dłużej trwający częstoskurcz istotnie częściej powoduje omdlenia. [2]

Czas trwania kołatania pomierzono w skali porządkowej. Zastosowano następujące kodowanie (skala własna):

- 1 – od kilku sekund do kilku minut
- 2 – 10 do 30 min
- 3 – 30 do 60 min
- 4 – 1 do 2 h

- 5 – 2 do 4 h
- 6 – 4 do 12 h
- 7 – powyżej 12 h

Maksymalny czas trwania kołatania analizowano także w następujących grupach (skala własna):

- do 120 min
- 120 – 360 min
- ponad 360 min

### ***3.2.4.3 Czynniki opisujące aktywność autonomicznego układu nerwowego***

Nagromadzona wiedza kliniczna wskazuje na związek pomiędzy aktywnością autonomicznego układu nerwowego i chorobami układu sercowo-naczyniowego. Wysoka aktywność układu współczulnego i/lub obniżenie aktywności nerwu błędnego (układ przywspółczulny) pozostaje w związku przyczynowym z występowaniem między innymi groźnych arytmii i nagłych zgonów sercowych. [30] Poprzez te mechanizmy przewaga układu sympatycznego sprzyja rozwojowi miażdżycy, przerostu mięśnia sercowego i mięśniówki naczyń. Każdy z wymienionych stanów zwiększa ryzyko wystąpienia incydentu wieńcowego i nagłego zgonu.

W niniejszej pracy udokumentowano napadowe częstoskurcze, które obserwowano w czasie aktywności dziennej, bądź w godzinach nocnych. Aktywność dzienna jest reprezentantem zwiększonej aktywności układu współczulnego, natomiast godziny nocne – to aktywność nerwu błędnego (układ przywspółczulny). Dlatego w niniejszej pracy analizowano dane z wywiadu dotyczące pory występowania omdleń. W bazie rozróżniano 4 grupy:

- a) częstoskurcze pojawiające się jedynie w ciągu dnia;
- b) częstoskurcze pojawiające się głównie w ciągu dnia, ale także w nocy;
- c) częstoskurcze pojawiające się głównie w nocy, ale także w ciągu dnia;
- d) częstoskurcze pojawiające się jedynie w nocy.

Następnie dokonano podziału pacjentów na dwie grupy:

- I. grupa „dzień” - kołatania w tej grupie pojawiały się tylko w ciągu dnia, można więc opisać tę grupę jako poddaną widocznemu działaniu układu współczulnego;
- II. grupa „noc” - byli to pacjenci, u których kołatania pojawiały się nocą bądź zarówno w dzień, jak i w nocy; osoby te stanowiły grupę bez widocznej przewagi działania układu współczulnego.



### 3.2.4.4 Częstoskurcz

#### (1) Częstotliwość HR częstoskurczu

Częstotliwość rytmu serca (HR, ang. heart rate) określa liczbę cykli częstoskurczu występujących w jednostce czasu (tu: minuta). Jest zasadniczym mechanizmem modelującym pracę układu sercowo-naczyniowego. Ponieważ jest pochodną aktywności komórek węzła zatokowego, podlega wpływowi m.in. układu autonomicznego, niektórym hormonom i miejscowemu stężeniu jonów. [27]

W warunkach spoczynkowych u dorosłego człowieka częstotliwość rytmu serca (nazywana też tętnem czy pulsem) wynosi przeciętnie 60–90 uderzeń na minutę. Oczywiście pod wpływem silnych bodźców stresowych, takich jak lęk, radość, itp. może następować zmiana częstotliwości akcji serca. Napadowy częstoskurcz (ang. paroxysmal tachycardia) to forma zaburzeń rytmu serca z szybką, ale miarową akcją serca. Występuje wówczas odczucie szybkiego bicia serca – nazywane kołataniem serca. Częstoskurcz rozpoznawany jest wtedy, gdy częstotliwość rytmu serca przewyższa 100 uderzeń na minutę. [2] [12]

W analizowanej próbie badawczej dla każdego chorego zawarto co najmniej dwie wartości częstotliwości rytmu serca (HR) w czasie częstoskurczu. Jako pierwszą odnotowano napadowy częstoskurcz udokumentowany w wywiadzie (EKG, Holter-EKG) z określeniem jego HR częstoskurczu. Natomiast drugą wartością jest HR wyzwolonego częstoskurczu w czasie badania EP. Dodatkowo utworzono nowe zmienne opisujące wartości standaryzowane i należne częstotliwości rytmu serca częstoskurczu.

- I. Częstoskurcz z wywiadu.
- II. Częstoskurcz z badania EP (u niektórych pacjentów HR częstoskurczu była podwyższona podaną doraźnie Isoprenaliną i.v. w celu jego wyzwolenia).
- III. Częstoskurcz standaryzowany - został utworzony, aby wyeliminować wpływ leku (Isoprenaliny i.v.) na wartość częstotliwości HR częstoskurczu. W przypadku podania leku (i.v.) podczas badania EP jest to wartość HR częstoskurczu odnotowana z wywiadu, natomiast dla pozostałych pacjentów nieotrzymujących leku i.v. w badaniu EP, jest to maksymalna wartość HR częstoskurczu z wywiadu i częstoskurczu zarejestrowanego podczas badania EP.
- IV. Należna wartość HR (przewidywana maksymalna częstotliwość rytmu serca) uwzględniająca wiek badanej osoby

Z wiekiem związane jest zwolnienie maksymalnej częstotliwości

rytmu serca rocznie o około 1/min. Podczas wykonywania próby wysiłkowej na bieżni ruchomej często stosuje się regułę teoretyczną: "220 minus wiek" [35] [36] dla oszacowania maksymalnej częstotliwości rytmu serca. Jednakże stosowanie tego wskaźnika jako ścisłej, jednoznacznej miary maksymalnego wysiłku jest ograniczone głównie ze względu na fakt, że wielu badanych stosuje leki wpływające na częstotliwość rytmu serca.

W niniejszej pracy zastosowano modyfikację cytowanego wcześniej wzoru [37] obliczającego maksymalną dla danego wieku częstotliwość rytmu serca, pozwalającą lepiej doszacować osoby starsze:

$$210 - (\text{wiek} \times 0,65)$$

V. Cztery grupy wg procentów należnej HR

- a. osoby, które nie przekroczyły 80% wartości częstoskurczu należnego;
- b. wartość częstoskurczu była w przedziale (80%;100%);
- c. przekroczone do 20%;
- d. przekroczone powyżej 20%.

(2) Długość cyklu częstoskurczu CL

Długość cyklu częstoskurczu, czyli CL (ang. cycle length), zapisana jest w ms. Jest to więc wartość opisująca częstoskurcz z większą precyzją, dokładniej niż częstotliwość HR częstoskurczu (zaokrąglona do całkowitej liczby cykli, tzn. „uderzeń”, przypadających na minutę). Dłuższy czas długości cyklu częstoskurczu oznacza krótszą jego częstotliwość HR. W pracy opisano CL w sposób analogiczny do częstotliwości HR częstoskurczu:

- I. CL częstoskurczu z wywiadu.
- II. CL częstoskurczu z badania EP (u niektórych pacjentów po podaniu Isoprenaliny i.v. czas trwania CL był skrócony).
- III. CL częstoskurczu standaryzowana - utworzona, aby wyeliminować wpływ leku (Isoprenaliny) na częstoskurcz. (W przypadku otrzymania leku podczas badania EP jest to wartość z wywiadu, natomiast dla nieotrzymujących leku, jest to minimalna wartość CL z wywiadu i z badania EP).

### **3.2.4.5 Pozostałe potencjalne czynniki ryzyka dla epizodów omdleń**

(1) Poziomy TSH, FT3 i FT4 [38]

(2) Choroby współistniejące

W pracy dokonano oceny wpływu występowania dodatkowych chorób na ryzyko wystąpienia epizodów omdleń oraz na częstotliwość HR częstoskurczu u chorych z wąskim zespołem QRS. Zastosowano dwa różne sposoby klasyfikacji pacjentów:

- (a) Choroby współwystępujące typu I - analizowano osoby z jakąkolwiek zgłoszoną w wywiadzie chorobą dodatkową wobec osób bez żadnych jawnych objawów choroby współistniejącej.
- (b) Choroby współwystępujące typu II - w tym wypadku analizowano osoby ze współistniejącymi objawami choroby, którą uznano za istotną z klinicznego punktu widzenia (podejrzenie, iż chorzy mogą doświadczać epizodu omdlenia częściej, nawet bez częstoskurczu) wobec osób, u których nie stwierdzono objawów tychże chorób. Wybrano cztery następujące grupy schorzeń:
  - (i) cukrzyca [39] [40]
  - (ii) nadciśnienie tętnicze [41] [42] [43]
  - (iii) osoby po zawale serca [44] [45] [46]
  - (iv) osoby z dławicą piersiową [47]

W analizowanej grupie pacjentów z cukrzycą u wszystkich poziom cukru był wyrównany oraz wszyscy prowadzili regularne leczenie. Podobnie pacjenci z nadciśnieniem tętniczym - stosowali regularnie leki przeciwnadciśnieniowe. Wśród pacjentów z dławicą piersiową i/lub po zawale serca nie obserwowano nawracających epizodów bólu wieńcowego, pacjenci prezentowali stabilny stan kliniczny, regularnie stosowali niezbędne leki.

(3) Czynniki biochemiczne

W pracy przeanalizowano poziomy sodu i potasu. [48] [49]

## **3.3 Metody**

Analiza czynników ryzyka omdleń wiąże się z rozpatrywaniem sytuacji, w której zmienną objaśnianą jest zmienna dychotomiczna, określająca występowanie bądź brak występowania omdlenia. [50] [51] [52] Do oceny omawianego problemu klinicznego można zastosować wiele metod zarówno statystycznych, jak i eksploracji danych.

Wstępna statystyka opisowa przeprowadzona została z zastosowaniem

testów parametrycznych i nieparametrycznych porównujących dwie grupy oraz testów Chi-kwadrat dla przeanalizowania związków pomiędzy cechami nominalnymi. [53] [54] Cechy charakteryzujące wielkość HR częstoskurczu przeanalizowano używając regresji wielorakiej. [55]

Ze względu na łatwą interpretację współczynników modelu bardzo popularne w badaniach medycznych stały się metody regresji logistycznej. [20] W niniejszej pracy wykorzystano wielowymiarowe modele logistyczne w celu zbadania bardziej złożonych efektów np. interakcyjnych i mediacyjnych oraz oceny zmiennych wnikających. Jakość modeli oceniana była na podstawie statystyki testu Hosmera-Lemeshowa oraz współczynników determinacji, istotność statystyczna całego modelu sprawdzano za pomocą testu ilorazu wiarygodności (test LR ang. likelihood ratio), natomiast istotność statystyczna konkretnej zmiennej w modelu na podstawie testu Walda. [56] Testy te pozwalają wybrać ten model, który lepiej obrazuje związki zachodzące w rzeczywistości. Przydatność klasyfikacyjną modelu oceniono także poprzez ocenę istotności statystycznej dla pola pod krzywą ROC (ang. Receiver Operating Characteristic). [57] [58]

W celu ograniczenia zbyt dużej liczby zmiennych branych pod uwagę przy budowie modeli utworzono nowe zmienne na bazie istniejących za pomocą analizy czynnikowej. [52] Tym samym została zredukowana wymiarowość badanych modeli. Decyzję co do liczby czynników, podjęto na podstawie wartości własnych wielkości wariancji danego czynnika. Następnie przeanalizowano modele logistyczne utworzone dla nowych zmiennych będących kombinacją liniową pierwotnie rozpatrywanych cech.

Dla wybranych modeli przeanalizowano testy dokładne (exact), które wymagają dużej mocy obliczeniowej. [59] Główną zaletą takich obliczeń jest fakt, iż dokładna istotność jest zawsze wiarygodna, bez względu na wielkość, rozkład, licznosc lub stopień zrównoważenia danych. Znacznie szybciej estymatory parametrów modelu można uzyskać metodami symulacji Monte Carlo (MC). [60] W przeciwieństwie do metod dokładnych, nie oszacowują one wartości p-value na podstawie wszystkich możliwych elementów zestawu referencyjnego. Wykorzystują za to losowy podzbiór tych elementów, pobrany zgodnie z rozkładem prawdopodobieństwa charakteryzującym badane zmienne. Metody te zostały w niniejszej pracy wybrane jako najwłaściwsze oszacowanie analizowanego problemu medycznego, ze względu na ich możliwości rozwiązywania problemów deterministycznych poprzez przekształcanie ich do problemów probabilistycznych.

Najbardziej praktycznym rozwiązaniem eksploracji danych wydaje się z kolei schemat blokowego postępowania np. w postaci drzewa decyzyjnego. [23]

Najogólniej, celem analizy z zastosowaniem algorytmu budowy drzew, jest znalezienie zbioru logicznych warunków podziału, typu: „jeżeli..., to...”, prowadzących do jednoznacznego zaklasyfikowania obiektów. W pracy zastosowano drzewa interakcyjne typu C&RT/CHAID (Chi-squared Automatic Interaction Detector - automatyczny detektor interakcji za pomocą Chi-kwadrat), które to opierają się na teście Chi-kwadrat jako kryterium wyznaczenia kolejnego najlepszego podziału w każdym kroku. Jednym z podejść do weryfikacji drzew klasyfikacyjnych, zastosowanym w niniejszej pracy, jest tzw. sprawdzian krzyżowy (inaczej walidacja krzyżowa lub kroswalidacja) polegający na zastosowaniu drzewa utworzonego na podstawie jednego zbioru obserwacji (próby uczącej) do innego, niezależnego zbioru obserwacji (próba testowa). Często stosuje się dodatkowo tzw. k-krotną walidację, gdzie oryginalna próba jest dzielona na k podzbiorów. [21] Każdy z nich brany jest kolejno jako zbiór testowy, a pozostałe (razem) jako zbiór uczący. Analiza wykonywana jest więc k razy, a następnie k rezultatów uśrednia się w celu uzyskania jednego, wspólnego wyniku. Kroswalidacja pomaga ocenić, czy model będzie dobrze działał dla danych, które nie były wykorzystywane do jego konstruowania (czy nie wystąpi zjawisko tzw. przeuczenia, ang. overfitting).

W niniejszej pracy utworzono również tzw. modele punktacji (modele scoringowe) [61], które okazują się bardzo przydatne w analizie danych medycznych. Jedną z zalet takiego modelu klasyfikacyjnego (oprócz przejrzystej wstępnej analizy danych – np. dyskretyzacji) jest możliwość wytworzenia skali (oceny) jako wyniku działania modelu. Ocena ta wyraża prawdopodobieństwo lub szansę wystąpienia analizowanego zdarzenia. [62] Jako metodę analityczną dla scoringu wybrano regresję logistyczną.

Z uwagi na duże niezbalansowanie analizowanych danych oraz chęć poprawnego rozpoznawania klasy mniejszościowej (osób zagrożonych wystąpieniem epizodu omdlenia), przeprowadzono również analizę omawianego problemu z wykorzystaniem teorii zbiorów przybliżonych. [63] Ponadto podjęto również próbę rozwiązania tego problemu poprzez zastosowanie algorytmów nadlosowania klasy mniejszościowej (ang. oversampling). [64]

## 3.4 Opis metod

### 3.4.1 Zastosowanie wstępnej analizy danych

Ponieważ poprawne zbudowanie analitycznych modeli w dużym stopniu uzależnione jest od jakości i rzetelności zebranych danych, w niniejszej pracy utworzono także modele logistyczne po zastosowaniu wcześniejszej wstępnej analizy predyktorów (zmiennych objaśniających). Były to następujące metody

- 1) wybór predyktorów
- 2) wybór wiązki reprezentantów
- 3) dyskretyzacja zmiennych

Przed przystąpieniem do dalszych analiz wyliczono tzw. „wagę dowodu” (WoE ang. Weight of Evidence) dla predyktorów. [65] Miara ta jest wyznacznikiem mocy predykcyjnej poszczególnych wartości lub przedziałów wartości zmiennych prognozujących obliczanym na podstawie OR (ilorazu szans). Wyższe wartości tej miary informują o większym prawdopodobieństwie wystąpienia pozytywnego zjawiska. (Klasa negatywna, to wystąpienie epizodu omdlenia.) Wartość WoE będzie równa 0, jeżeli iloraz szans wynosi 1:

$$WoE = \left[ \ln \left( \frac{\%good_i}{\%bad_i} \right) \right] * 100$$

$\%good_i$  - procent przypadków klasy pozytywnej w przedziale wyników  $i$ .

$\%bad_i$  - procent przypadków klasy negatywnej w przedziale wyników  $i$ .

Dzięki zastosowaniu „wagi dowodu” (WoE) dla zakodowania różnych predyktorów, są one kodowane w tej samej skali, więc można bezpośrednio porównywać parametry liniowego równania regresji logistycznej. [62]

#### 3.4.1.1 Wybór predyktorów

Jest to eliminacja cech diagnostycznych, które są nieistotne z punktu widzenia wpływu na analizowaną zmienną zależną (tu: omdlenie).

Miarą umożliwiającą wykonanie rankingu predyktorów jest moc informacyjna zmiennej (IV, ang. Information Value), czyli bezwzględna suma wartości WoE dla wszystkich  $k$  przedziałów analizowanej zmiennej. [62]

$$IV = \sum_{i=1}^k [(\%good_i - \%bad_i) * WoE]$$

$\%good_i$  - procent przypadków klasy pozytywnej w przedziale wyników  $i$ .

$\%bad_i$  - procent przypadków klasy negatywnej w przedziale wyników  $i$ .

Metoda WoE pozwala ocenić tzw. „dobroć podziału” i siłę predykcijną każdego z atrybutów badanej cechy, natomiast IV bada siłę predykcijną całej zmiennej. Im bardziej oddalone są od siebie rozkłady wyników klasy negatywnej i pozytywnej, tym wyższy jest współczynnik IV i tym większy jest potencjał prognostyczny rozpatrywanej zmiennej.

W modelu staramy się uwzględnić zmienne o wysokim współczynniku IV, pamiętając jednocześnie o tym, aby ten współczynnik nie był zbyt wysoki, bo taka zmienna może zdominować model i niesie ryzyko spadku stabilności modelu. Wykres WoE dla zmiennej powinien układać się w logiczny trend. Kategoryzacja powinna dać wystarczające udziały procentowe dla każdego przedziału.

Oprócz miary IV można wykorzystywać także miarę V Cramera oraz Gini’ego, na podstawie których również można ograniczyć wejściowy zbiór danych jedynie do zmiennych istotnie wpływających na badane zdarzenie. (Współczynnik Gini’ego służy do pomiaru przewagi karty punktowej nad systemem losowo podejmowanych decyzji, a współczynnik V Cramera pokazuje miarę siły związku pomiędzy zmiennymi.) [61]

W niniejszej pracy ograniczono się do wyznaczania współczynnika IV i zastosowano następującą jego interpretację:

- $> 0,3$  silna moc predykcyjna
- $<0,1;0,3$  średnia moc predykcyjna
- $<0,02;0,1$  słaba moc predykcyjna

#### **3.4.1.2 Wybór reprezentantów - tworzenie wiązki zmiennych**

Wybór reprezentantów [62], poszukiwanie cech nadmiernie ze sobą skorelowanych, w modelach scoringowych tożsamy jest z analizą składowych głównych, która została omówiona szczegółowo w rozdziale 3.5.3.3. Analiza ta przeprowadzana jest z rotacją (varimax znormalizowana) i powoduje wyodrębnienie niezależnych czynników (wymiarów, wiązek) zmienności oraz przypisanie do tych czynników zmiennych, które są z nimi najmocniej skorelowane. Dzięki temu analizowane zmienne pogrupowane zostaną w wiązki podobnych (w sensie korelacji) zmiennych, które zostaną przypisane do odpowiedniego czynnika. Korelację pomiędzy wyodrębnionym czynnikiem a pierwotną zmienną nazywamy ładunkiem. [61] W pracy ustalono, iż ładunki przekraczające poziom 0,7 określać będą reprezentantów danego czynnika.

#### **3.4.1.3 Dyskretyzacja zmiennych**

Istnieje wiele algorytmów pozwalających realizować proces dyskretyzacji w sposób automatyczny. Dyskretyzację można przeprowadzać uwzględniając

wartości atrybutu decyzyjnego (metody z nadzorem) lub ich nie uwzględniając (metody bez nadzoru). Można stosować np. podział dziedziny atrybutu na przedziały o równej szerokości, czy na przedziały o porównywalnej liczbie obserwacji. W niniejszej pracy korzystano z dyskretyzacji „z nadzorem”, tworzącej grupy rekordów o podobnej wartości zmiennej celu.

Kategorie mające bardzo zbliżoną wartość „wagi dowodu” (WoE) można łączyć do wspólnej kategorii. Przyjmuje się, że drobne różnice w wartościach WoE mogą wynikać z niedoskonałości analizowanej próby, a nie z rzeczywistych zmian wpływu na ryzyko. Takie wyróżnienie grup jednorodnych ze względu na ryzyko wystąpienia omdlenia wykonywano w celu przygotowania zmiennych pochodnych, które następnie były wykorzystywane do finalnej analizy. Jest to również naturalny sposób obsługi braków danych pozwalający lepiej zrozumieć charakter analizowanych zmiennych, wygładzić szумы informacyjne oraz wyeliminować wpływ obserwacji odstających. [65]

### 3.4.2 Regresja liniowa i logistyczna

Metodą powszechnie wykorzystywaną w celu przybliżenia zależności pomiędzy ciągłą zmienną celu a zbiorem zmiennych objaśniających jest **regresja liniowa**. [66] Wartość oczekiwana  $Y$  wyraża się przez liniową kombinację zmiennych  $X$ , które są liniowo niezależne i mają rozkład zgodny z rozkładem normalnym.

$$y = a_0 + \sum_{i=1}^k a_i x_i$$

gdzie:

$y$  - zmienna zależna,

$x_1, x_2, \dots, x_k$  - zmienne niezależne; wpływy poszczególnych zmiennych  $x_i$  na zmienną zależną są addytywne,

$a_1, a_2, \dots, a_k$  - współczynniki (nieznane parametry modelu) estymowane metodą najmniejszych kwadratów w oparciu o obserwowane zmienne niezależne oraz o zmienną zależną

Liczebność grupy ( $n$ ) musi być dostatecznie duża [55], co oznacza, że dla  $k$  parametrów musi być spełniony warunek, iż liczba obserwacji musi być większa lub równa liczbie szacowanych parametrów modelu:

$$n \geq k + 1$$

Podstawowym narzędziem szacującym istotność wszystkich zmiennych w modelu jest test analizy wariancji (test F). Do oszacowania czy zmienna niezależna, dla której wyznaczono współczynnik, wywiera istotny wpływ na zmienną zależną służy test t-Studenta. Błąd standardowy estymacji (SE) jest miarą, która opiera się na resztach modelu, czyli rozbieżności pomiędzy



rzeczywistymi wartościami zmiennej zależnej w próbie a wartościami zmiennej zależnej wyliczonej na podstawie zbudowanego modelu. Model dobrze dopasowany powinien mieć jak najmniejszy błąd SE. Na podstawie współczynnika determinacji wielorakiej  $R^2$  wylicza się jego wartość poprawioną ( $R^2$  skorygowane), która uwzględnia ilość zmiennych w modelu i licznosc próby. Skorygowana wartość  $R^2$  określa, jaki procent zmienności zmiennej zależnej (objaśnianej) jest wyjaśniany za pomocą zmiennej niezależnej, bądź modelu statystycznego. [67]

W medycynie dużo częściej analizuje się wartości zmiennych objaśniających, które wskazują na wystąpienie lub brak wystąpienia pewnego zdarzenia, które chcemy prognozować. Jedną z metod regresji stosowanych w takich przypadkach gdy zmienna objaśniana jest na skali dychotomicznej, tzn. przyjmuje tylko dwie wartości, jest **regresja logistyczna** (LR – ang. logistic regression). [68] Wiąże ona w prosty sposób parametry modelu z ilorazem szans (OR, ang. odds ratio), który jest intuicyjny i łatwy w interpretacji.

Szansa (ang. odds) jest to funkcja oparta na wartości prawdopodobieństwa. Zamiast wyliczania klasycznego prawdopodobieństwa, czyli stosunku liczby sukcesów do liczby wszystkich prób, wyliczamy stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki. Niech  $o(A)$  oznacza szansę wystąpienia zdarzenia A oraz  $p(A)$  prawdopodobieństwo sukcesu. [56] Wtedy:

$$o(A) = \frac{p(A)}{1-p(A)} \quad \text{oraz:} \quad p(A) = \frac{o(A)}{1+o(A)}$$

Iloraz szans (ang. odds ratio) stosuje się w przypadku porównywania dwóch klas obserwacji. Jest to iloraz szans, że dane zdarzenie zajdzie w pierwszej grupie elementów oraz że zajdzie ono również w drugiej. Opisane jest wzorem:

$$OR = \frac{o_1(A)}{o_2(A)}$$

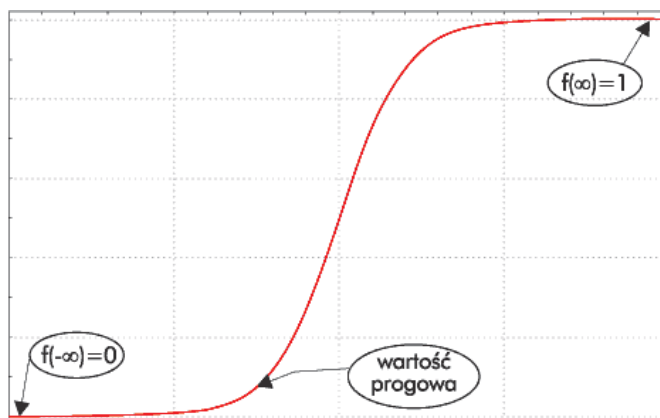
Prawdopodobieństwo zdarzenia  $p \in (0, 1)$ , więc szansa przyjmuje wartości z zakresu  $(0, \infty)$ , a jej logarytm wartości z zakresu  $(-\infty, \infty)$ . Zatem:

$$\text{logit}(p) = \ln \frac{p(A)}{1-p(A)} = a_0 + \sum_{i=1}^k a_i x_i$$

Funkcja przekształcająca prawdopodobieństwo na logarytm szansy zwana jest logitem. Funkcja logit często wybierana jest dla dwumianowego rozkładu zmiennej zależnej. Jej funkcja odwrotna opisana jest wzorem:

$$p = \frac{e^{\text{logit}(p)}}{1 + e^{\text{logit}(p)}} = \frac{1}{1 + e^{-\text{logit}(p)}}$$

Kształt funkcji przypomina rozciągniętą literę S. Funkcja logistyczna przyjmuje wartości od 0 do 1. Model może opisywać wartości prawdopodobieństwa, które są zawsze zawarte między 0 a 1 (Rycina 4).



Rycina 4. Postać funkcji logistycznej (źródło: [68])

Zmiany wartości funkcji są minimalne, jeśli wartości zmiennych są mniejsze od pewnej wartości progowej. Natomiast, gdy próg zostanie przekroczony, wówczas wartość funkcji zaczyna gwałtownie rosnąć osiągając asymptotycznie wartość 1. [66] [68]

Model regresji logistycznej wymaga właśnie liniowej zależności między logarytmem szans a wektorem zmiennych objaśniających. [20]

$$P(Y = 1 | x_1, x_2 \dots x_k) = \frac{e^{a_0 + \sum_{i=1}^k a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^k a_i x_i}}$$

gdzie:

$y$  - dychotomiczna zmienna zależna,

$P(Y = 1 | x_1, x_2 \dots x_k)$  – warunkowe prawdopodobieństwo, że zmienna  $Y$  przyjmie wartość 1 dla wartości zmiennych niezależnych  $x_1, x_2, \dots, x_k$

$x_1, x_2, \dots, x_k$  - zmienne niezależne,

$a_1, a_2, \dots, a_k$  - współczynniki modelu LR estymowane metodą największej wiarygodności (Przedziały ufności dla oszacowań współczynników w modelu regresji logistycznej konstruuje się na podstawie statystyki testu Walda [20])

Równanie regresji logistycznej może służyć do wyznaczenia prawdopodobieństwa, że u osobnika wystąpi badane zdarzenie. Współczynnik  $a_1$  mówi nam o zmianie wartości logarytmu szansy związanej ze zmianą o jedną jednostkę czynnika opisanego przez  $x_1$ . Przechodząc z logarytmu szansy

na terminy szansy,  $e^{a_i}$  oznacza relatywną zmianę możliwości wystąpienia zdarzenia pod wpływem czynnika opisanego przez zmienną  $x_i$ . EkspONENTA poszczególnego współczynnika:  $e^{a_i}$  jest oszacowaniem jednostkowego ilorazu szans  $OR_i$

- dla  $OR_i > 1$  - czynnik opisywany przez zmienną  $x_i$  ma stymulujący wpływ na wystąpienie badanego zjawiska (zwiększona szansa wystąpienia zdarzenia, gdy  $x_i$  wzrasta o jednostkę)
- dla  $OR_i < 1$  – czynnik działa ograniczająco
- dla  $OR_i = 1$  – czynnik nie ma wpływu na opisywane zdarzenie

Ponieważ model regresji logistycznej jest dopasowany na skali logarytmicznej, wpływy poszczególnych wartości  $x_i$  na wystąpienie badanego zjawiska są multiplikatywne (są iloczynem ich oddzielnych wpływów) [53].

Formalnie model regresji logistycznej jest szczególnym przypadkiem uogólnionego modelu liniowego (GLM- ang. general linear model), w którym użyto logitu jako funkcji wiążącej. Funkcja łącząca  $g(\mu)$  opisuje związek między wartością oczekiwaną zmiennej objaśnianej  $EY = \mu$  a modelem liniowym (kombinacją liniową zmiennych objaśniających):

$$g(\mu) = a_0 + \sum_{i=1}^k a_i x_i$$

Model regresji logistycznej nie wymaga niektórych założeń koniecznych dla regresji liniowej. Wektory zmiennych objaśniających i reszty nie muszą mieć rozkładu normalnego, dopuszczalna jest heteroskedastyczność (niejednorodność wariancji). [56] Jednak konieczne jest spełnienie innych warunków:

- w danych nie może występować silna współliniowość, gdyż może to prowadzić do zwiększenia błędów estymowanych współczynników, przez co uniemożliwić poprawne wnioskowanie.
- liczebność grupy ( $n$ ) musi być dostatecznie duża, co oznacza, że dla  $k$  parametrów musi być spełniony warunek:

$$n > 10(k + 1)$$

W celu znalezienia estymatorów  $a_0, a_1, \dots, a_k$  stosujemy metodę największej wiarygodności (metoda najmniejszych kwadratów nie jest możliwa, ze względu na niespełniony warunek o stałości wariancji dla zmiennej dychotomicznej  $Y$ ). Do oceny istotności estymatorów (istotności ilorazów szans)

używa się testu Walda. Ocenia on, na podstawie współczynnika oraz jego błędu szacunku, czy zmienna niezależna, dla której ten współczynnik został oszacowany, wywiera istotny wpływ na zmienną zależną. Natomiast podstawowym narzędziem szacującym istotność wszystkich zmiennych w modelu jest test ilorazu wiarygodności.

#### **3.4.2.1 Jakość modelu**

Jakość zbudowanego modelu regresji logistycznej możemy ocenić za pomocą odpowiednika współczynnika determinacji wielorakiej  $R^2$  wyznaczanego dla liniowej regresji wielorakiej. Dla regresji logistycznej najczęściej stosuje się poprawioną wartość pseudo  $R^2$  ( $R^2$  Nagelkerka lub  $R^2$  Coxa-Snella). Innym sposobem oceny jakości zbudowanego modelu regresji logistycznej jest test Hosmera-Lemeshowa. [56] Porównuje on wartości oszacowanego prawdopodobieństwa i zaobserwowane wartości występowania badanego zjawiska (hipoteza zerowa mówi o dobrym dopasowaniu modelu).

Jakość klasyfikacyjną modelu możemy również zobrazować za pomocą krzywej ROC (ang. Receiver Operating Characteristic), a dokładniej pola pod tą krzywą: AUC (ang. Area Under Curve). [57] [58] Metoda ta została opisana w podrozdziale 3.4.7 dotyczącym sposobów porównania modeli.

#### **3.4.2.2 Wybór zmiennych niezależnych**

Zazwyczaj opisując dane zjawisko biologiczne badacz ma do czynienia z ogromną liczą zmiennych opisujących złożoność tego zdarzenia. Jest to spowodowane dużą zmiennością między- i wewnątrzsobniczą organizmów żywych. Aby dokonać właściwego wyboru zmiennych niezależnych, które mają zostać wykorzystane w modelach, potrzebne są specjalne procedury statystyczne. W analizie regresji istnieje pomocnicza procedura wyboru zmiennych niezależnych – regresja krokowa [67]. W zależności od kierunku działania, wyróżniamy dwa przeciwstawne podejścia – analiza krokowa postępująca i wsteczna. Kierunek postępujący zakłada krokowe (aż do uzyskania „najlepszego” modelu) dołączanie do listy zmiennych objaśniających, uwzględnionych w modelu, tych zmiennych, które mają najistotniejszy wpływ na zmienną zależną. Przeciwny kierunek oznacza stopniową eliminację zmiennych z modelu pełnego, w taki sposób, aby utrzymać model z najwyższą wartością współczynnika determinacji, przy zachowaniu istotności parametrów. W praktyce najczęściej się zdarza, że każda z tych procedur „wybierze” inny zestaw predyktorów w modelu regresji. Do automatycznej eliminacji zmiennych należy dołączyć modyfikację modelu zgodnie ze zdrowym rozsądkiem badacza i tzw. wiedzą ekspercką [67], gdyż różnice w wybranych zestawach pozwalają

wybrać układy optymalne, a więc np. najmniej inwazyjne, najszybsze w opracowaniu, najtańsze itp.. W niniejszej pracy wykorzystano także analizę składowych głównych jako metodę redukcji wymiaru (podrozdział 3.4.3).

#### **3.4.2.3 Modele z interakcjami**

Model regresji można rozszerzyć wprowadzając do niego sztucznie stworzone predyktory np. iloczyny dwóch lub większej liczby zmiennych objaśniających. Pozwala to na uwzględnienie tzw. interakcji pomiędzy zmiennymi, czyli zmiany siły wpływu jednej ze zmiennych (na zmienną zależną) przy różnych wartościach innej zmiennej. [55]

Rozpatrywanie interakcji jest równoważne z szukaniem zmiennych będących moderatorem zasadniczo rozpatrywanej zależności.

#### **3.4.3 Metody redukcji wymiaru**

Ze względu na dużą liczbę zmiennych, które mogą mieć pośredni wpływ na zmienną zależną (wystąpienie epizodu omdlenia), zdecydowano się na zastosowanie metod redukcji wymiaru. Metody te umożliwiają przedstawienie obserwacji w przestrzeni o zadanym wymiarze, niższym niż w przypadku oryginalnych danych. Redukcja wymiaru często jest pośrednim etapem w zagadnieniu klasyfikacji czy regresji. W określonych sytuacjach pozwala na poprawę skuteczności tych metod, zwiększa stabilność, a czasem pozwala na uwzględnienie w analizach dużej liczby zmiennych. Jest też popularnie wykorzystywaną metodą do wizualizacji wielowymiarowych zmiennych, dane są redukowane do przestrzeni dwuwymiarowej, w której już łatwo je przedstawić na wykresie.

Jedną z metod redukcji danych jest analiza czynnikowa (ang. factor analysis), która stosowana jest także do wykrywania struktury w związkach między zmiennymi (to znaczy klasyfikacji zmiennych) [69]. W analizie tej wykorzystuje się fakt, iż kilka zmiennych objaśniających może w naturalny sposób utworzyć jeden komponent (inaczej czynnik lub składnik), który będzie opisywał ten sam aspekt danych.

Istnieje wiele metod analizy czynnikowej. [52] W niniejszej pracy wykorzystano jedną z nich: analizę składowych głównych (PCA, ang. Principal Component Analysis). Przy omawianiu wyników stosowano termin „analiza czynnikowa”, mając na uwadze zarówno analizę składowych głównych, jak i analizę czynników głównych.

Niezależnie od metody, analizę czynnikową rozpoczynamy od budowy macierzy korelacji i sprawdzenia, czy możliwe jest zastosowanie analizy

czynnikowej. Jeśli otrzymamy wysoką korelację między dwoma wskaźnikami, możemy wnioskować, że są one redundantne. [56] Gdyby zmienne nie były skorelowane ze sobą nawzajem, wówczas każda zmienna stanowiłaby już odrębny czynnik i nie byłaby potrzebna analiza czynnikowa. Do weryfikacji hipotezy o tym, że współczynniki korelacji pomiędzy zmiennymi są zerowe, wykorzystano test Bartletta.

Zakładając, iż pewne grupy zmiennych losowych reprezentują zmienność tych samych czynników, czyli zmienne losowe w danej grupie są od siebie w pewnym stopniu zależne, można zredukować dużą liczbę zmiennych losowych do mniejszego zbioru. Dla każdej składowej głównej i dla każdej zmiennej, wyliczane są wagi składowych, które reprezentują korelację cząstkową pomiędzy zmienną a składową główną. Wagi te są więc wartościami z przedziału  $\langle -1; 1 \rangle$ .

Uzyskana macierz ładunków czynnikowych nie jest jedynym możliwym rozwiązaniem analizy czynnikowej. Poprzez obrót układu wzajemnie ortogonalnych osi - czynników głównych - można wygenerować nieskończenie wiele różnych macierzy ładunków. Dokonanie takiej rotacji pozwala często na takie ustalenie osi, aby odpowiadająca mu macierz ładunków zapewniła możliwie najłatwiejszą interpretację czynników. [61]

Można zaproponować różne strategie rotacji. Celem wszystkich tych strategii jest uzyskanie przejrzystego układu ładunków, to znaczy czynników, które są w jakiś sposób jasno wyróżnione przez wysokie ładunki dla niektórych zmiennych i niskie ładunki dla innych. Typowe strategie rotacji to varimax, quartimax oraz equamax. [70]

Wskutek rotacji zasoby zmienności wspólnej określające udział wszystkich czynników wspólnych w wyjaśnianiu wariancji zmiennej nie ulegają zmianie. Varimax upraszcza interpretację czynników (minimalizuje liczbę zmiennych potrzebnych do wyjaśnienia danego czynnika), natomiast quartimax upraszcza interpretację zmiennych (minimalizuje liczbę czynników potrzebnych do wyjaśnienia danej zmiennej). W niniejszej pracy stosowano głównie rotację varimax poszukującą takiego zestawu ładunków czynnikowych, że każdy czynnik ma pewne ładunki bliskie zero, a pewne ładunki bliskie -1 lub +1.

Kwadrat wagi składowej określa zatem część całkowitej zmienności zmiennych, która jest wyjaśniona przez daną składową ( $r^2$ ). Aby waga składowej miała jakieś praktyczne znaczenie, powinna być wyższa niż 0,50 (aby przynajmniej 25% zmienności danej zmiennej było tłumaczone przez daną składową).

Najczęściej stosowanymi kryteriami do określania liczby składowych głównych są: kryterium wykresu osypiskowego lub kryterium części wariancji wyjaśnionej przez składowe główne.

#### ***3.4.3.1 Kryterium wartości własnej i wykresu osypiskowego***

Kryterium wykresu osypiskowego powiązane jest z pojęciem wartości własnej. Wartość własna równa 1 oznaczałaby, że składowa wyjaśnia zmienność „równoważną” jednej oryginalnej zmiennej. Dlatego racjonalnym kryterium wartości własnej jest to, aby pozostawić do dalszej analizy jedynie składowe o wartościach własnych przekraczających 1 [71]. Jest to kryterium zaproponowane przez Kaisera (1960). Graficznie jest to prezentowane na tzw. wykresie osypiskowym (metoda Cattella). Wykres ten ma charakterystyczny kształt osypiska, ponieważ prezentuje wartości własne poszczególnych składowych, przy czym pierwsza składowa wyjaśnia zwykle najwięcej zmienności, a kolejne stopniowo coraz mniej. Należy znaleźć miejsce na wykresie - tzw. koniec osypiska, czyli koniec osypywania się informacji o zmiennych pierwotnych, jaką niosą składowe główne. Jest to moment, kiedy linia malejąca przechodzi w poziomą. Wówczas składowe znajdujące się na prawo od punktu kończącego osypisko reprezentują znikomą wariancję i przedstawiają w większości losowy szum informacyjny.

#### ***3.4.3.2 Kryterium części wariancji wyjaśnionej przez składowe główne***

W tym wypadku to badacz określa procent zmienności, która ma zostać wyjaśniona przez składowe główne i wybiera tyle kolejnych składowych, aż zostanie osiągnięta minimalna wartość wyjaśnionej zmienności, a żaden następny czynnik nie tłumaczy więcej niż 5% wariancji. W przypadku niniejszej pracy, składowe główne mają być wykorzystane jako zmienne zastępujące dla oryginalnego zbioru danych i służyć w kolejnych modelach do dalszego wnioskowania. Z tego powodu część wyjaśnionej zmienności powinna być maksymalnie duża, przyjęto więc próg minimum 70%.

Stosowanie kryterium Kaisera sprawia, że czasami zostaje zbyt wiele czynników, podczas gdy technika osypiska może pozostawiać ich zbyt mało. [71] [70] W praktyce, dodatkowym ważnym aspektem jest to, na ile rozwiązanie poddaje się interpretacji. Dlatego zazwyczaj bada się kilka rozwiązań z większą lub mniejszą liczbą czynników, a następnie wybiera się to, które wydaje się najbardziej "sensowne".

### **3.4.3.3 Idea analizy składowych głównych**

Celem metody składowych głównych jest znalezienie transformacji układu współrzędnych, która lepiej opisze zmienność pomiędzy obserwacjami. Wyodrębnianie składowych głównych jest równoznaczne z rotacją maksymalizującą wariancję (varimax) wyjściowej przestrzeni zmiennych. [52] Po znalezieniu takiej linii, w przypadku której wariancja jest maksymalna, wokół tej linii pozostanie trochę zmienności. W analizie składowych głównych po wyodrębnieniu pierwszego czynnika, to znaczy po nałożeniu na dane pierwszej linii, możemy dalej definiować inną linię, która maksymalizuje pozostałą zmienność itd. W ten sposób wyodrębnia się kolejne czynniki. Ponieważ każdy następny czynnik jest definiowany tak, aby maksymalizować zmienność, która nie została objęta przez poprzedni czynnik, kolejne czynniki są od siebie niezależne. Innymi słowy kolejne czynniki są nieistotnie skorelowane. Analiza składowych głównych poszukuje takiego ortogonalnego przekształcenia, które opisze całkowitą zmienność systemu. Jej celem jest wyjaśnienie zmienności, a nie budowa modelu. Budowa modelu oparta na odnalezionych składowych jest natomiast celem analizy czynnikowej.

### **3.4.4 Testy dokładne (exact)**

Testy statystyczne bazują na rozkładach teoretycznych. Aby można było je stosować wymagana jest odpowiednio duża liczność próby, tabele powinny być gęsto wypełnione i dobrze zrównoważone. Jest to spowodowane tym, iż statystyka testowa przyjmuje odpowiedni teoretyczny rozkład asymptotycznie, tzn., gdy dysponujemy wystarczająco dużą próbą. Wyniki asymptotyczne uzyskane z małych zbiorów danych lub z rzadkich czy niezrównoważonych tabel, mogą być błędne. Za pomocą testów dokładnych możliwe jest uzyskanie dokładnego poziomu istotności bez konieczności przyjmowania założeń, których dane mogą nie spełniać. Dokładna istotność jest zawsze wiarygodna, bez względu na wielkość, rozkład, licznosc lub stopień zrównoważenia danych [54].

Testy dokładne wymagają dużej liczby obliczeń, bardzo czasochłonnych i niezwykle obciążających dla pamięci podręcznej. [59] Zazwyczaj mogą być szybko wykonywane jedynie dla prób o wielkości poniżej 30. Z uwagi na powyższe ograniczenia tylko część analiz zastosowanych w niniejszej pracy została przeliczona testami dokładnymi (exact).



### 3.4.5 Metoda Monte Carlo

Z uwagi na duże ograniczenia związane ze stosowaniem testów dokładnych (exact), w niniejszej pracy poszukiwano najwłaściwszego sposobu oszacowania analizowanego problemu medycznego.

Metoda Monte Carlo (MC) jest niezwykle pożyteczną metodą umożliwiającą rozwiązywanie problemów deterministycznych poprzez przekształcanie ich do problemów probabilistycznych. Jej istotą jest losowanie przypadkowych wielkości charakteryzujących model. Metoda Monte Carlo jest stosowana do modelowania matematycznych procesów zbyt złożonych (obliczania całek, łańcuchów procesów statystycznych), aby można było przewidzieć ich wyniki za pomocą podejścia analitycznego. Istotną rolę w metodzie MC odgrywa losowanie (wybór przypadkowy) wielkości charakteryzujących proces, przy czym losowanie dokonywane jest zgodnie z rozkładem, który musi być znany. W ostatnich dziesięcioleciach obserwujemy dynamiczny wzrost wykorzystania metod MC. Stało się to możliwe ze względu na znaczne zwiększenie się możliwości obliczeniowych komputerów.

Z matematycznego punktu widzenia, w sensie formalnym, rachunki metody Monte Carlo równoważne są całkowaniu, przy czym głównym zadaniem jest estymacja wartości oczekiwanej pewnej zmiennej losowej. [60]

Zgodnie z prawem wielkich liczb, przy odpowiednio dużej liczbie losowań, z prawdopodobieństwem dostatecznie bliskim jedności, szukana wartość jest w przybliżeniu równa średniej arytmetycznej symulowanych wartości.

Wyniki metody Monte Carlo można otrzymać znacznie szybciej niż wyniki metody dokładnej. [59] Estymator Monte Carlo, to estymator nieobciążony dokładnego poziomu istotności, liczony metodą powtarzalnego losowania z referencyjnego zestawu tabel o takich samych wymiarach i prawdopodobieństwach brzegowych jak tabela zawierająca obserwacje. Metoda Monte Carlo pozwala na szacowanie dokładnego poziomu istotności bez przyjmowania założeń, wymaganych dla metod asymptotycznych. Dodatkowo umożliwia oznaczenie przedziałów ufności dla estymowanych wielkości. Metoda ta jest użyteczna, gdy zbiór danych jest zbyt duży, aby obliczać dokładną istotność, lecz dane nie spełniają założeń metody asymptotycznej. Należy jednak pamiętać, iż metoda deterministyczna produkuje zawsze ten sam wynik, metoda niedeterministyczna natomiast, taka jak Monte Carlo, produkuje różne wyniki zależnie od konkretnych realizacji zmiennych losowych. Błąd standardowy metody związany jest z liczebnością losowanej próbki. Aby zmniejszyć błąd standardowy  $k$  razy, liczebność próby musimy zwiększyć  $k^2$  razy. [60]

### 3.4.6 Eksploracja danych (data mining)

Oprócz technik eksploracji reprezentujących podejście czysto statystyczne, istnieją także metody oparte na sztucznej inteligencji i teorii przetwarzania sygnałów, przeznaczone głównie do badania dużych zasobów danych. [22] Techniki te wraz z drzewami regresyjno-klasyfikacyjnymi (zwanymi również drzewami decyzyjnymi) składają się na metodykę badawczą określaną jako „data mining” - czyli metody „drażenia” lub „zglobiania” danych. W przypadku tych metod ocena modelu nie dokonuje się na podstawie istotności statystycznej lecz na podstawie poprawności na zbiorze testowym. To dane, a nie „a priori” sformułowane hipotezy, stanowią punkt wyjścia przy budowie modelu. Podejście to znajduje zastosowanie przede wszystkim w tych obszarach badań empirycznych, gdzie z różnych względów nie jest możliwe przeprowadzanie ściśle zaplanowanych i dobrze kontrolowanych eksperymentów (np. ze względu na koszty lub powody etyczne) lub występuje brak wystarczająco uzasadnionych teorii, czy też złożoność zjawisk jest zbyt duża. Obszarów stosowania eksploracji danych jest oczywiście wiele. Również w przypadku baz medycznych eksploracja danych znajduje obecnie coraz większe uznanie.

Eksploracja danych (data mining) jest stosunkowo nową dziedziną powstałą u zbiegu kilku różnych dyscyplin naukowych. Jej rozwój stymulowany był gwałtownym wzrostem wielkich baz danych i rosnącą potrzebą odkrywania ukrytych w nich prawidłowości [72] [73]. W dzisiejszym świecie, ze względu na wielki postęp technologiczny w zakresie cyfrowego generowania i gromadzenia informacji, systemy gromadzące dane stosuje się prawie we wszystkich dziedzinach życia. Ogromne zbiory danych gromadzone są w tzw. hurtowniach danych. Oczywiście pojawiła się potrzeba analizy tych danych w celu odkrycia nieznanej dotąd wiedzy. Data mining, jako proces automatycznego lub półautomatycznego „zglobiania” danych, postrzegany jest jako szansa na relatywnie tanie dokonywanie odkryć naukowych, przez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych [74]. W przypadku wielkich baz danych nie byłoby to możliwe bez wykorzystania szybkości komputerów.

Metody eksploracji danych są grupą bardzo zróżnicowaną i stale wzbogacaną. Wywodzą się one ze statystyki, informatyki, analizy sygnałów, matematyki, grafiki komputerowej. Uzyskiwane rezultaty takich analiz mogą być reprezentowane i przechowywane w różnych formach. W dziedzinach uczenia maszynowego i sztucznej inteligencji dla potrzeb przechowywania wiedzy stosowane są struktury takie jak: sieci neuronowe, drzewa decyzyjne, sieci

semantyczne, proste i złożone reguły logiczne. Nie wszystkie z tych struktur spełniają wymagania narzucane przez problematykę odkrywania wiedzy ze zbiorów danych. Podstawowym wymogiem jest prostota opisu i czytelność reprezentowanej wiedzy dla badacza, na tym polega ich praktyczna użyteczność. Najpopularniejszymi i najprostszymi w interpretacji metodami reprezentacji wiedzy odkrywanej w bazach danych, są drzewa decyzyjne i reguły logiczne.

Wstępny proces eksploracji danych to przygotowanie danych obejmujące czyszczenie i przekształcanie danych oraz zredukowanie liczby analizowanych zmiennych do poziomu pozwalającego efektywnie wykonywać analizy (poziom ten zależy od stosowanych metod data mining). Przy doborze zmiennych wybiera się podzbiór predyktorów z dużej listy kandydatów zakładając, że związki pomiędzy predyktorami a zmiennymi zależnymi są liniowe lub monotoniczne. Po przygotowaniu danych, dalszy przebieg eksploracji zależy od konkretnego problemu, który chcemy rozwiązać. Eksploracja może obejmować bardzo różne metody, od prostego wyboru predyktorów za pomocą regresji liniowej do wyrafinowanego badania danych różnymi metodami graficznymi i statystycznymi. Do redukcji danych wykorzystuje się np. analizę skupień. (Zastosowane w niniejszej pracy metody wstępnej analizy danych opisane zostały szczegółowo w rozdziale 3.5.1)

Mimo, iż eksploracyjne metody analizy danych wymagają zazwyczaj bardzo dużej liczby obserwacji, a ponadto ich wyniki niełatwo dają się uogólniać na populację, obserwuje się stały rozwój tego typu metod, jak i coraz większy zakres praktycznych zastosowań. [71] [74]

#### ***3.4.6.1 Modele punktacji (scoringowe)***

Modele punktacji, nazywane dalej modelami scoringowymi, to modele klasyfikacyjne, w których zmienna zależna przyjmuje z reguły dwa stany (albo zdarzenie zaistnieje, albo nie). Przyznają one punkty poszczególnym cechom badanego zjawiska, a następnie na podstawie otrzymanych rezultatów, konwertują wynik na prawdopodobieństwo zaistnienia badanego zdarzenia. Rezultatem ich działania jest ocena (scoring) wyrażająca prawdopodobieństwo lub szansę zajścia modelowanego zdarzenia. [62]

Do budowy modeli scoringowych można użyć zarówno metod statystycznych, jak i metod eksploracji danych. Do pierwszej grupy najczęściej wybierane są: regresja logistyczna, probit, analiza dyskryminacyjna lub model proporcjonalnego hazardu Coxa. Najpopularniejszymi metodami data mining używanymi w scoringu są z kolei: sieci neuronowe, drzewa wzmacniane

(boosted trees), losowy las (random forest) oraz metoda wektorów nośnych (support vector machines). Pierwsza grupa metod charakteryzuje się prostotą interpretacji, łatwością opisaną kierunku i siły wpływu poszczególnych cech na model. Metody data mining natomiast, działające na zasadzie „czarnej skrzynki”, są dużo bardziej skomplikowane. [61] Mają jednak większą siłę predykcyjną, gdyż umożliwiają modelowanie nieliniowych wzorców bez potrzeby dodawania nowych zmiennych reprezentujących zidentyfikowane interakcje.

Bardzo często modele scoringowe wykorzystywane są w bankowości, sektorze ubezpieczeń, czy kampaniach marketingowych. Znajdują zastosowanie przy rozwiązywaniu trudności związanych z określeniem wiarygodności klienta, czy wybieraniu klientów do akcji marketingowej w celu uzyskania najwyższej skuteczności. Można ich jednak używać również w wielu innych dziedzinach badawczych takich, jak medycyna czy meteorologia, aby oszacować prawdopodobieństwo zaistnienia analizowanego zdarzenia.

#### 3.4.6.1.1 Szukanie reguł – drzewa losowe

Jednym ze sposobów poszukiwania reguł opisujących zmienną zależną, może być klasyfikacja przy użyciu drzew losowych. [23] Moduł „Losowy las” generuje 100 losowych drzew klasyfikacyjnych z ustaloną maksymalną liczbą poziomów. Każde z tych drzew opisane jest pewną regułą, dla której wyliczono wartość wskaźnika zdarzeń niepożądanych („bad rate”), oznaczającego stosunek liczby „złych” przypadków – tu: tracących przytomność - do liczby wszystkich w danym przedziale/podziale. W przypadku małej liczebności klasy, w której występowały rozpatrywane, niepożądane zjawiska, nie uzyskujemy nigdy wysokiego „bad rate”. Uzyskujemy za to duże prawdopodobieństwo wykluczenia wystąpienia analizowanego zdarzenia.

Reguły decyzyjne opisano bardziej szczegółowo w rozdziale 3.4.6.2.2. Odpowiednikiem wskaźnika zdarzeń niepożądanych („bad rate”) jest tam dokładność reguły dla decyzji odpowiadającej wystąpieniu analizowanego, niepożądanego zdarzenia (tu: omdlenia).

Losowy las jest techniką baggingu [75] (agregacja wyników modeli budowanych w oparciu o próbki danych losowanych ze zwracaniem), opartą o technikę prostych drzew decyzyjnych. Końcowa klasyfikacja oparta jest na równoprawnym głosowaniu wszystkich drzew. W przeciwieństwie do procesu budowy pojedynczego drzewa decyzyjnego, podział w węźle realizowany jest na podstawie najlepszego atrybutu z wylosowanej każdorazowo puli atrybutów. Drzewa te nie są przycinane, co może powodować ich nadmierne dopasowanie

się do zbioru.

#### 3.4.6.1.2 Wyliczanie punktacji w modelach punktacji (scoringowych)

Modele punktacji (scoringowe), można określić jako system automatycznej i obiektywnej oceny pacjentów, wprowadzony w rezultacie doświadczenia, uzasadniony przykładami i statystykami. [62] W analizowanym zbiorze danych ocena punktowa oparta jest na zmiennych dotyczących danych zgromadzonych z wywiadu i/lub dotyczących stanu zdrowia pacjenta w przeszłości. Założeniem scoringu jest fakt, iż przyszłość jest odzwierciedleniem przeszłości.

W niniejszej pracy tablicę punktacji (scoringową) otrzymujemy w wyniku przekształcenia ocen parametrów regresji logistycznej, która została wybrana jako metoda analityczna dla eksploracji danych. Każdej kategorii zmiennych w modelu została przypisana określona liczba punktów. Wynikiem modelu jest suma punktów dla poszczególnych kategorii, do których trafiły wyniki analizowanego pacjenta.

Często, dla wygody interpretacji, przyjmuje się przedział punktowy od 0 do 100. Można wówczas wynikowi nadać interpretację procentową.

Wartość punktu odcięcia (ang. cut off) reprezentuje próg ryzyka przy klasyfikacji nowego obiektu. Jego wybór powinien być związany z minimalizacją udziału błędnych przypadków i zapewnić pożądany poziom akceptowalności wniosków. Punkt ten można odnaleźć analizując nowopowstałą zmienną punktującą: należy przeprowadzić dyskretyzację tej zmiennej, czyli odnaleźć grupy jednorodne ze względu na ryzyko wystąpienia niepożądanego zjawiska (tu: omdlenia) lub ocenić model za pomocą krzywej ROC.

### 3.4.6.2 Zbiory przybliżone

Zagadnienia dodatkowe związane z podstawą teorii zbiorów przybliżonych zostały umieszczone w Załączniku nr 2 niniejszej pracy [76].

#### 3.4.6.2.1 Teoria budowania reguł metodą zbiorów przybliżonych

Algorytmy oparte na teorii zbiorów przybliżonych (RS ang. rough sets) należą do rodziny algorytmów budujących wiedzę w postaci reguł. Teoria ta [77] skupiona jest na obsłudze sprzeczności w danych, które mogą pojawiać się w analizowanym zbiorze danych. Takie sytuacje są niezwykle częste w przypadku danych medycznych ze względu na dużą zmienność między- i wewnątrzsobniczą organizmów żywych.

Teoria zbiorów przybliżonych proponuje nowe spojrzenie na wiedzę i jej reprezentację, u podstaw którego leży przekonanie, iż wiedza to zdolność do klasyfikacji. Przez klasyfikację rozumie się zdolność do odróżniania obiektów - elementów otaczającej rzeczywistości.

W teorii zbiorów przybliżonych, do zdefiniowania wiedzy mamy pewien zbiór obiektów, które chcemy klasyfikować oraz zbiór cech, które te obiekty opisują.

W niniejszej pracy ustalono klasyfikację na omawianym zbiorze pacjentów ze względu na wystąpienie (lub nie) omdlenia. Podział analizowanego zbioru będzie się więc składał z dwóch pojęć.

Zbiory przybliżone analizują własności nie pojedynczych obiektów, ale całych klas obiektów, które są wzajemnie nierozróżnialne lub podobne. Ma to znaczenie zwłaszcza tam, gdzie pojęcia nie mogą być zdefiniowane w sposób jednoznaczny, a można jedynie podać przybliżone definicje pojęć.

Zbiór jest przybliżony (ang. rough set), jeśli obszar brzegowy jest niepusty, w przeciwnym przypadku zbiór jest nazwany dokładnym (ang. crisp set). [77]

W teorii zbiorów przybliżonych wprowadza się pojęcie reduktu, który jest najmniejszym zestawem atrybutów warunkowych prowadzących do klasyfikacji o takiej samej jakości, jaką daje klasyfikacja z uwzględnieniem wszystkich dostępnych w systemie informacyjnym predyktorów. Najmniejszy niezależny zbiór predyktorów dający taką samą jakość klasyfikacji co redukt, nazywany jest zbiorem minimalnym. System informacyjny może mieć kilka reduktów i/lub kilka zbiorów minimalnych. [76] Wspólna część wszystkich dostępnych zbiorów minimalnych (zespół predyktorów należących do wszystkich bez wyjątku zbiorów minimalnych) nazywana jest rdzeniem. Elementy należące do rdzenia nie mogą być pominięte w analizie danych bez utraty jakości klasyfikacji.

Na podstawie atrybutów warunkowych i przypisanej im wartości atrybutu decyzyjnego buduje się zdania logiczne określane jako reguły decyzyjne. Zbiór reguł decyzyjnych, określane też mianem algorytmu decyzyjnego, w maksymalnie oszczędny sposób przedstawia zależności przyczynowo skutkowe występujące w systemie informacyjnym.

Zbiory obiektów opisanych identycznymi wartościami deskryptorów nazywamy atomami. Przypadki nierozróżnialne mają silny wpływ zarówno na dokładność klasyfikacji, jak i na ostateczną strukturę reguł decyzyjnych opisujących procedurę klasyfikacyjną. Generowane reguły decyzyjne muszą być kompletne i zgodne. Oznacza to, że każdy obiekt, który w rzeczywistości należy do określonej badanej klasy decyzyjnej, musi być rozpoznany jako należący do tej klasy, natomiast każdy obiekt nienależący w rzeczywistości do badanej klasy, nie może być rozpoznany jako element tej klasy. W języku statystyki oznacza to, że liczba przypadków fałszywie dodatnich i fałszywie ujemnych musi być równa zero. [76]

#### 3.4.6.2.2 Reguły decyzyjne

Reguły decyzyjne mają postać:

$$r : (\text{Warunek}) \rightarrow (\text{Teza})$$

gdzie:

Warunek: koniunkcja testów na atrybutach

Teza: klasa decyzyjna

Zbiór reguł:  $R = \{r_1, r_2, \dots, r_n\}$

Reguła  $r$  pokryje obiekt  $x$ , jeśli wartości jego atrybutów spełniają warunek reguły, tzn. jeśli obiekt spełnia wszystkie deskrytory w  $R$ . [77] Liczba obiektów pokrytych przez regułę  $R$  określana jest mianem wsparcia reguły (ang. support).

Nie wszystkie przypadki muszą być pokryte przez wygenerowany zbiór reguł. W większości metod dany obiekt może być pokryty przez kilka reguł (reguły nie są rozłączne). Jedynie w wypadku generowania reguł z drzew decyzyjnych uzyskujemy reguły rozłączne.

Reguły decyzyjne dzieli się na dwie klasy:

- deterministyczne, gdy reguła prowadzi do jednoznacznego zakwalifikowania obiektu do danej klasy na podstawie wartości użytych w niej predyktorów
- probabilistyczne – w przeciwnym przypadku.

Tablica decyzyjna jest deterministyczna (dobrze określona, spójna), gdy wszystkie reguły w niej zawarte są deterministyczne, w przeciwnym przypadku jest niedeterministyczna.

Reguły decyzyjne generowano w niniejszej pracy za pomocą algorytmów typu LEM2 (ang. Learning from Examples Module, version 2), polegających na wyliczaniu lokalnych pokryć na podstawie przybliżeń zbiorów przybliżonych [78].

#### 3.4.6.2.3 Ocena skuteczności reguł i klasyfikatorów

Mogą pojawić się przypadki, w których obiekt, który spełnia warunek, ma decyzję równą (tzw. pozytywny przykład), bądź różną od  $d$  (tzw. negatywny przykład). [79] Dlatego do opisu reguły często dodawana jest jej dokładność (ang. accuracy ratio), wyliczaną wg wzoru:

$$\text{dokładność reguły} = \frac{\text{liczba pozytywnych przykładów reguły}}{\text{wsparcie reguły}}$$

Dokładność przybliżenia dowolnego zbioru obiektów przy użyciu określonego zbioru atrybutów warunkowych w analizowanym systemie informacyjnym określa liczba z przedziału  $\langle 0,1 \rangle$ . Jest ona wyznaczona jako stosunek liczebności zbioru dolnego przybliżenia do liczebności zbioru górnego przybliżenia. (Więcej definicji znajduje się w Załączniku nr 2)

W niniejszej pracy używano także pojęcia pewności reguły, które definiujemy w sposób następujący:

$$\text{pewność reguły} = \frac{\text{liczba pozytywnych przykładów reguły}}{\text{liczba wszystkich obiektów z decyzją } d}$$

Jakość klasyfikacji stanowi natomiast stosunek liczby wszystkich poprawnie sklasyfikowanych obiektów do liczby wszystkich obiektów w systemie. [77]

Ponieważ reguły tworzone są, by odnaleźć prawidłowości dotyczące nie tylko analizowanej grupy badawczej, ale całej populacji, należy zweryfikować tzw. zdolności predykcyjne algorytmu decyzyjnego. [63] Należy ocenić skuteczność odnalezionych reguł w klasyfikowaniu obiektów, które do tej pory nie były rozpatrywane przez system klasyfikacyjny.

Trafność klasyfikatora obliczana jest za pomocą  $V$ -krotnej walidacji. (rozdział 3.4.7)



#### 3.4.6.2.4 Problem niezrównoważonych danych

Analizowana próba, opisująca dane pacjentów ze zdiagnozowaną arytmia, zawiera ok. 20% przypadków z odnotowanymi epizodami omdleń. W takim przypadku, gdy klasa mniejszościowa (ang. minority class) zawiera wyraźnie mniej przykładów niż inne klasy, dane nazywane są niezrównoważonymi lub niezbalansowanymi (ang. imbalanced). [80] Niedoreprezentowane są zazwyczaj klasy reprezentujące obiekty wyjątkowe, na których rozpoznaniu najbardziej zależy badaczom. Standardowe miary nie uwzględniają dysproporcji pomiędzy klasami, co powoduje zwiększenie wagi klasy większościowej w ocenie skuteczności [81].

Ze względu na niezbalansowanie analizowanych danych podjęto próbę znalezienia metod regulowych, optymalnych dla wykrywania przypadków osób z epizodami omdleń w analizowanej bazie.

W niniejszej pracy zaproponowano metodę wstępnego przetwarzania danych niezrównoważonych. Takie zabiegi ingerujące w zbiór danych, mające na celu uzyskanie lepszego klasyfikatora, można podzielić [64] na dwie główne podgrupy:

- a) Nadlosowywanie (ang. oversampling) przykładów klasy mniejszościowej. Może się to odbywać przez duplikację istniejących przykładów lub tworzenie nowych, sztucznych przykładów.

Jedną z metod nadlosowywania przykładów jest metoda SMOTE - (ang. synthetic minority oversampling technique). Polega ona na dogenerowaniu nowych przykładów klasy mniejszościowej pomiędzy przykładami pierwotnymi.

- b) Usuwanie przykładów z klasy większościowej (ang. undersampling)

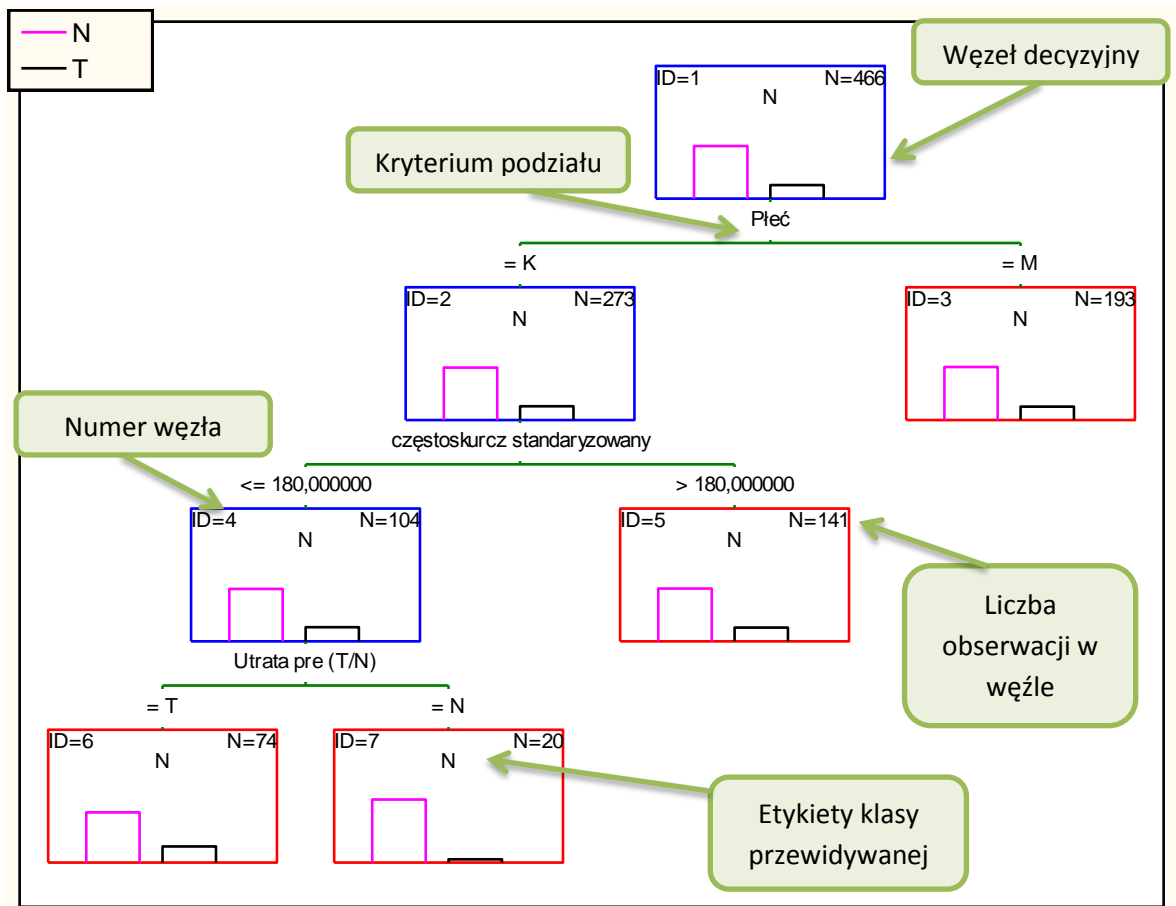
#### 3.4.6.3 Drzewa klasyfikacyjne

Drzewa klasyfikacyjne stosuje się najczęściej do rozwiązywania problemów z dużą licznością danych. Za ich pomocą wyznacza się przynależność przypadków lub obiektów do klas jakościowej zmiennej zależnej na podstawie pomiarów jednej lub wielu zmiennych objaśniających. [23] Na popularność drzew składają się prawdopodobnie: możliwość przedstawiania danych graficznie i łatwość interpretacji oraz hierarchiczna natura i elastyczność.

##### 3.4.6.3.1 Budowa drzewa.

Drzewo decyzyjne jest grafem o strukturze drzewiastej. Jest to graf spójny i acykliczny. Krawędzie takiego grafu nazywane są *gałęziami*, natomiast wierzchołki, z których wychodzi co najmniej jedna krawędź, nazywane są

węzłami. Wierzchołki niebędące węzłami to *liście* [21]. Węzeł drzewa, który nie ma żadnych węzłów macierzystych jest *korzeniem*. Drzewo „rośnie” od góry do dołu (od korzenia do liści).



Rycina 5. Przykładowe drzewo klasyfikacyjne dla zmiennej decyzyjnej „Omdlenie” przyjmującej dwa stany: TAK (T) lub NIE (N)

Celem analizy z zastosowaniem algorytmu budowy drzew jest znalezienie zbioru logicznych warunków podziału, typu „jeżeli...to...”, prowadzących do optymalnego prawidłowego zaklasyfikowania obiektów. (Rycina 5)

Zaobserwowane elementy badanej próby przechodzą wzdłuż gałęzi przez węzły. W węzłach podejmowane są decyzje o wyborze gałęzi, wzdłuż której trwa dalszy podział elementów docierających na podgrupy (względem zapisanego w nim kryterium podziału – testu). Przesuwanie trwa do momentu napotkania liścia. [21]

Dla każdego liścia istnieje dokładnie jedna droga łącząca go z korzeniem. Zbiór wszystkich dróg może być przekształcony do zbioru reguł klasyfikujących przykłady. Drzewo decyzyjne jest więc zbiorem reguł rozłącznych.

Drzewa klasyfikacyjne szczególnie narażone są na problem nadmiernego dopasowania. Przejawia się to bardzo małym błędem klasyfikacji na próbie

uczącej, lecz zbyt dużym błędem rzeczywistym. Tak stworzone drzewo, poprzez bardzo złożoną strukturę, może odzwierciedlać przypadkowe zależności występujące w zbiorze uczącym. Można uniknąć tego błędu poprzez tworzenie uproszczonych drzew, nazywane *przycięciem*. Drzewo przycięte ma prostszą strukturę, co daje krótszy czas klasyfikacji, ale konsekwencją tego działania jest pogorszenie dokładności klasyfikacji zbioru uczącego [70].

#### 3.4.6.3.2 Typy drzew.

Najważniejsze typy drzew klasyfikacyjnych stosowane w niniejszej pracy to [70]:

- a) **Drzewa typu CHAID** (ang. Chi-squared Automatic Interaction Detector, czyli automatyczny detektor interakcji za pomocą chi-kwadrat). Jest to jedna z najstarszych metod drzew klasyfikacyjnych. Buduje drzewa, w których z węzłów mogą wychodzić więcej niż dwie gałęzie. Metoda ta opiera się na teście Chi-kwadrat jako kryterium wyznaczenia kolejnego najlepszego podziału w każdym kroku. W przypadku problemów typu regresyjnego (ciągła zmienna zależna) program wylicza wartość testu F. W kolejnym kroku program wybiera predyktor o najniższej wartości poziomu p, tzn. ten, który daje najbardziej istotny podział. Jeśli wartość poziomu p (z poprawką Bonferroniego) dla każdego predyktora jest niższa niż poziom p dla podziału, to dalsze podziały nie są wykonywane i węzeł jest liściem drzewa. Proces jest powtarzany aż do momentu, gdy nie będą możliwe dalsze podziały (przy przyjętych wartościach p dla łączenia i dzielenia).
  
- b) **Drzewa typu C&RT** (ang. Classification and Regression Trees, czyli modele drzew klasyfikacyjnych i regresyjnych). W problemach klasyfikacyjnych C&RT daje możliwość wybrania różnych miar niejednorodności: Indeks Giniego lub Chi-kwadrat. Indeks Giniego „zanieczyszczenia” węzła (inaczej: niejednorodności w węźle) jest miarą najczęściej używaną w problemach typu klasyfikacyjnego. Jako miara zanieczyszczenia przyjmuje wartość zero tylko wtedy, gdy w węźle są przypadki z dokładnie jednej klasy. Miara Chi-kwadrat podobna jest do standardowej statystyki Chi-kwadrat obliczanej dla licznosci oczekiwanych i obserwowanych. W problemach typu regresyjnego automatycznie wykorzystywane jest kryterium najmniejszych kwadratów.

- c) **Drzewa interakcyjne typu C&RT/CHAID** Budowanie drzew w sposób interakcyjny wykorzystuje wiedzę doświadczonego analityka o danym zagadnieniu. W metodzie tej dokonuje się "ręcznie", interakcyjnie, wyborów w trakcie konstruowania drzewa, mających na celu uzyskanie najlepszej zdolności predykcyjnej modelu. Użytkownik może zdecydować o wykorzystaniu pewnych zmiennych oraz o tym, jakich podziałów na nich dokonać podczas tworzenia gałęzi. Umożliwia to eksperymentowanie ze zmiennymi, z różnymi podejściami do zagadnienia, pozwalając zdobyć unikalną wiedzę na temat analizowanego problemu badawczego, wynikającą z interakcji własnego doświadczenia z wiedzą uzyskaną z algorytmów. W praktyce, często łączy się podejście w pełni automatyczne budowania drzew, z wiedzą ekspercką w danej dziedzinie. Budowa niektórych części drzew odbywa się w sposób automatyczny, natomiast później otrzymany model może być dodatkowo doprecyzowany przez eksperta. Niekiedy powodem ingerencji badacza w automatyczny mechanizm tworzenia drzewa są braki danych, np. trudność w zdobyciu pewnych wyników badań, które algorytm wybrał do budowy niektórych gałęzi. Można wówczas wybrać jakąś zastępczą zmienną, taką, która w jakimś stopniu mogłaby zastąpić tę wybraną automatycznie [23].

#### 3.4.6.3.3 Sprawdzanie poprawności drzewa.

Jednym z typów testów tworzących drzewa jest sprawdzian krzyżowy. Jest to podejście polegające na zastosowaniu drzewa utworzonego na podstawie jednego zbioru obserwacji (próby uczącej) do innego, niezależnego zbioru obserwacji (próba testowa). Jeśli większość lub wszystkie podziały wyznaczone przez analizę próby uczącej powstały na podstawie losowego szumu informacyjnego, to predykcja w próbie testowej będzie bardzo słaba.

Inną metodą sprawdzania poprawności drzewa jest V-krotny sprawdzian krzyżowy. Jest on stosowany wtedy, gdy nie ma możliwości dysponowania próbą testową, a próba ucząca jest za mała, aby wyodrębnić z niej taką próbę. Badacz określa wartość V dla V-krotnego sprawdzianu krzyżowego. Wartość V wyznacza liczbę podprób losowych wyodrębnianych z próby uczącej. Podpróby powinny być możliwie równe sobie wielkością. Drzewo klasyfikacyjne określonej wielkości jest obliczane V razy. Za każdym razem opuszcza się w obliczeniach jedną z podprób i wykorzystuje się ją jako próbę testową w sprawdzaniu krzyżowym. Zatem każda podpróba jest użyta V-1 razy w próbie uczącej i tylko jeden raz w charakterze próby testowej. Następnie dla każdej z V prób

testowych obliczane są koszty sprawdzianu krzyżowego, a te są uśrednione i otrzymujemy V-krotną ocenę kosztów sprawdzianu krzyżowego, która może być podawana, razem z błędem standardowym. [21]

Niestety w dostępnych pakietach statystycznych V-krotna walidacja krzyżowa bazuje jedynie na standardowych drzewach, a nie na podziałach użytkownika, tak więc nie można jej było zastosować w przypadku drzew interakcyjnych..

### 3.4.7 Sposoby porównywania modeli

Modele regresji logistycznej oceniane są na podstawie wielkości pseudo  $R^2$  (czyli  $R^2$ Nagelkerka lub  $R^2$ Coxa-Snella) oraz wyniku testu ilorazu wiarygodności porównującego kolejno powstające (zagnieżdżone, ang. nested) modele. [20] Jeśli porównywane modele nie różnią się istotnie, wówczas wybierany jest model z mniejszą liczbą zmiennych. Brak różnicy oznacza, że zmienne, które są w modelu pełnym, a nie ma ich w modelu zredukowanym, nie wnoszą istotnej informacji. Jeśli natomiast różnica jest istotna statystycznie, oznacza to, że jeden z nich (ten z większą liczbą zmiennych, czyli o większym  $R^2$ ) jest istotnie lepszy niż drugi.

W przypadku regresji wielorakiej, również, jeśli różnica pomiędzy porównywanymi modelami jest istotna statystycznie, wówczas model pełny jest istotnie lepszy niż model zredukowany. Oznacza to, iż badana zmienna nie jest nadmiarowa, wywiera istotny wpływ na dany model i dlatego nie powinna być z niego usuwana. W tym wypadku do porównania modeli stosuje się test F. [67]

Aby porównać modele zbudowane różnymi metodami potrzebne jest wyliczenie miar oceny jakości klasyfikacji. Wymaga to zdefiniowania możliwych odpowiedzi klasyfikatora:

Tabela 1. Tabela klasyfikacji na podstawie modelu.

Klasyfikacja uzyskana z modelu	Poprawność klasyfikacji	
	+	-
+	TP	FP
-	FN	TN

- TP - true positive (liczba wyników prawdziwie dodatnich)
- TN - true negative (liczba wyników prawdziwie ujemnych)
- FP - false positive (liczba wyników fałszywie dodatnich)
- FN - false negative (liczba wyników fałszywie ujemnych)

Miary najczęściej wykorzystywane do oceny modeli [82] [83] [58]:

1. **Czułość** (ang. sensitivity, recall)

$$= \frac{TP}{TP + FN}$$

- zdolność testu do wykrycia choroby w analizowanej grupie chorych

2. **Swoistość** (ang. specificity)

$$= \frac{TN}{TN + FP}$$

- zdolność testu do wykluczenia pacjentów zdrowych

3. **Precyzja** (ang. precision) lub wartość predykcyjna dodatnia (PPV ang. positive predictive value)

$$= \frac{TP}{TP + FP}$$

- dokładność klasyfikacji w obrębie rozpoznanej klasy, czyli prawdopodobieństwo, że pacjent miał chorobę mając pozytywny wynik testu; wartość zależna od rozpowszechnienia choroby

4. **Wartość predykcyjna ujemna**

$$= \frac{TN}{FN + TN}$$

(NPV ang. negative predictive value)

– prawdopodobieństwo, że pacjent nie miał choroby mając negatywny wynik testu. Jeśli więc badana osoba otrzymała negatywny wynik testu, to NPV daje jej informację na ile może być pewna, że nie cierpi na daną chorobę; wartość zależna od rozpowszechnienia choroby

5. **Iloraz wiarygodności**

(LR ang. likelihood ratio)

$$= \frac{\frac{TP}{TP + FN}}{\frac{FP}{FP + TN}}$$

$$\text{lub} = \frac{\text{Czułość}}{1 - \text{Swoistość}}$$

- iloraz dwóch szans: szansy na to, że pozytywny wynik testu otrzyma osoba z grupy chorych do szansy, że ten sam efekt będzie obserwowany wśród osób zdrowych; niezależny od rozpowszechnienia choroby

## 6. Skuteczność

(dokładność, ang. accuracy)

$$= \frac{TP + TN}{TP + FP + FN + TN}$$

- prawdopodobieństwo prawidłowej diagnozy przy wykorzystaniu testu diagnostycznego
- miara całkowicie nieprzydatna w przypadku klas niezbalansowanych

Ponieważ w przypadku nierównowagi klas wymagany jest stosunkowo wysoki wskaźnik prawidłowej wykrywalności w klasie mniejszościowej, należy wykorzystać miary najlepiej dopasowane do tego typu problemów. Prosta predykcyjna dokładność nie jest właściwa w takiej sytuacji. Dlatego w pracy zastosowano do oceny modeli także miary odzwierciedlające dobroć klasyfikatora w obecności klas rzadkich [82]. Spośród wielu miar tworzonych typowo dla danych niezrównoważonych [84], wybrano podstawowe, najczęściej używane miary bazujące na czułości i odwrotnie proporcjonalnej do niej precyzji:

## 7. F-miara

(ang. F-measure)

$$= (1 + \beta^2) * \frac{\text{Precyzja} * \text{Czułość}}{\beta^2 * \text{Precyzja} + \text{Czułość}}$$

- średnia harmoniczna czułości i precyzji;
- można ją także ważyć w zależności od ustalonego priorytetu. Podstawienie  $\beta=1$  oznacza równą wagę czułości i precyzji, natomiast  $\beta>1$  stosuje się w przypadku konieczności nadania większej wagi czułości ( $\beta$ -razy ważniejsza jest czułość od precyzji [83]).

W niniejszej pracy wyliczono (oprócz F1) miary F2 i F3, gdyż ustalono nierówną wagę dla czułości i swoistości. Koszty błędnych decyzji FN uznano za trzykrotnie wyższe niż FP, stąd najwłaściwszą miarą używaną przy porównaniach modeli powinno być F3.

## 8. G-miara

$$= \sqrt{\text{Czułość} * \text{Swoistość}}$$

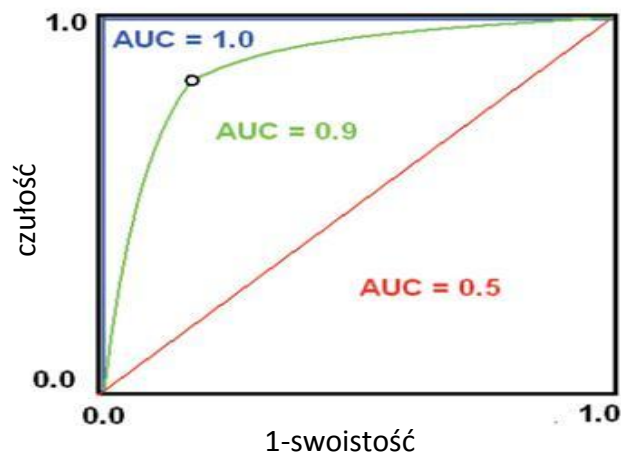
$$\text{lub} = \sqrt{\text{Czułość} * \text{Precyzja}}$$

- geometryczna średnia czułości i precyzji;
- może być rozpatrywana jako korelacja pomiędzy dwoma wskaźnikami; duża wartość tej miary wystąpi, gdy oba wskaźniki będą wysokie, natomiast będzie ona małą, gdy chociażby jeden ze wskaźników będzie posiadał małą wartość. [80]

## 9. Krzywa ROC (ang. Receiver Operating Characteristic) i pole pod krzywą (AUC ang. area under curve)

Krzywa ROC budowana jest w oparciu o wartość zmiennej zależnej oraz jej przewidywane prawdopodobieństwo (uzyskane np. w analizie regresji logistycznej lub w modelu zbudowanym inną metodą). Gdy krzywa ROC pokrywa się z przekątną  $y = x$ , to decyzja o przyporządkowaniu przypadku do wybranej klasy (+) lub (-), podejmowana na podstawie modelu, jest jednoznaczna z losowym podziałem badanych przypadków. Każdy punkt na tej krzywej ma współrzędne (1 - specyficzność, czułość). [58]

Pole pod krzywą ROC (AUC) jest miarą jakości danej metody w taki sposób, że pole 0,5 oznacza jakość klasyfikacji porównywalną do losowego rzutu monetą, zaś pole 1,0 – jest to idealna, bezbłędna klasyfikacja (Rycina 6). Jakość klasyfikacyjna modelu jest dobra, gdy krzywa znajduje się znacznie powyżej przekątnej  $y = x$ , czyli gdy pole pod krzywą ROC jest istotnie większe niż 0,5.



Rycina 6. Krzywa ROC

Oprócz zbudowania modelu równie ważne jest wybranie odpowiedniego punktu odcięcia, czyli takiej wartości  $k$  z przedziału  $(0,1)$ , że jeśli  $y < k$ , to obiekt przyporządkujemy do klasy kodowanej przez -, w przeciwnym wypadku, gdy  $y \geq k$ , do klasy kodowanej przez +. [57]

Jeśli natomiast krzywa ROC przebiega poniżej przekątnej  $y = x$ , to decyzja o przyporządkowaniu przypadku do wybranej klasy (+) lub (-), podejmowana na podstawie modelu, powinna być odwrotna, do opisanej powyżej.



Punkt odcięcia  $k$  wyznacza granicę klasyfikacji. Domyślnie jest to wartość 0,5. Badacz może zmienić tę wartość na dowolną inną wartość z przedziału (0, 1). W niniejszej pracy punkt odcięcia do klasyfikacji na podstawie modelu logistycznego ustalano na poziomie sugerowanym przez konkretną krzywą ROC (np. 0,3 w modelu A; patrz Tab.XX - Załącznik nr 1). Jeśli przyjmujemy równe koszty błędnych klasyfikacji, to optymalnym punktem odcięcia, sugerowanym przez konkretną krzywą ROC, jest punkt krzywej znajdujący się najbliżej punktu o współrzędnych (0,1). Wartość ta obliczana jest na podstawie czułości, swoistości oraz wielkości nachylenia stycznej do krzywej ROC w analizowanym punkcie. (Ważne są także koszty błędnych decyzji i współczynnik chorobowości - obliczany jako iloraz liczby osób chorych i liczby osób „narażonych” na chorobę.)

W niniejszej pracy, przy konstruowaniu krzywych ROC dla jednej zmiennej decyzyjnej, ustalono koszty błędnych decyzji FN na poziomie trzykrotnie wyższym niż FP (tzn. ustalono, iż klasyfikowanie pacjenta jako nienależącego do grupy ryzyka, podczas, gdy w rzeczywistości jest on realnie zagrożony omdleniem, jest dużo gorszym błędem niż odwrotnie). Współczynnik chorobowości wyznaczany był automatycznie z próby i wynosił ok. 0,2.

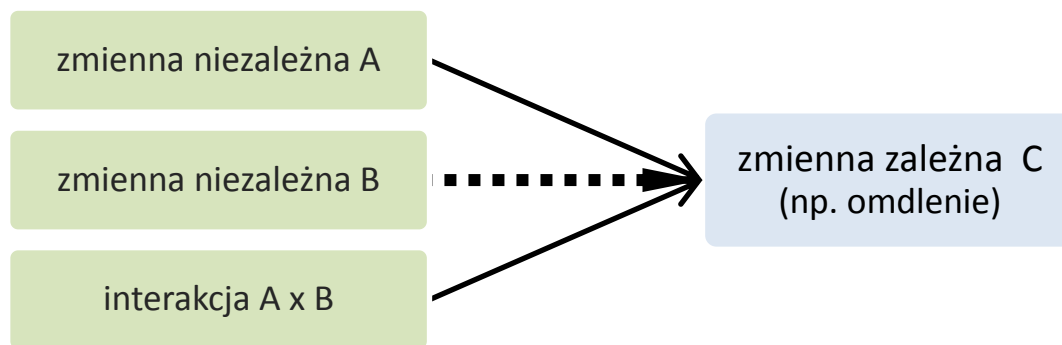
#### **3.4.8 Zmienne wikłające, interakcyjne i mediacyjne**

Z uwagi na złożoność relacji zmiennych będących potencjalnymi czynnikami występowania epizodów omdleń w analizowanej populacji, szczególną uwagę zwrócono na zmienne wikłające, interakcyjne i mediacyjne. Zmienne, mogące powodować uwikłanie wyniku, wybierane były zgodnie z intuicją badacza oraz wiedzą ekspercką dotyczącą klinicznego znaczenia poszczególnych czynników – parametrów modelu.

##### **a) Moderator**

Jeśli występuje słaba, mało istotna zależność, czasami wręcz zanikająca, poszukiwany jest tzw. moderator tej zależności. Jest to zmienna, której wartość decyduje o kierunku lub sile zasadniczej zależności [85] [86]. Obecność istotnego moderatora można sprawdzić analizując istotność interakcji.

Moderatory to takie zmienne, które opisują warunki konieczne do występowania danego efektu lub charakteryzują grupę osób szczególnie podatnych na analizowany efekt. Zazwyczaj odpowiadają na pytania: „kto?”, „kiedy?”, „w jakich warunkach?”.



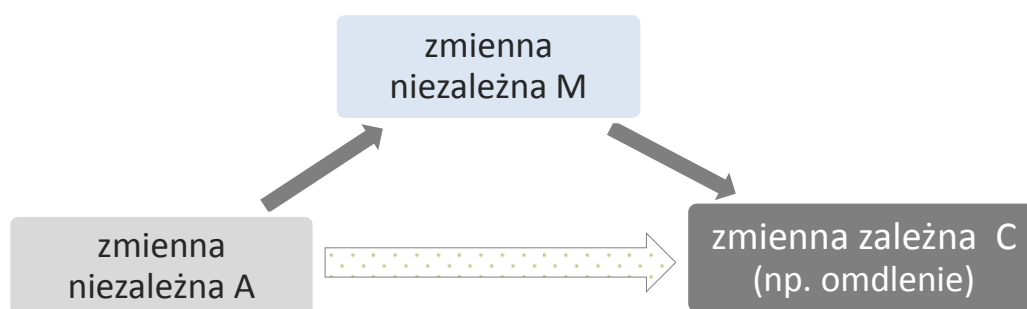
Rycina 7. Moderacja. Zmienna niezależna B jest moderatorem, tzn. wpływa na zależność zasadniczą A i C.

Moderator ma zawsze duże znaczenie praktyczne, pozwala bowiem określić warunki, w których jakaś zależność występuje i odróżnić je od warunków związanych z całkowitym brakiem tej zależności. Rycina 7 obrazuje sytuację, gdy zmienna B moderuje związek pomiędzy A i C. Strzałki ciągłe oznaczają istotne zależności w przypadku moderacji, natomiast strzałka kreskowana oznacza zależność, która nie musi być istotna. Najważniejszym dowodem istnienia moderatora jest istotność interakcji zmiennych A x B [86].

#### b) Mediator

Jest to zmienna będąca stanem lub procesem pośredniczącym między jakąś zmienną niezależną a zmienną zależną, pomiędzy którymi istnieje związek [87] [55]. Mediator zależności odpowiada na pytanie: „dlaczego ona występuje?”. Znalezienie mediatora istniejącej, znanej zależności jest więc równoznaczne z jej wyjaśnieniem.

Po uwzględnieniu wpływu zmiennej mediującej (pośredniczącej), zależność zasadnicza spada (w idealnych warunkach do zera).



Rycina 8. Mediacja. Zmienna niezależna M jest mediatorem, tzn. pośredniczy w zależności zasadniczej A i C.

Na Rycinie 8 zilustrowano mediację, która wskazuje, iż wpływ zmiennej niezależnej (A) na zmienną zależną (C) jest przekazywany poprzez trzecią zmienną (M). Mówimy więc, że zmienna A wpływa na zmienną C poprzez zmienną M. Kierunki strzałek oznaczają kierunek mechanizmu przyczynowo -

skutkowego zależności, natomiast strzałka kropkowana oznacza zależność, która zostaje osłabiona lub wygasa po dodaniu mediatora do modelu.

c) Zmienna wikłająca

W pracy opisano również zmienne zakłócające, jednak niebędące ani moderatorem, ani mediatorem. Są to zmienne, które „wikłają” proste zależności pomiędzy analizowanymi zmiennymi i mogą wpływać na wynik badania w sposób niekontrolowany [55].

### 3.4.9 Zastosowane oprogramowanie

Wybór oprogramowania wspomagającego dla poszukiwania rozwiązania omawianego problemu klinicznego zdeterminowany został przez wybrane metody analizy danych.

Do obliczeń użyto pakietów statystycznych:

- Statistica 10 (plus Zestaw Scoringowy)
- PQStat 1.4.8
- StatXact 9

Część analiz związanych z eksploracją danych przeprowadzono z wykorzystaniem programów:

- ROSE 2 wersja 2.2
- WEKA
- RSES2
- Implementacja algorytmu BRACID (metody uczenia reguł dla zbiorów niezbalansowanych - Bottom-up Induction of Rules And Cases from Imbalanced Data) autorstwa zespołu profesora Jerzego Stefanowskiego z Politechniki Poznańskiej, a w szczególności dzięki publikacjom dr Krystyny Napierały [88].

### **3.5 Zastosowana metodologia**

Ocena czynników ryzyka występowania epizodów omdleń, w grupie ze zdiagnozowaną arytmia, wymagała przeanalizowania dostępnych cech, które podejrzewane były o ewentualny związek z analizowanym zdarzeniem. Poszukiwano również efektów interakcyjnych i mediacyjnych oraz oceniano zmienne wnikające. Szczególną uwagę zwrócono na częstotliwość HR częstoskurczu, dlatego poszukiwano cech mogących mieć związek z podwyższeniem wartości tej cechy.

#### **3.5.1 Spełnianie założeń**

Dla analizowanych danych zbadano zgodność z rozkładem normalnym (test Lillieforsa) zmiennych ciągłych wykorzystywanych w dalszych analizach (Załącznik nr 1; Tab. I). Większość rozkładów zmiennych wykazywała odchylenia od rozkładu normalnego, jednak tylko nieliczne zmienne miały wysokie współczynniki skośności i kurtozy. W dalszej części pracy cytowane są wyniki zarówno nieparametrycznych, jak i parametrycznych analiz (dla zmiennych wykazujących niewielkie odchylenia od rozkładu normalnego).

W modelach regresji wielorakiej i logistycznej nie uwzględniano jednocześnie cech nadmiernie ze sobą skorelowanych. Dla modeli regresji wielorakiej zbadano normalność rozkładu reszt modelu (test Lillieforsa), aby nie zaburzyć oceny istotności współczynników poszczególnych zmiennych modelu. Dla wybranych modeli wyznaczono także wartości miar DFFITS w celu identyfikacji przypadków wpływowych. [70] Za wpływowe uznano te obserwacje, których wartość DFFITS przekraczała wartość 1.

#### **3.5.2 Wstępna charakterystyka badanej populacji**

- 1) Porównanie obu rodzajów arytmii (AVNRT i AVRT).
  - a) Porównania zmiennych ilościowych dokonano za pomocą testu parametrycznego t-Studenta lub Manna-Whitneya (w przypadku braku zgodności z rozkładem normalnym).
  - b) Różnice dotyczące cech opisanych na skali nominalnej zbadano stosując testy Chi-kwadrat oraz test Fishera-Freemana-Haltona (w zależności od liczności w tabeli wielodzielczej)
- 2) Dokonano wstępnego przetwarzania danych:
  - a) wyboru predyktorów,
  - b) reprezentantów,
  - c) dyskretyzacja zmiennych.

### 3.5.3 Czynniki powiązane z częstotliwością HR częstoskurczu

- 1) Statystyki jednowymiarowe:
  - a) Wpływ poszczególnych zmiennych interwałowych na wartość częstotliwości HR częstoskurczu standaryzowanego przeanalizowany został poprzez wyznaczenie współczynników korelacji Spearmana. Istotność tych współczynników zbadano testem t-Studenta na istotność współczynnika korelacji.
  - b) Do analizy danych nominalnych powiązanych z częstoskurczem użyto testu porównującego dwie grupy (test Manna-Whitneya).
- 2) Przeprowadzono również analizę wartości częstotliwości HR częstoskurczu po podaniu leku, czyli niestandaryzowanego, w kontekście współwystępowania dodatkowych chorób (test Manna-Whitneya do porównania dwóch grup) oraz wartości należnych HR częstoskurczu, wyznaczonych na podstawie wieku pacjenta (porównania z HR częstoskurczu obserwowanego dokonano za pomocą testu kolejności par Wilcoxon). W celu oceny zależności HR należnego z pozostałymi zmiennymi obliczono współczynniki korelacji Spearmana. Analizę czterech podgrup wg procentów należnej HR dokonano za pomocą testu ANOVA Kruskala-Wallis oraz testu Fishera-Freemana-Haltona.
- 3) Przeprowadzono serię analiz regresji wielorakiej w celu oceny jednoczesnego wpływu na częstotliwość HR częstoskurczu wielu zmiennych.
  - a) Dla każdego modelu dokonano jego porównania, testem F, z modelem zawierającym jedynie wiek.
  - b) Dokonano oceny zmiennych wnikających, interakcyjnych i mediacyjnych: obok płci, rodzaju arytmii oraz wieku pacjenta, przeanalizowano także wpływ wieku częstoskurczu oraz współwystępowania innych, jawnych chorób. Oceny tej dokonano poprzez dołączanie tych dodatkowych zmiennych do wyliczanych modeli regresji wielorakiej, poprzez dodawanie interakcji do tych modeli lub poprzez przeliczanie powyższych analiz w podgrupach.
  - c) Analizowane modele regresji wielorakiej przeliczono zarówno dla całej populacji, jak i dla wybranej, najstarszej podgrupy (mężczyzn z AVNRT).
  - d) Dla wszystkich modeli regresji wielorakiej oceniono zarówno istotność całego modelu (test F), istotność poszczególnych

zmiennych w modelu (test t-Studenta), jak i jakość dopasowania modeli (skorygowana wartość  $R^2$ , wartość SE). Zbadano również normalność reszt modelu (test Lillieforsa), a w przypadkach wymagających dodatkowej oceny obecności ewentualnych obserwacji wpływowych wyliczono także wartości miar DFFITS [70].

#### 3.5.4 Czynniki powiązane z występowaniem epizodów omdleń

##### 3) Statystyki jednowymiarowe:

Wystąpienie w przeszłości epizodu omdlenia zostało zapisane w bazie jako zmienna dychotomiczna (T/N).

- a) Celem przeprowadzenia oceny poszczególnych zmiennych interwałowych zastosowano testy porównujące dla dwóch grup: parametryczne i nieparametryczne (w zależności od zgodności rozkładu cechy z rozkładem normalnym).
  - b) Ewentualny wpływ cech opisanych na skali nominalnej na występowanie omdleń, zbadano stosując testy Chi-kwadrat. W przypadku wykazania istotnej statystycznie zależności wyliczono iloraz szans (OR) opisujący, jak bardzo zmiana stanu analizowanej cechy wpływa na szansę zaistnienia zjawiska niepożądanego, czyli omdlenia.
  - c) Przeanalizowano otrzymane wyniki w celu odnalezienia zmiennych wnikających, interakcyjnych i mediacyjnych.
- 4) W celu zbadania bardziej złożonych efektów wpływających na występowanie epizodów omdleń, wykorzystano wielowymiarowe techniki statystyczne i eksploracji danych:
- a) Regresję logistyczną danych pierwotnych:
    - i. Zbudowano serię modeli analizujących cechy będące potencjalnymi klinicznymi czynnikami ryzyka epizodów omdleń w analizowanej grupie oraz cechy sugerowane przez wyniki statystyk prostych.
    - ii. Uwzględniono podgrupy związane ze zmiennymi wnikającymi.
    - iii. Porównywano modele rozszerzone z ich wersją zredukowaną za pomocą testu ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele.
    - iv. Przeliczono modele stworzone po wstępnej analizie danych (dyskretyzacja, wybór reprezentantów i predyktorów)

- v. Oceniono jakość modeli
- b) Regresję logistyczną danych po zastosowaniu analizy składowych głównych:
- i. Na podstawie struktury korelacyjnej wybrano zmienne do analizy czynnikowej.
  - ii. „Kompresja” zmiennych dokonana została w trzech grupach zmiennych (Wyznaczono czynniki reprezentujące zmienne, będące kombinacją liniową zmiennych w poszczególnych grupach).
  - iii. Określono liczbę składowych głównych używając kryterium minimalnego zasobu zmienności wspólnej, kryterium wykresu osypiska oraz kryterium minimalnego zasobu zmienności wspólnej.
  - iv. W celu uzyskania czytelnego obrazu układu wartości czynnikowych, uzyskane wyniki poddano rotacji varimax
  - v. Zbudowano i oceniono model logistyczny zawierający nowe zmienne.
- c) Modele scoringowe
- i. Wybrano predyktory, czyli usunięto nieistotne dane
  - ii. Utworzono wiązki zmiennych, a następnie wybrano dla nich reprezentantów, czyli usunięto nadmiernie skorelowane zmienne
  - iii. Zastosowano regresję logistyczną jako metodę analityczną przy budowaniu modeli scoringowych
  - iv. Poszukiwano reguł dla epizodów omdleń, czyli zastosowano klasyfikację przy użyciu drzew losowych
  - v. Oceniono powstałe modele scoringowe
- d) Drzewa klasyfikacyjne
- i. Drzewa interakcyjne typu C&RT/CHAID
  - ii. Oceniono jakość drzew
- e) Zbiory przybliżone
- i. Zastosowano teorię zbiorów przybliżonych do podstawowego zbioru danych i do zbioru po dyskretyzacji.
  - ii. Ze względu na występujące niezrównoważenie danych dokonano wstępnego przekształcenia zbioru danych (nadlosowania)

- 4) Podczas większości analiz poszukiwano zmiennych wnikających, interakcyjnych i mediacyjnych. Szczególną uwagę zwrócono na następujące dodatkowe zmienne: płeć, rodzaj arytmii, wiek pacjenta, a także wpływ czasokresu, który upłynął od pierwszego udokumentowanego częstoskurczu (w latach) oraz ewentualny wpływ współwystępowania innych, jawnych chorób. Oceny tej dokonano poprzez dołączanie dodatkowych zmiennych do wyliczanych modeli regresji wielorakiej, poprzez dodawanie interakcji do tych modeli lub poprzez przeliczanie powyższych analiz w podgrupach.
- 5) Dla wybranych przypadków przeanalizowano testy dokładne (exact) oraz metody symulacji Monte Carlo.
- 6) Wszystkie modele regresji logistycznej miały ocenioną jakość (test Hosmera-Lemeshowa oraz współczynniki determinacji), istotność statystyczną całego modelu (test ilorazu wiarygodności - LR) oraz konkretnych zmiennych w modelu (test Walda).

### 3.5.5 Porównanie modeli

Otrzymane wyniki analiz, dokonanych przy użyciu różnych technik, porównano za pomocą wybranych miar oceny jakości klasyfikacji (Rozdział 3.4.7).

Oszacowanie parametrów modelu i wygenerowanie prognozy powinno odbywać się na niezależnej próbie testowej. Jednak, w przypadku posiadania małej liczby przypadków niezbalansowanych danych, taki podział zbioru danych na podzbiory (uczący i testowy) jest niekorzystny. Aby rozwiązać ten problem można zastosować m.in. wielokrotne próbkowanie elementów z dostępnej próby [23] (np. metoda bootstrap, polegająca na n-krotnym losowaniu ze zwracaniem przypadków z dostępnego zbioru danych i przypisywaniu wylosowanych elementów do zbioru uczącego) lub usuwanie kolejno po jednym z elementów próby (np. metoda Jackknife). Niestety, nie wszystkie techniki analityczne oferują próbkowanie.

W przypadku niniejszej pracy przedstawiono oceny modeli po V-krotnej walidacji. Najczęściej zbiór analizowanych obiektów dzielony jest na 10 w przybliżeniu równych podzbiorów i przeprowadzana jest 10-krotną walidacją krzyżową. W każdym z 10 kroków przeprowadzany jest eksperyment polegający na zastosowaniu modelu lub generowaniu reguł decyzyjnych na podstawie dziewięciu podzbiorów tworzących zbiór uczący i sprawdzaniu ich poprawność w odniesieniu do zbioru testującego. Wypadkowa trafność klasyfikowania obliczana jest na podstawie liczby poprawnie sklasyfikowanych obiektów w poszczególnych krokach.



## 4 Wyniki

### 4.1 Wstępna charakterystyka badanej populacji

Cały zbiór analizowanych danych zawierał informacje o 517 pacjentach. Częstotliwość HR częstoskurczu standaryzowanego wyznaczono dla 425 pacjentów, z których jedynie 407 miało odnotowaną informację w wywiadzie o zaistniałym, bądź nie, w przeszłości epizodzie omdlenia. W analizowanej grupie chorych u 19,92% kobiet i 18,93% mężczyzn wystąpiły w przeszłości epizody omdleń (Tabela 3). Tabela 2 zawiera charakterystykę osób o znanym HR częstoskurczu, z podziałem na grupy płci i występowania lub nie w przeszłości epizodu omdleń. Zarówno dla całej bazy (Tabela 3;  $p=0,90770$ ), jak i dla poszczególnych arytmii (Tabela 2;  $p=0,86363$  oraz  $p=0,56496$ ), nie wykazano istotnej statystycznie zależności pomiędzy płcią a występowaniem epizodów omdleń.

Tabela 2. Wstępna charakterystyka badanej populacji osób o znanym HR częstoskurczu - zależność omdleń od płci w podgrupach arytmii.

	AVNRT (N=269)			WPW (N=156)		
	Omdlenia	Bez omdleń	Razem z brakami danych	Omdlenia	Bez omdleń	Razem z brakami danych
Kobiety (N=256)	34	146	186	17	48	70
% kobiet	70,83%	68,87%	69,14%	48,57%	42,86%	44,87%
% omdleń	18,28%	78,49%	100,00%	24,29%	68,57%	100,00%
Mężczyźni (N=169)	14	66	83	18	64	86
% mężczyzn	29,17%	31,13%	30,86%	51,43%	57,14%	55,13%
% omdleń	16,87%	79,52%	100,00%	20,93%	74,42%	100,00%
Razem:	48	212	269	35	112	156
% omdleń	17,84%	78,81%		22,44%	71,79%	17,84%
p testu Chi <sup>2</sup> *	p=0,86363			p=0,56496		

\* Zależność omdlenia od płci

Tabela 3. Wstępna charakterystyka badanej populacji osób o znanym HR częstoskurczu - zależność omdleń od płci dla obu arytmii łącznie.

	Obie arytmie razem		
	Omdlenia	Bez omdleń	Razem z brakami danych
Kobiety	51	194	256
% kobiet	61,45%	59,88%	60,24%
% omdleń	19,92%	75,78%	100,00%
Mężczyźni	32	130	169
% mężczyzn	38,55%	40,12%	39,76%
% omdleń	18,93%	76,92%	100,00%
Razem:	83	324	425
% omdleń	20,4%	79,6%	
p testu Chi <sup>2</sup> *	p=0,90770		

\* Zależność omdlenia od płci

#### 4.1.1 Porównanie populacji arytmii AVNRT i AVRT (zespół WPW)

Obie arytmie, AVNRT i AVRT, łączy zaburzenie szerzenia impulsu elektrycznego w mechanizmie krążącego pobudzenia nawrotnego (reentry). Różnią się natomiast lokalizacją arytmogennego podłoża w sercu oraz odmienną częstością występowania w zależności od płci. Częstoskurcz typu AVNRT występuje częściej u kobiet (ponad 69%), czego nie obserwuje się w grupie AVRT (tam wskazuje się nieznaczną przewagę wśród mężczyzn). Wybrane, najważniejsze różnice pomiędzy obydwojema rodzajami arytmii, przedstawiono w Tabeli 4. (Więcej szczegółów, dotyczących porównań cech ilościowych dla obu arytmii, znajduje się w Załączniku nr 1; Tab.XII.)

Tabela 4. Wybrane różnice pomiędzy AVNRT i AVRT

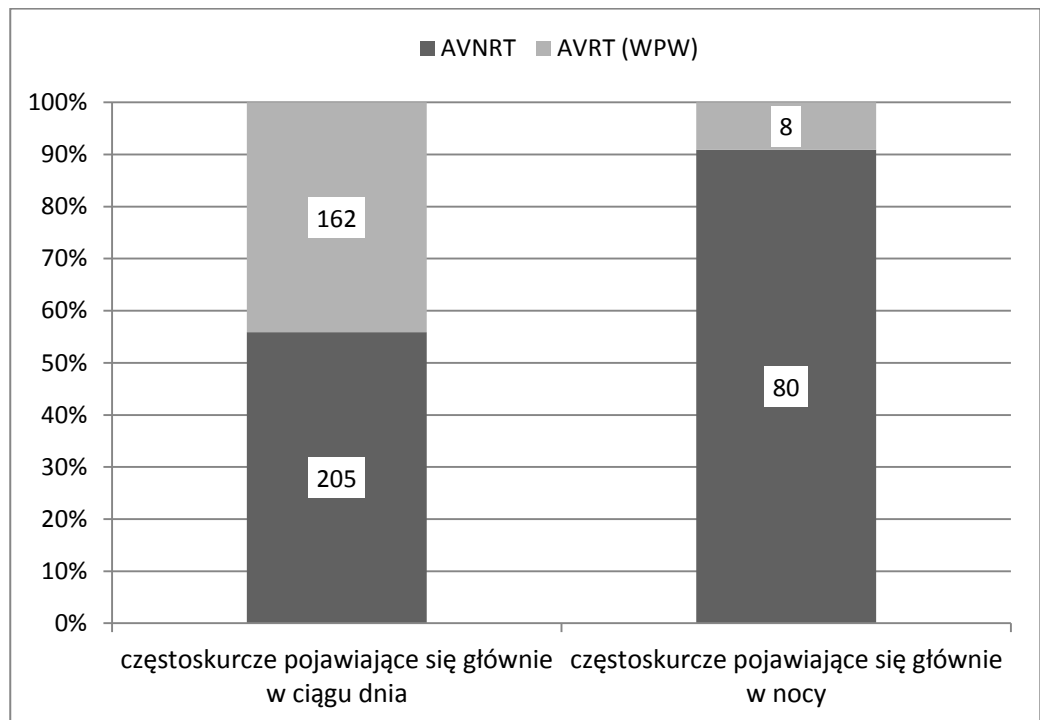
	Wartości średnie			Porównanie dwóch arytmii (p testu Manna-Whitneya)		
	AVNRT N=269	AVRT N=159	Ogółem N=425	Dla kobiet	Dla mężczyzn	Razem
<b>wiek</b>	44,33	38,46	42,18	0,002342*	0,000008*	0,000001*
<b>HR częstoskurczu standaryzowanego</b>	188,09	187,53	187,88	0,223049	0,149814	0,830719
<b>LK (mm)</b>	45,22	46,23	45,59	0,759671	0,048264*	0,025037*
<b>PMK (mm)</b>	9,43	9,93	9,61	0,999330	0,130347	0,002133*
<b>LP (mm)</b>	31,61	32,92	32,08	0,914028	0,952891	0,010843*
<b>PK(mm)</b>	24,17	24,30	24,23	0,021283*	0,543270	0,944153
<b>Ao (mm)</b>	28,37	29,30	28,71	0,974304	0,349185	0,036570*
<b>BMI (kg/m<sup>2</sup>)</b>	25,19	25,32	25,24	0,700613	0,066381	0,678462
<b>TSH (mU/l)</b>	1,68	1,47	1,63	0,458374	0,226134	0,724051

\*wyniki istotne statystycznie (p<0,05)

Najważniejsze fakty wynikające z porównania obydwu grup arytmii:

- Opisywane rodzaje arytmii nie różniły się częstością występowania epizodów omdleń (test Chi<sup>2</sup>: p=0,21851); także w podgrupach kobiet i mężczyzn (p=0,08312 oraz p=0,94928).
- W przypadku AVNRT było istotnie więcej kobiet (test Chi<sup>2</sup> p<0,000001).
- Grupa AVNRT była istotnie starsza od AVRT (test Manna-Whitneya p=0,000001).
- Arytmie nie różniły się wysokością częstotliwości HR częstoskurczu (test Manna-Whitneya p=0,830719).
- Brak różnic pomiędzy arytmią a liczbą zarejestrowanych częstoskurczów (test Manna-Whitneya p=0,591578).

- W obu arytmach nie było istotnej zależności pomiędzy występowaniem omdleń a indukowalnością częstoskurczu (p testu Chi<sup>2</sup> wynosiło p=0,51358 dla AVNRT oraz p=0,98027 dla AVRT).
- Występują istotnie statystyczne różnice pomiędzy arytmiami w zakresie wpływu układu współczulnego i przywspółczulnego, przy spontanicznym pojawianiu się częstoskurczu (w AVRT prawie wszystkie odnotowane częstoskurcze pojawiały się w ciągu dnia - Rycina 9; test Chi<sup>2</sup> p<0,000001).



Rycina 9. Pora dnia pojawiania się częstoskurczów typu AVNRT i AVRT.

- Więcej osób (27,68%) z grupy AVNRT miało odnotowane istotne klinicznie objawy jawnej choroby typu II (pacjenci z cukrzycą, nadciśnieniem tętniczym, po zawale serca lub z dławicą piersiową); w grupie WPW było 17,11% takich osób (test Chi<sup>2</sup>: p=0,00456).
- W grupie AVRT (WPW) pacjenci z współwystępującymi objawami jakichkolwiek jawnych chorób lub chorób typu II, nie różnili się wysokością częstotliwości HR częstoskurczów w porównaniu z grupą bez takich dodatkowych objawów (test Manna-Whitneya p<0,000001 w obu przypadkach), choć pacjenci z dodatkowymi chorobami byli istotnie starsi (test Manna-Whitneya p=0,000997 i <0,000001 odpowiednio). W grupie AVNRT natomiast, częstoskurcz był istotnie wolniejszy w starszych, chorujących grupach. (Załącznik nr 1; Tab. III)

- W przypadku AVNRT, czyli pacjentów z nawrotnym częstoskurczem węzłowym – zawsze dążono do sprowokowania częstoskurczu podczas ablacji. Inaczej było w przypadku zespołu WPW – tam indukowalność częstoskurczu wynosiła 64%. (test  $\text{Chi}^2$   $p < 0,000001$ )

#### 4.1.2 Przetwarzanie wstępne

Analizowany zbiór danych został poddany wstępnej analizie, czyli metodom mającym na celu zawęzić zbiór zmiennych jedynie do tych, które mogą mieć najsilniejszy związek z występowaniem omdlenia. Rozpoczęto od wyboru predyktorów, następnie dokonano wyboru wiązek reprezentantów, a na koniec dokonano dyskretyzacji zmiennych. Na Rycinie 10 przedstawiono schemat postępowania związanego ze wstępną analizą danych. Wynikowy zbiór danych zawierał 26 najważniejszych zmiennych po dyskretyzacji.



Rycina 10. Wstępna analiza zmiennych.

W kolejnych podrozdziałach przedstawiono szczegółowe rezultaty poszczególnych etapów analizy, prowadzące do powstania nowego, zawężonego zbioru danych. Wynikowy zbiór posiadał 466 obiektów (pacjentów), dla których zdefiniowana była zmienna zależna, czyli informacja o występowaniu omdleń.

##### 4.1.2.1 Wybór predyktorów

W niniejszej pracy tzw. ranking predyktorów wykonany został na podstawie obliczonej mocy informacyjnej dla każdej zmiennej (IV, ang. Information Value). Polega to na zsumowaniu wartości tzw. wag dowodu (WoE) dla wszystkich przedziałów analizowanej zmiennej (Rozdział 3.4.1.1). Wyniki rankingu predyktorów przedstawiono w Tabeli 5, która oprócz miary IV zawiera także dwie inne miary: V Cramera oraz współczynnik Giniego. Po wyeliminowaniu cech diagnostycznych, które są nieistotne z punktu widzenia wpływu na analizowaną zmienną zależną (tu: omdlenie), w zbiorze danych pozostawiono 38 zmiennych. Zastosowane kryterium: pozostawiono zmienne o mocy

informacyjnej (IV) co najmniej 0,02. Zauważono, iż przy zastosowanym kryterium zmienną, która najmocniej wpływa na wykrycie ryzyka omdlenia jest FT4 (Tabela 5). Niestety jest to zmienna, która ma w analizowanej bazie liczne braki danych (jedynie 105 osób miało odnotowaną wartość FT4). Z tego względu zmienna ta nie mogła być włączona do większości planowanych analiz.

Tabela 5. Zmienne pozostawione w zbiorze danych po przeprowadzeniu rankingu predyktorów.

	Zmienna	IV	V Cramera	Gini
1	FT4 (pmol/l)	0,5	0,28	0,12
2	EP: po-CI-Wen- retro	0,4	0,22	0,03
3	EP: po-HR-Wenck-retro	0,35	0,23	0,04
4	FT3 (pmol/l)	0,3	0,33	0,16
5	EP: CI-Wen-ante	0,29	0,23	0,01
6	EP: HR-Wenck-ante	0,29	0,23	0,01
7	EP: po-CI-Wen-ante	0,27	0,2	0,01
8	CL* (ms)	0,25	0,17	0,07
9	K+ (mmol/l)	0,23	0,19	0,15
10	PK (mm)	0,21	0,17	0,14
11	EP: pre-ERP-Avnr	0,19	0,16	0,02
12	EP: CI-Wen- retro	0,18	0,18	0,03
13	TSH (mU/l)	0,18	0,17	0,11
14	LP (mm)	0,16	0,16	0,13
15	Ao (mm)	0,14	0,14	0,06
16	waga (kg)	0,13	0,15	0,05
17	EP: pre-ERP-Avna	0,13	0,13	0,06
18	EP: po-ERP-Avnr	0,12	0,15	0,01
19	PMK (mm)	0,11	0,14	0,1
20	EP: po-ERP-Avna	0,11	0,13	0,09
21	LK (mm)	0,11	0,13	0,09
22	różnica należna HR i z standaryzowanego	0,11	0,13	0,03
23	EP-AVNRT-HR max	0,11	0,12	0,01
24	BMI (kg/m <sup>2</sup> )	0,1	0,13	0,1
25	grupa CL*	0,1	0,12	0,14
26	wzrost (cm)	0,08	0,12	0,02
27	Wiek	0,08	0,11	0,08
28	Wywiad-udokument-AVNRT (HR)	0,07	0,11	0,05
29	Na+ (mmol/l)	0,07	0,11	0,02
30	EP: HR-Wenck-retro - wartości	0,07	0,11	0
31	kiedy pierwszy częstoskurcz	0,06	0,1	0,08
32	częstoskurcz standaryzowany	0,06	0,1	0,01
33	kategorie BMI 2	0,06	0,09	0,07
34	EP: po-HR-Wen-ante- wartości	0,04	0,08	0,04
35	Choroby współistniejące typu II	0,03	0,07	0,08
36	Pora dnia	0,03	0,06	0,01
37	Choroby współistniejące typu I	0,02	0,06	0,07
38	Naprężenie zespołu QRS	0,02	0,05	0,05

\*CL – dotyczy wartości CL częstoskurczu podczas badania EP

#### 4.1.2.2 Wybór reprezentantów

Na podstawie kryterium minimalnego zasobu zmienności wspólnej (analiza czynnikowa omówiona została w rozdziale 3.5.3), wybrano 9 czynników (inaczej wiązek danych) i przypisano do nich te zmienne, które najmocniej z nimi korelują. W Załączniku nr 1 zawarto tabelę wartości własnych poszczególnych czynników – Tab. XIII. Ładunki dla poszczególnych zmiennych wyliczono z zastosowaniem analizy czynnikowej ze znormalizowaną rotacją czynników (Varimax) - Tabela 6. W zbiorze danych pozostawiono, co najwyżej, po dwie zmienne z każdego czynnika (z każdej wiązki), które miały najwyższe ładunki (powyżej 0,7), tzn. były najsilniej skorelowane z danym czynnikiem.

Tabela 6. Wybór reprezentantów z zastosowaniem analizy składowych głównych – ładunki zmiennych.

	Czynnik	Zmienna	Ładunek	IV	Wybór
1	1	waga (kg)	0,89	0,27	+
2	1	różnica należna HR i z standaryzowanego	0,85	0,03	+
3	1	EP: HR-Wenck-retro -	-0,929	0,13	-
4	2	EP: po-ERP-Avnr	-0,95	0,02	+
5	2	EP: po-HR-Wen-ante-	-0,836	0,03	+
6	2	EP: po-CI-Wen- retro	-0,934	0,13	-
7	2	EP: po-HR-Wenck-retro	-0,792	0,19	-
8	3	Pora dnia	-0,902	0,03	+
9	3	LK (mm)	0,861	0,1	+
10	3	Wywiad-udokument-AVNRT (HR)	-0,889	0,06	-
11	4	FT4 (pmol/l)	-0,709	0,01	-
12	4	FT3 (pmol/l)	0,774	0,29	-
13	4	Wiek	0,806	0,18	+
14	5	EP: CI-Wen-ante - wartości	0,975	0,12	+
15	5	EP: po-CI-Wen-ante - wartości	0,953	0,1	-
16	5	K+ (mmol/l)	0,737	0,17	-
17	6	częstoskurcz standaryzowany	0,903	0,07	+
18	6	wzrost (cm)	0,863	0,02	+
19	7	Ao (mm)	0,825	0,1	+
20	7	Na+ (mmol/l)	0,713	0,09	+
21	8	PK (mm)	0,971	0,17	+
22	8	CL* (ms)	-0,743	0,24	+
23	8	Naprzemienność zespołu QRS	0,96	0,02	-
24	9	Choroby współistniejące typu I	0,82	0,02	+
25	9	Choroby współistniejące typu II	0,708	0,03	+
26	9	kiedy pierwszy częstoskurcz	-0,798	0,06	-

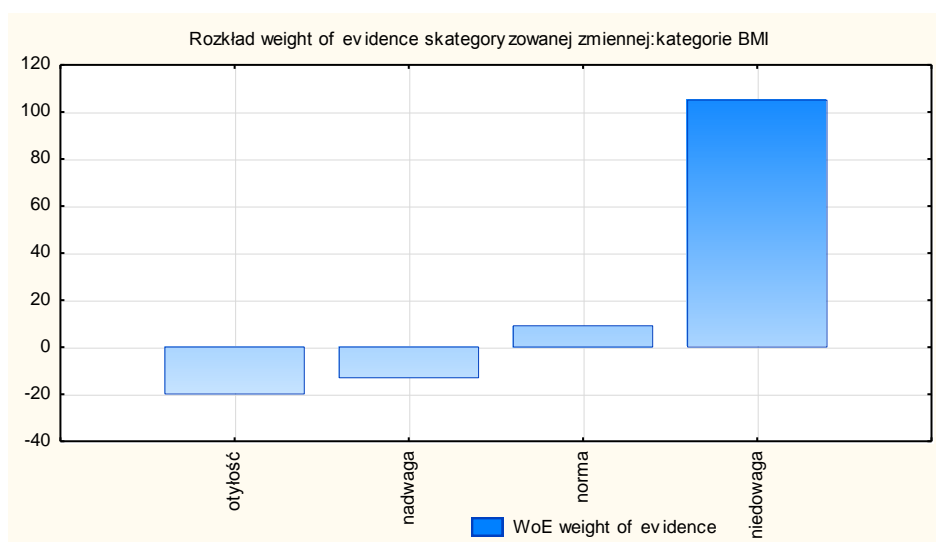
„+” - wybrano zmienną do dalszej analizy; „-” - nie wybrano zmiennej do dalszej analizy

\* CL – dotyczy wartości CL częstoskurczu podczas badania EP

Następnym krokiem było przeanalizowanie macierzy korelacji wewnątrz każdego z czynników w celu wyeliminowania z dalszej analizy zmiennych nadmiernie ze sobą skorelowanych. Macierze korelacji przedstawiono w Załączniku nr 1. Wykluczone zostały zmienne o sile korelacji powyżej 0,4. W przypadku czynnika 4 pozostawiono jedynie zmienną wiek ze względu na liczne braki danych związane z wartościami FT3 i FT4. Po wyborze 16 reprezentantów (kolumna „Wybór”, Tabela 6), łącznie pozostawiono w analizowanym zbiorze 27 zmiennych.

#### 4.1.2.3 Dyskretyzacja zmiennych

Dyskretyzacja jest procesem zamiany atrybutów liczbowych na atrybuty symboliczne typu porządkowego. Wyznaczono więc w obrębie wszystkich zmiennych grupy jednorodne ze względu na ryzyko wystąpienia omdleń. Zmienne ilościowe podzielono na kategorie, rozpoczynając od podziału co 10 procentyli, a następnie korzystając z wykresów wartości WoE połączono najbliższe sobie kategorie. Granice przedziałów były generowane automatycznie wg rozkładu zmiennej i nie były modyfikowane wg klinicznych norm. Wyjątek stanowią cechy, które już wcześniej przekodowane były na kategorie wg ustalonych wcześniej schematów (np. BMI - oceniając zmienną BMI zrezygnowano z jej postaci interwałowej, natomiast pozostawiono skategoryzowaną wg norm WHO). Na Rycinie 11 przedstawiono wykres wartości WoE dla kategorii BMI, dla zmiennej stanu *omdlenie*:

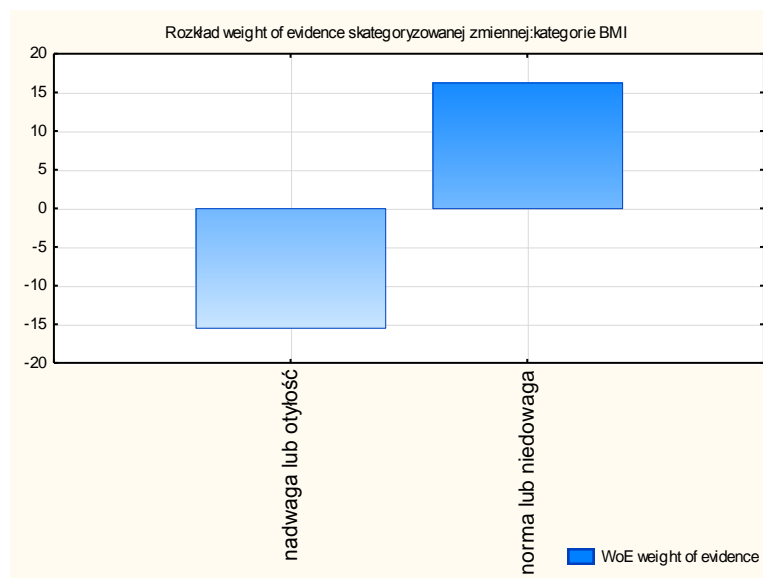


Rycina 11. Wartości WoE dla zmiennej kategorii BMI przed dyskretyzacją.

Wyższe wartości WoE informują o większym prawdopodobieństwie braku wystąpienia omdlenia. Na Rycinie 11 można zauważyć, iż kategorie „otyłość” i



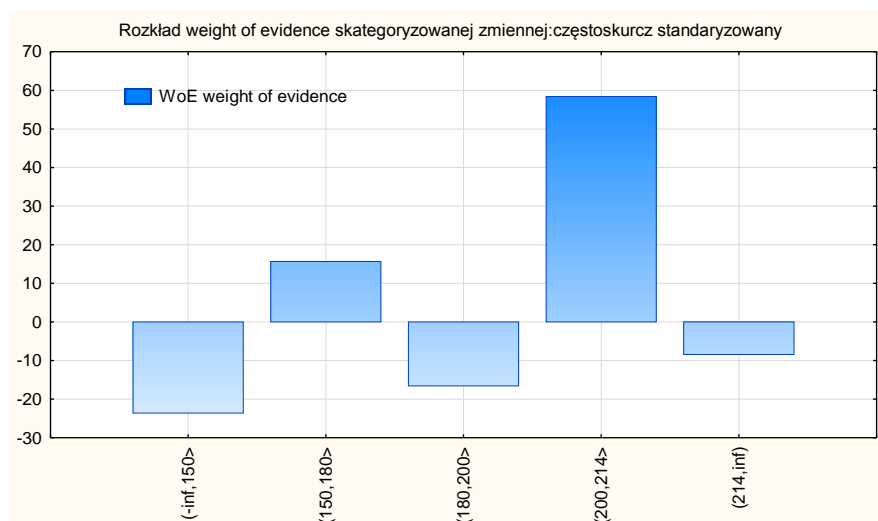
„nadwaga” mają podobne zakresy WoE, wskazujące na podobne (aczkolwiek niewielkie) ryzyko wystąpienia omdlenia. Takie kategorie mogą zostać połączone w jedną wspólną. Dwie pozostałe kategorie, czyli BMI w normie oraz „niedowaga”, również można połączyć, gdyż w tym przypadku jest szansa na to, iż omdlenie nie wystąpi. Warto zauważyć, iż dla osób z niedowagą mamy największe prawdopodobieństwo, że nie wystąpią tam omdlenia. Po dyskretyzacji zmienna *kategorie BMI* wygląda zatem w sposób zaprezentowany na Rycinie 12.



Rycina 12. Wartości WoE dla zmiennej kategorii BMI po dyskretyzacji.

Po przeanalizowaniu wykresów wartości WoE dla pozostałych, wybranych cech (Załącznik nr 1; Tab. XIV) stwierdzono, że w analizowanej bazie pacjentów większość cech nie wykazywała regularności podobnej do tej zaprezentowanej na Rycinie 12 dla zmiennej BMI. Występujące podgrupy jednorodnego ryzyka miały małą liczebność i ich bezwzględnie niewielkie wartości WoE na zmianę były dodatnie i ujemne.

Na Rycinie 13 zaprezentowano wykres dotyczący wartości WoE dla częstotliwości HR częstoskurczu standaryzowanego – jednego z potencjalnych czynników ryzyka omdlenia. Wykres ten obrazuje naprzemiennie dodatnie i ujemne wartości wagi dowodu w kolejnych podgrupach częstotliwości HR częstoskurczu.



Rycina 13. Wartości WoE dla zmiennej HR częstoskurczu standaryzowanego.

Taki brak regularności na wykresie (Rycina 13) świadczy o słabym związku analizowanej cechy ze zmienną zależną (omdleniem). W Tabeli 7 przedstawiono szczegółowe charakterystyki przedziałów po dyskretyzacji utworzonych dla zmiennej częstotliwość HR częstoskurczu standaryzowanego.

Tabela 7. Kategoryzowana zmienna: HR częstoskurczu standaryzowanego – charakterystyka przedziałów.

częstoskurcz standaryzowany - przedziały po dyskretyzacji	N	T	liczność	% N	% T	Procent ogółem	IV	WoE
(-inf,150>	37	12	49	11,42%	14,46%	12,04%	0,01	-23,59
(150,180>	105	23	128	32,41%	27,71%	31,45%	0,01	15,66
(180,200>	86	26	112	26,54%	31,33%	27,52%	0,01	-16,57
(200,214>	35	5	40	10,80%	6,02%	9,83%	0,03	58,40
(214,inf)	61	17	78	18,83%	20,48%	19,16%	0,00	-8,42
<b>Ogół</b>	<b>324</b>	<b>83</b>	<b>407</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>	<b>0,05</b>	

Wykresy wartości WoE dla pozostałych zmiennych po dyskretyzacji dołączono w Załączniku nr 1 - Tab. XIV.

## 4.2 Czynniki powiązane z HR częstoskurczu

Analizy opisane w kolejnych podrozdziałach biorą pod uwagę częstoskurcz standaryzowany, by wyeliminować wpływ leku (Isoprenaliny) na wartość HR częstotliwości częstoskurczu. Przeanalizowano dodatkowo wpływ ewentualnych zmiennych wnikających: m.in. wieku pacjentów oraz długości czasokresu od pierwszego udokumentowanego częstoskurczu w latach.

### 4.2.1 Zależności proste

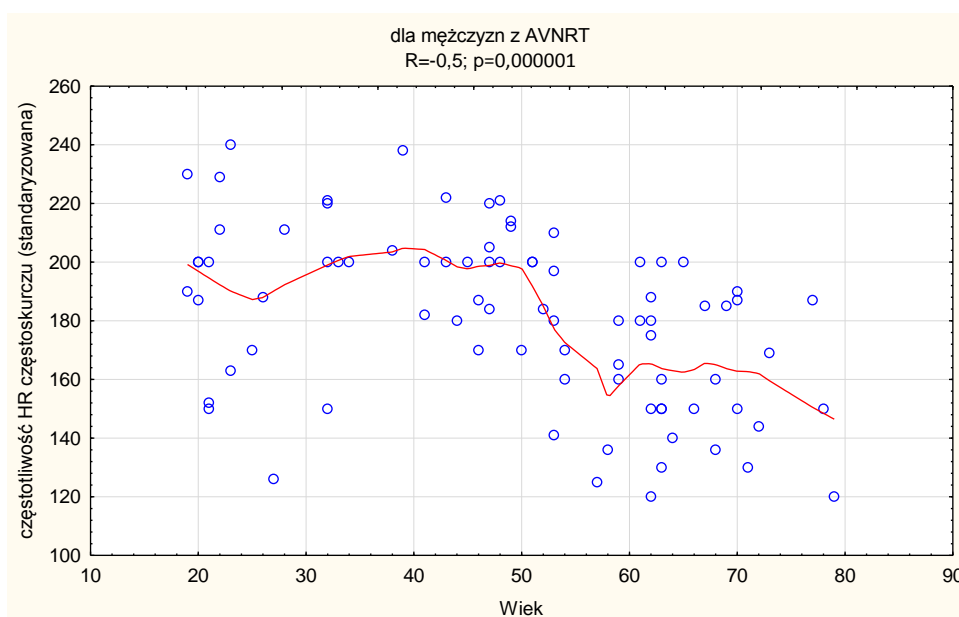
Najważniejszą, powszechnie znaną, zależnością częstotliwości HR częstoskurczu jest spadek jego wartości wraz z wiekiem. [89]

Tabela 8. Współczynniki korelacji Spearmana wieku i standaryzowanej wartości HR częstoskurczu w podgrupach płci i rodzaju arytmii.

arytmia	Płeć	liczność (N)	R wsp.Spearmana	p - testu na istotność wsp.korelacji
całość		425	-0,318	0,000000*
AVNRT	K	186	-0,301	0,000030*
AVNRT	M	83	-0,503	0,000001*
AVRT	K	70	-0,263	0,027566*
AVRT	M	86	-0,253	0,018638*

\*wyniki istotne statystycznie ( $p < 0,05$ )

Z Tabeli 8 wynika, iż zależność wieku i standaryzowanej wartości HR częstoskurczu najsilniej widoczna jest w grupie mężczyzn z AVNRT – istotny statystycznie ( $p = 0,000001$ ) współczynnik korelacji Spearmana wynosi  $-0,50$  (Rycina 14). Dla całej analizowanej grupy istotny współczynnik korelacji Spearmana wynosi  $-0,32$ .

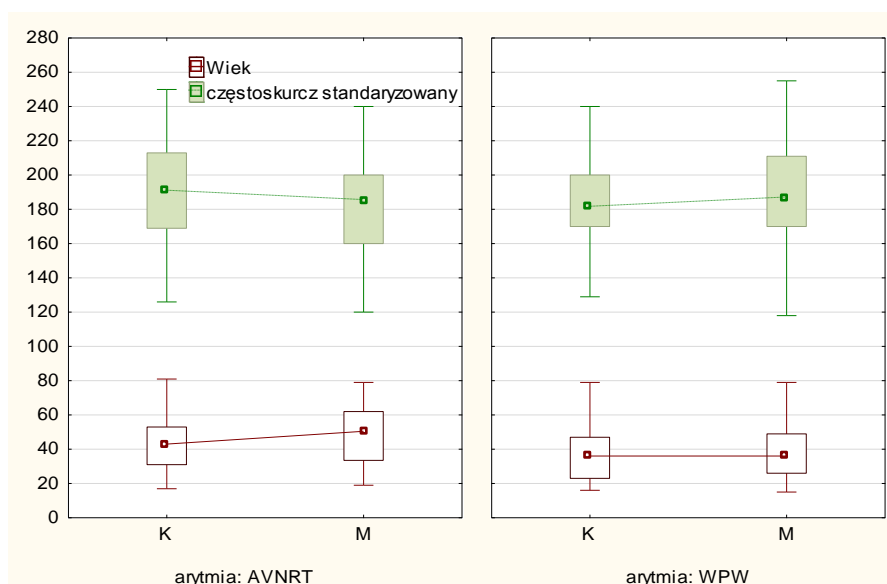


Rycina 14. Wykres rozrzutu częstotliwości HR częstoskurczu wobec wieku mężczyzn z grupy AVNRT.

Na Rycinie 14 zastosowano metodę wygładzania LOWESS (ważona regresja lokalnie wielomianowa) w celu zaprezentowania ogólnej zależności między dwiema analizowanymi zmiennymi.

Istotna zależność wieku i częstotliwości HR częstoskurczu powoduje, iż w następnych etapach analizy, wiek był jednym z czynników branych pod uwagę przy tworzeniu modeli. Większość zmiennych, które są skorelowane z HR częstoskurczu, zależy także od wieku. (Więcej wartości współczynników korelacji dla HR częstoskurczu i wieku przedstawiono w Załączniku nr 1; w Tab. IX i Tab. X.)

W całej populacji kobiety i mężczyźni nie różnili się istotnie ani wiekiem ( $p=0,397061$ ), ani HR częstoskurczu ( $p=0,239445$ ). Analizując osobno dwa rodzaje arytmii wykazano, iż grupa AVRT (WPW) nie różni się istotnie wiekiem pomiędzy płciami ( $p=0,595734$ ), natomiast w AVNRT mężczyźni byli istotnie starsi ( $p=0,003798$ ) i mieli również znacząco wolniejsze częstoskurcze ( $p=0,02895$ ). Sytuację tę obrazuje Rycina 15.



Rycina 15. Mediana wieku i HR częstoskurczu w podgrupach płci, skategoryzowana względem rodzaju arytmii.

Analizując cechy nominalne, stwierdzono istotnie niższe wartości HR częstoskurczu osób ze zgłoszonymi w wywiadzie chorobami dodatkowymi, wobec osób bez żadnych jawnych objawów choroby współistniejącej ( $p=0,010022$ ). Jednocześnie wykazano, iż pacjenci ze zgłoszonymi w wywiadzie chorobami dodatkowymi byli istotnie starsi ( $p<0,000001$ ). (Tabela 9 zawiera wyniki porównania dwóch grup, natomiast szczegóły charakterystyk opisowych zawarte zostały w Załączniku nr 1; w Tab. II i Tab. III. Również u pacjentów bez naprzemienności zespołu QRS odnotowano wolniejsze częstoskurcze ( $p<0,000001$ ), ale jednocześnie były to osoby starsze od tych z

naprzemiennością zespołu QRS ( $p < 0,000001$ ).

Tabela 9. Wyniki porównania (p testu Manna-Whitneya) (1) wieku oraz (2) częstotliwości HR częstoskurczu pomiędzy dwoma podgrupami: płci, współwystępowania dodatkowych, jawnych objawów innych chorób, naprzemienności zespołu QRS oraz arytmii.

		zmienna tworząca dwie podgrupy				
		płeć	choroba typu I (tak/nie)	choroba typu II (tak/nie)	naprzemienność zespołu QRS (tak/nie)	rodzaj arytmii
<b>(1)</b>	<b>Wiek</b>	0,397061	<0,000001*	<0,000001*	<0,000001*	0,000001*
	dla kobiet	-	<0,000001*	<0,000001*	0,000109*	0,002342*
	dla mężczyzn	-	<0,000001*	<0,000001*	0,000566*	0,000008*
	dla AVNRT	0,003798*	<0,000001*	<0,000001*	0,000199*	-
	dla WPW	0,595734	0,000997*	<0,000001*	0,001407*	-
<b>(2)</b>	<b>HR częstoskurczu standaryzowanego</b>	0,239445	0,010022*	0,000963*	<0,000001*	0,830719
	dla kobiet	-	0,034137*	0,014659*	<0,000001*	0,223049
	dla mężczyzn	-	0,198947	0,036751*	0,000002*	0,149814
	dla AVNRT	0,028951*	0,045524*	0,000051*	0,000051*	-
	dla WPW	0,463077	0,096892	0,777717	<0,000001*	-

\*wynik istotny statystycznie ( $< 0,05$ );

Zauważono także, iż pomimo tego, że w grupie AVNTR byli pacjenci istotnie starsi, w porównaniu z AVRT (WPW) ( $p = 0,000001$ ), nie odnotowano u nich istotnie wolniejszych częstoskurczów standaryzowanych, tzn. częstoskurczów niepodwyższonych sztucznie poprzez podanie leku ( $p = 0,830719$ ). (Najważniejsze wyniki porównań wieku i HR częstoskurczu przedstawiono w Załączniku nr 1; Tab. II i Tab.III, natomiast dodatkowe porównania w kluczowych podgrupach przedstawiono w Tabeli 9)

Tabela 10. Porównywane populacje, u których nie wystąpiły istotnie niższe wartości HR częstoskurczu w grupie starszej.

Grupa starsza	Grupa młodsza
AVNRT	WPW
WPW z chorobą typu I	WPW bez choroby typu I
WPW z chorobą typu II	WPW bez choroby typu II
Mężczyźni z chorobą typu I	Mężczyźni bez choroby typu I
dłuższe czasokresy częstoskurczów	„krótsze” czasokresy częstoskurczów
kobiety - częstoskurcze „nocne”	kobiety – częstoskurcze „dienne”

Podobna sytuacja miała miejsce w podgrupie AVRT (WPW) dla chorób typu I i II oraz w podgrupie kobiet dla częstoskurczów pojawiających się w ciągu dnia, a także w podgrupie mężczyzn w przypadku chorób typu I oraz w podgrupach z dłuższym i krótszym czasookresem od pierwszego pojawienia się częstoskurczu.

We wszystkich tych przypadkach starsi pacjenci nie mieli niższych częstotliwości HR częstoskurczu. Zestawienie podgrup, u których wpływ wieku na HR częstoskurczu został zniwelowany, przedstawiono w Tabeli 10.

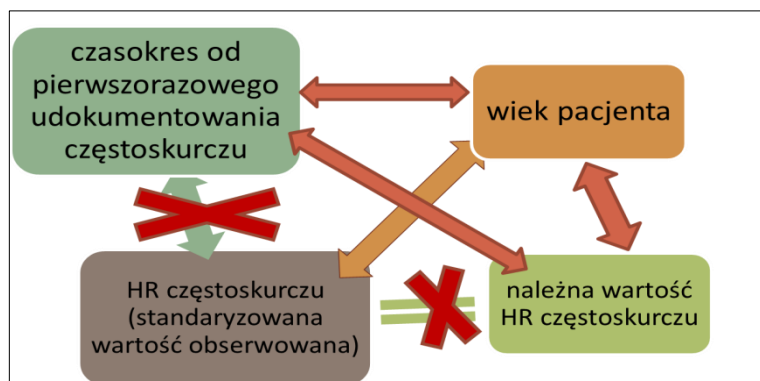
Czasokres od pierwszego udokumentowanego częstoskurczu był silnie skorelowany z wiekiem pacjenta. Starsi mają „dłuższe” czasokresy częstoskurczów (minęło więcej czasu od pojawienia się tych udokumentowanych epizodów) (Załącznik nr 1). Mimo to czasokres od pierwszego udokumentowanego częstoskurczu nie był istotnie skorelowany ze standaryzowaną wartością częstotliwości HR częstoskurczu - współczynnik korelacji Spearmana wynosił -0,03 i nie był istotny (wyjątek stanowiła grupa kobiet z AVNRT: istotny współczynnik korelacji -0,14).

Jest to związane z kolejną analizowaną zmienną: częstoskurczem należnym. Ponieważ jest on funkcją wieku (rozdział 3.2.4.4) jest z nim silnie skorelowany (wsp. korelacji Spearmana wynosił -0,49 i był istotny). Dla całej bazy, dla AVNTR i dla grupy kobiet częstoskurcz standaryzowany i wartość HR należąca różniły się istotnie – tzn. obserwowane częstoskurcze były istotnie szybsze w rzeczywistości niż ich wyliczone wg wieku należne odpowiedniki HR (Załącznik nr 1; Tab. XI). Natomiast w grupie WPW i w grupie mężczyzn, odnotowano częstotliwości HR częstoskurczów nieodbiegające istotnie od tych wyliczonych względem wieku.

Częstotliwość HR częstoskurczów wyzwalanych przy współudziale układu współczulnego była istotnie wyższa w całej analizowanej grupie ( $p=0,04889$ ), w podgrupie AVNRT ( $p=0,027830$ ) oraz dla mężczyzn ( $p=0,031705$ ). Jednocześnie pacjenci, u których częstoskurcze pojawiały się w ciągu dnia, byli istotnie młodsi (szczegóły w Tab. II i Tab. III; Załącznik nr 1), co może tłumaczyć ich szybsze częstoskurcze. Dla kobiet nie wykazano istotnych różnic w wartościach HR częstoskurczów wyzwalanych przy widocznym udziale układu współczulnego w porównaniu do częstoskurczów tzw. „nocnych” ( $p=0,324464$ ), mimo, iż „nocna” grupa była istotnie starsza ( $p=0,000057$ ).

Na Rycinie 16 przedstawiono opisane w tym rozdziale zmienne uwikłane, powiązane z częstotliwością HR częstoskurczu i wiekiem oraz zależności występujące pomiędzy nimi.

U osób, u których występowały stany poprzedzające omdlenia, odnotowano istotnie wyższe wartości częstotliwości HR częstoskurczu (test Manna-Whitneya  $p=0,023$ ). Jednocześnie wiadomo, że występowanie stanów przedomdleniowych nie było powiązane z wiekiem (test Manna-Whitneya  $p=0,134$ ).



Rycina 16. Standaryzowana wartość częstotliwości HR częstoskurczu - zmienne uwikłane.

Analizując zależności proste nie wykryto natomiast bezpośredniego związku częstości występowania omdleń z częstotliwością HR częstoskurczu ( $p=0,738$ ).

#### 4.2.2 Wyniki regresji wielorakiej

Analizując wybrane czynniki ciągle związane z wielkością HR częstoskurczu wykorzystano regresję wieloraką. W sposób oczywisty HR częstoskurczu tłumaczą modele, które zawierają wiek.

Tabela 11. Wybrane modele regresji wielorakiej dla HR częstoskurczu zawierające wiek i dodatkowe zmienne.

modele	F	SE	p testu F (istotność całego modelu)	N	Norma Iność reszt (p testu Lillieforsa)	R wielorakie	R <sup>2</sup> skorygowane	porównanie modeli z wersją zredukowaną (test F)
<b>A1</b>	49,18	28,70	<0,0000001*	425	0,389	0,32	10,20%	-
<b>B</b>	25,36	28,68	<0,0000001*	425	0,249	0,33	10,31%	0,2250
<b>C</b>	25,79	28,66	<0,0000001*	425	0,420	0,33	10,47%	0,1343
<b>A2</b>	21,05	26,32	0,000016*	83	0,250	0,45	19,7%	-

\*wynik istotny statystycznie (<0,05); SE - błąd standardowy estymacji modelu

A1 – model podstawowy zawierający jedynie wiek (dla całości);

A2 - model podstawowy dla mężczyzn z AVNRT;

B – model rozszerzony zawierający płeć (dla całości);

C – model rozszerzony zawierający rodzaj arytmii (dla całości)

W Tabeli 11 przedstawiono wybrane modele: podstawowe - typu A, które zawierają jedynie wiek oraz modele rozszerzone (B i C), zawierające dodatkowe zmienne niezależne. Wykazano, iż wytłumaczalność częstotliwości HR częstoskurczu wiekiem nie zwiększa się w sposób istotny, gdy dodane zostaną informacje o płci, czy o rodzaju arytmii. Następnie, ze względu na silną korelację wieku i HR częstoskurczu w najstarszej podgrupie, tzn. w grupie mężczyzn z AVNRT, analizowane modele regresji wielorakiej przeliczono zarówno dla całej populacji, jak i dla tej specyficznej podgrupy.

W Tab. XIX (Załącznik nr 1) zaprezentowano szczegółowe wyniki wszystkich modeli regresji wielorakiej badającej HR częstoskurczu standaryzowanego. Analizowane dodatkowe zmienne niezależne to: BMI, TSH, FT4, FT3, Na<sup>+</sup>, K<sup>+</sup>, LK, LP, PK, dolno-boczna ściana LK, PMK oraz Ao. Okazuje się, iż dodanie tych zmiennych niezależnych do zmiennej wiek, nie zwiększa jakości powstałych modeli. Porównań modeli pełnych z modelami po usunięciu dodatkowej zmiennej dokonano testem F, a test badający istotność poszczególnych składników modelu wykazał, w każdym z badanych przypadków, istotność jedynie dla zmiennej wiek. Przeanalizowane dodatkowe zmienne były skorelowane zarówno z HR częstoskurczu, jak również (słabo) z wiekiem (Załącznik nr 1).

### **4.3 Czynniki ryzyka omdleń**

#### **4.3.1 Statystyki jednowymiarowe:**

Poszukując czynników ryzyka omdleń, spośród zebranych danych klinicznych oraz wyników badania elektrofizjologicznego serca (EP), w niniejszej pracy przeprowadzono szereg analiz porównawczych dla zmiennych ciągłych oraz przeanalizowano tabele wielodzienne związane z czynnikami opisanymi na skali nominalnej. Szczegółowe wyniki tych analiz oraz niezbędne charakterystyki opisowe przedstawiono w Załączniku nr 1 (Tab. IV i Tab. V). W Tabeli 12 zaprezentowano jedynie te czynniki, które w omawianych analizach uznane zostały za zwiększające szansę wystąpienia epizodów omdleń. U mężczyzn mających wyższe BMI, istotnie częściej występowały niepożądane epizody omdleń ( $p=0,038257$ ), w porównaniu z mężczyznami mającymi BMI niższe. Grupa zwiększonego ryzyka występowania omdleń charakteryzowała się także wyższymi wartościami K<sup>+</sup> w porównaniu z grupą bez omdleń ( $p=0,026134$ ). Pacjenci ci mieli także większe LP ( $p=0,045787$ ) i PK ( $p=0,041396$ ). Natomiast grupa kobiet z omdleniami miała istotnie niższe wartości TSH ( $p=0,020533$ ) w porównaniu z grupą bez omdleń.



Dla czynników mierzonych na skali interwałowej wyznaczono krzywą ROC oraz przetestowano pole pod tą krzywą (Tabela 12 dla czynników interwałowych zawiera wartości pola pod krzywą ROC wraz z przedziałami ufności.). Zarówno BMI dla mężczyzn, jak i K+, LP i PK dla całej grupy, okazały się istotnymi stymulantami – wzrost wartości tych zmiennych prognozował większą szansę wystąpienia epizodu omdlenia. TSH dla kobiet uznano za destymulantę – ze wzrostem wartości TSH u kobiet spada szansa wystąpienia u nich epizodu omdlenia. Wszystkie pięć opisanych w Tabeli 12 krzywych ROC miało pola pod krzywą istotnie większe od 0,5.

W przypadku istotnych czynników ryzyka mierzonych na skali nominalnej, Tabela 12 zawiera ilorazy szans (OR) określające zwiększoną szansę wystąpienia omdleń w poszczególnych podgrupach wraz z odpowiadającymi im przedziałami ufności. Występowanie stanów przedomdleniowych istotnie zwiększało występowanie omdlenia zarówno dla całej grupy ( $p=0,00149$ ), jak i dla kobiet ( $p=0,00363$ ) - Załącznik nr 1 (Tab. IV). Iloraz szans dla całej grupy wynosił 3,75, a więc okazuje się, że szansa omdlenia osób mających w przeszłości stany przedomdleniowe jest niemal 4-krotnie wyższa. Natomiast w podgrupie kobiet szansa ta wzrasta aż do 6,65. Mężczyźni z krótszym (poniżej 280,5 ms) CL częstoskurczu, mierzonym podczas badania EP (a więc ewentualnie obciążonym podaniem leku), mieli ponad 4-krotnie większe szanse na wystąpienie omdlenia. Krótszy CL częstoskurczu oznacza wyższą częstotliwość HR częstoskurczu. (Szczegóły wszystkich analiz znajdują się w dodatku: Załącznik nr 1; Tab. IV i Tab. V)

Tabela 12. Podsumowanie czynników ryzyka różnicujących grupy z omdleniami i bez omdleń w zależności od płci.

Czynniki różnicujące grupy z i bez omdlenia	Razem (kobiety i mężczyźni)	kobiety	mężczyźni
<b>BMI</b> (stymulanta)	-	-	<b>AUC=0,61</b> (0,51;0,71); p= 0,0381* pkt.odcięcia 31,44
<b>K+</b> (stymulanta)	<b>AUC=0,58</b> (0,51;0,64); p= 0,0263* pkt.odcięcia 4,33	-	-
<b>LP</b> (stymulanta)	<b>AUC=0,57</b> (0,50;0,64); p= 0,0461* pkt.odcięcia 42	-	-
<b>PK</b> (stymulanta)	<b>AUC=0,57</b> (0,50;0,64); p= 0,0420* pkt.odcięcia 30	-	-
<b>TSH</b> (destymulanta)	-	<b>AUC=0,63</b> (0,53;0,72); p= 0,0204* pkt.odcięcia 1,79	-
stany przedomdleniowe	OR=3,75 (1,56-10,9); p=0,00149**	OR=6,65 (1,56-28,44) p=0,00363**	-
krótszy cykl częstoskurczu (CL) z EP #	-	-	OR=4,66 (1,83;11,84) p=0,00059**

\* wynik istotny statystycznie (<0,05); p-testu sprawdzającego, czy AUC jest istotnie większe niż pole pod prostą  $y = x$ , czyli niż 0,5.

\*\* wynik istotny statystycznie (<0,05); p-testu  $\chi^2$

# „krótszy” cykl częstoskurczu oznacza grupę z  $CL \leq 280,5$ ;

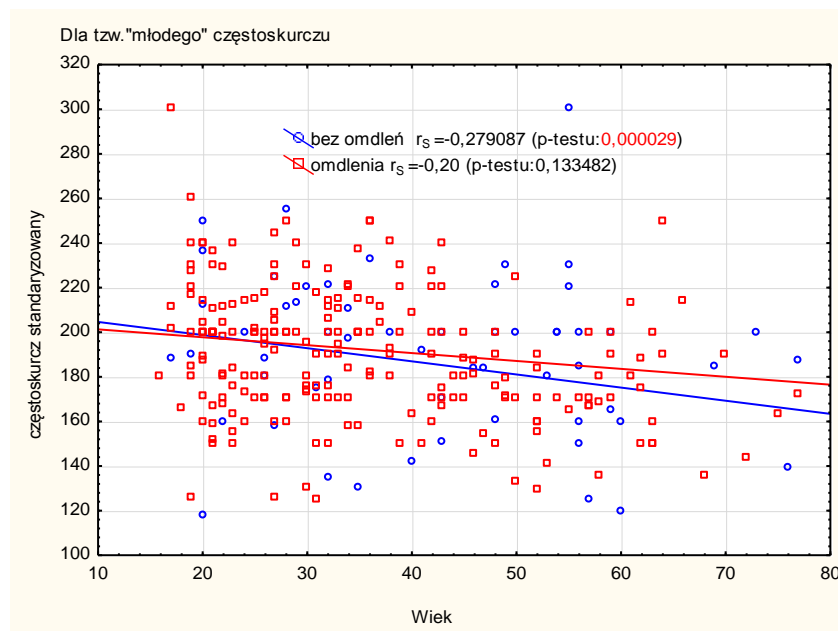
„-” oznacza brak istotnego czynnika ryzyka w analizowanej podgrupie;

istotny czynnik ryzyka występowania epizodów omdleń w analizowanej populacji opisano dla parametrów mierzonych w skali interwałowej za pomocą pola powierzchni pod krzywą ROC (AUC wraz z przedziałami ufności), natomiast dla parametrów mierzonych w skali nominalnej za pomocą ilorazu szans, wraz z przedziałami ufności.

Pozostałe analizowane zmienne nie różnicowały epizodów omdleń w sposób bezpośredni, jednak ze względu na ich szczególne znaczenie kliniczne oraz ewentualne powiązania z częstotliwością HR częstoskurczu, przeanalizowano je dokładniej. W kolejnych podrozdziałach opisano ścieżki analizy tych zmiennych.

#### 4.3.1.1 Czasokres od pierwszego udokumentowanego częstoskurczu

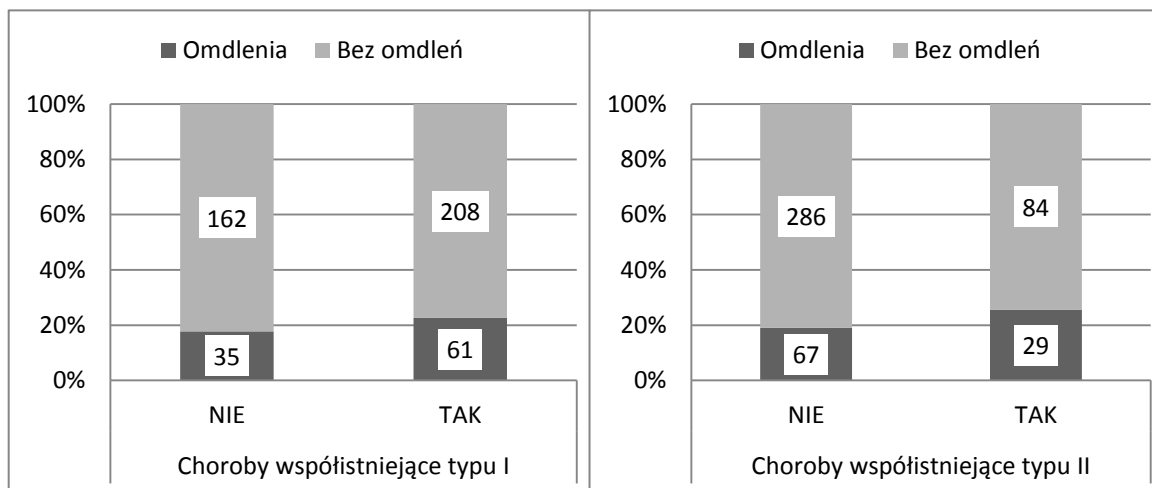
Najwięcej osób miało „krótkie” (stosunkowo krótko trwające, tzn.  $\leq 20$  lat) czasokresy częstoskurczów (68,85%). Grupy z „dłuższym” i „krótszym” czasokresem objawów kołatania i częstoskurczu nie różnicowały częstości występowania omdleń w analizowanej grupie (test  $\chi^2$ :  $p=0,54136$ , Załącznik nr 1). Zaobserwowano natomiast różnice w zależnościach HR częstoskurczu od wieku pacjenta, w grupie z omdleniem i bez omdleń, w zależności od grupy czasokresu częstoskurczu (Rycina 7). Znany z literatury spadek wartości HR częstoskurczu wraz z wiekiem pacjenta widoczny jest, bez względu na czasokres częstoskurczu, jedynie dla osób bez omdleń. Natomiast grupa, w której występowały epizody omdleń, a jednocześnie pierwszy częstoskurcz pojawił się nie dalej niż 20 lat temu (był to więc „krótszy” czasokres częstoskurczu), nie ma istotnej statystycznie zależności HR częstoskurczu od wieku badanych.



Rycina 17. Zależność HR częstoskurczu od wieku pacjenta w grupie z „krótszym” czasokresem częstoskurczu dla grupy z omdleniami i bez omdleń.

#### 4.3.1.2 Dodatkowe choroby współistniejące

Analizując podgrupy ze współistniejącymi jawnymi chorobami dodatkowymi, wobec osób bez objawów chorób współwystępujących (Rycina 18), nie stwierdzono istotnej statystycznie różnicy w występowaniu epizodów omdleń. (Test Chi<sup>2</sup> dla chorób typu I: p=0,19544 oraz dla chorób typu II: p=0,12628; Załącznik nr 1; Tab. IV)



Rycina 18. Epizody omdleń wobec współwystępujących dodatkowych objawów innej choroby.

W grupie z jawnymi objawami chorób współistniejących były osoby istotnie starsze, co tłumaczy niższe wartości HR częstoskurczów w tej grupie. Natomiast analizując HR częstoskurczu niestandardyzowanego, ale podwyższonego poprzez podanie leku podczas badania EP, okazało się, że HR częstoskurczu było istotnie szybsze w grupie z epizodami omdleń pod warunkiem, że osoby te nie miały dodatkowej jawnej choroby współistniejącej. Pamiętając o zmiennej wnikającej – wiek, wykonano także analizy dodatkowe, które wykazały, iż grupa z omdleniami i bez (w przypadku braku jawnych objawów chorób dodatkowych) nie różni się istotnie wiekiem.

Tabela 13. Częstotliwość HR częstoskurczu po leku wobec epizodów omdleń dla obu arytmii razem.

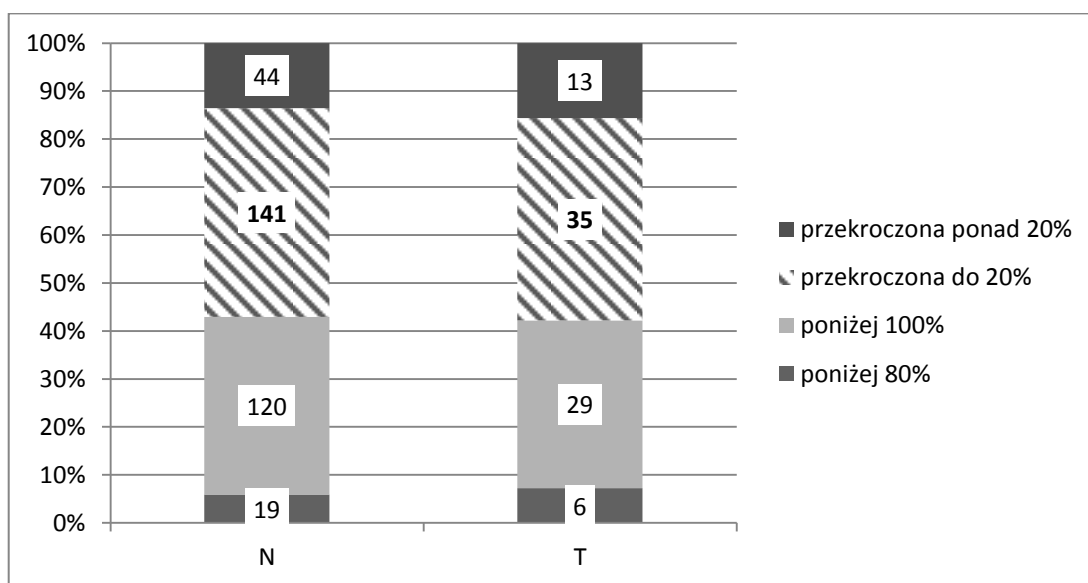
		omdlenia	brak omdleń	Porównanie dwóch grup (p testu Manna-Whitneya)
bez objawów jakiegokolwiek dodatkowej choroby jawnej	Liczność	8	30	
	Średnia	231,63	204,17	p=0,010*
	Mediana	232	202,5	
współistniejąca choroba	Liczność	13	50	
	Średnia	181,85	189,58	p=0,329
	Mediana	177	186,5	
bez względu na chorobę współistniejącą	Liczność	21	80	
	Średnia	200,81	195,05	p=0,555
	Mediana	208	191,5	

\* wynik istotny statystycznie (p<0,05)

Szybsze częstoskurcze nie były zatem tłumaczone młodszym wiekiem, więc wyższa częstotliwość HR częstoskurczu po leku mogłaby być zidentyfikowana jako czynnik ryzyka omdlenia w grupie bez dodatkowej jawnej choroby współistniejącej. W Tabeli 13 przedstawiono niektóre wyniki porównań dla obu arytmii łącznie, natomiast szczegóły zamieszczono w Załączniku nr 1.

Przedstawiona w niniejszej pracy ocena częstotliwości HR częstoskurczu, po podaniu leku w czasie badania EP, prezentowana została we Wrocławiu, na XXIV Konferencji Sekcji Rytmu Serca Polskiego Towarzystwa Kardiologicznego (POLSTIM 2013) [90], gdzie przedstawione w dyskusji wnioski stanowiły pierwszą, publiczną prezentację częściowych wyników z niniejszej (pisanej wówczas) pracy.

Należna wartość częstoskurczu posłużyła do podziału danych na 4 podgrupy: dwie bliższe wartości należnej HR (w zakresie 20%) oraz dwie dalsze. Na Rycinie 19 można zauważyć, iż dwie środkowe grupy częstoskurczu są najbardziej liczne zarówno pośród osób z odnotowanymi epizodami omdleń, jak i bez nich.



Rycina 19. Liczności, w 4 podgrupach, wg należnego częstoskurczu, w zależności od występowania (T) lub nie (N) epizodów omdleń.

Rycina 19Tabela 14 ukazuje, iż w grupie AVNRT, gdzie przekroczono wartość częstoskurczu więcej niż o 20% względem wartości należnej, odsetek osób z odnotowanymi epizodami omdleń był największy (20,93%). W grupie WPW najwięcej osób z omdleniami było w grupie, u której odnotowano częstoskurcze dużo wolniejsze (o więcej niż 20%) niż max wartość przewidywana (36,36%).

Tabela 14. Należna wartość częstoskurczu w 4 podgrupach; w zależności od rodzaju arytmii i wystąpienia epizodów omdleń.

Rodzaj arytmii	wartość HR częstoskurczu (4 grupy wg wartości należnej)	Omdlenie	
		T	N
AVNRT	całość	<b>17,84%</b>	<b>78,81%</b>
	poniżej 80%	12,50%	81,25%
	poniżej 100%	15,96%	78,72%
	przekroczona do 20%	18,97%	80,17%
	przekroczona ponad 20%	20,93%	74,42%
	p testu Fishera-Freemana-Haltona	0,8657	
WPW	całość	<b>22,44%</b>	<b>71,79%</b>
	poniżej 80%	36,36%	54,55%
	poniżej 100%	21,88%	71,88%
	przekroczona do 20%	20,31%	75,00%
	przekroczona ponad 20%	23,53%	70,59%
	p testu Fishera-Freemana-Haltona	0,6024	

Analizując proporcje osób z epizodami omdleń w 4 grupach należnego częstoskurczu, test Fishera-Freemana-Haltona nie wykazał zależności pomiędzy rodzajem grupy częstoskurczu a omdleniem: zarówno w całej grupie ( $p=0,916213$ ), jak i osobno dla kobiet i mężczyzn ( $p= 0,9043$  i  $p= 0,1745$  odpowiednio). Wyniki te znajdują się w Załączniku nr 1; Tab. IV). Ponadto przeanalizowano zależność grupy częstoskurczu i epizodów omdleń także osobno dla obu rodzajów arytmii – Tabela 14. Również w tym wypadku zależności te okazały się nieistotne statystycznie (p testu Fishera-Freemana-Haltona dla AVNRT wynosiło 0,8657, natomiast dla WPW 0,6024).

#### 4.3.1.3 Czas trwania kołatania

W Tabeli 15 przedstawiono charakterystykę czasu kołatania serca wobec występowania epizodów omdleń w grupie arytmii WPW. Niższe wartości oznaczają krótsze czasy kołatania (patrz rozdział 3.2.4.11).

Tabela 15. Mediana czasu kołatania w zależności od rodzaju arytmii.

	Wstępny czas kołatania	Końcowy czas kołatania	Max czas kołatania
Bez omdleń	1	3	90
Omdlenie	1	3	90

Porównywane grupy (z i bez omdleń) nie różniły się istotnie czasem trwania kołatania (test Manna-Whitneya). Wykazano więc brak istotnych statystycznie różnic w czasie kołatania pomiędzy pacjentami z i bez odnotowanych epizodów omdleń.

Przechodząc na skalę nominalną i stosując dokładne testy Fishera-Freemana-Haltona również nie odnotowano zależności pomiędzy czasem trwania kołatania a wystąpieniem epizodów omdleń. W przypadku maksymalnego czasu trwania kołatania dokonano zgrupowania danych do 3 grup (do 120 min, 120 - 360 min oraz ponad 360 min), i przetestowano również testem Fishera-Freemana-Haltona oraz porównano parami poszczególne częstości występowania omdleń w podgrupach mających różny max czas kołatania (testem Z dla dwóch niezależnych proporcji) – nigdzie nie wykazano związku pomiędzy czasem kołatania a omdleniem. Następnie dokonano podziału binarnego dla wstępnego i końcowego czasu kołatania. Przekodowanie dla czasu wstępnego dokonano wg schematu: kody równe 1 to „krótkie” epizody, pozostałe zaklasyfikowano jako „długie”. W przypadku max czasu kołatania: czasy  $\leq 120$  min zaklasyfikowano jako „krótkie” epizody, pozostałe przypadki – jako „długie”. Zastosowanie dokładnego testu Fishera również nie wykazało istotnej statystycznie zależności występowania omdleń i długości trwania kołatania.

#### ***4.3.1.4 Pora dnia występowania częstoskurczu***

Analizowane dwa rodzaje arytmii różniły się istotnie rolą układu współczulnego przy pojawiających się częstoskurczach: w przypadku WPW prawie wszystkie odnotowane przypadki częstoskurczów pojawiały się w dzień (95,25%), natomiast w grupie AVNRT takie przypadki stanowiły zaledwie 71,93%. Opisano także wpływ układu współczulnego na wartość częstotliwości HR częstoskurczu (rozdział 4.2.1).

W niniejszej pracy nie wykazano istotnej statystycznie zależności pomiędzy porą występowania częstoskurczów a epizodami omdleń (test Fishera-Freemana-Haltona  $p=0,6076$ ; Załącznik nr 1; Tab. IV).

### 4.3.2 Modele regresji logistycznej

W celu oceny potencjalnych czynników ryzyka wystąpienia epizodów omdleń, w grupie ze zdiagnozowaną arytmia, przeprowadzono liczne analizy regresji logistycznej. Do budowy modeli wykorzystywano po kolei cechy podejrzewane przez klinicystów o związek z występowaniem omdleń w analizowanej grupie oraz cechy sugerowane przez wyniki statystyk prostych. Szczegóły wyliczonych modeli przedstawiono w Załączniku nr 1. Przy konstruowaniu modeli szczególnie istotne było przestrzeganie założeń o liczbie obserwacji, która to powinna być dziesięciokrotnie większa lub równa liczbie szacowanych parametrów modelu. Ponieważ cały analizowany zbiór danych zawierał jedynie 466 przypadków, dla których zdefiniowano zmienną opisującą występowanie (lub nie) w przeszłości epizodów omdleń, w analizie regresji logistycznej można było brać jednocześnie pod uwagę co najwyżej trzy zmienne.

Tabela 16. Wybrane modele logistyczne opisujące wpływ wieku i częstotliwości HR częstoskurczu na omdlenia w zależności od czasookresu od pierwszego udokumentowanego częstoskurczu.

Model logistyczny	p testu ilorazu wiarygodności	R2 Nagelkerke	R2 Coxa-Snella	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)
<b>A</b>	0,12785	1,580%	1,006%	częstoskurcz standaryzowany	1,0025	0,57150
				wiek	1,0171	0,04375*
<b>A dla czasookresu (&lt;20lat)</b>	0,00888*	5,280%	3,377%	częstoskurcz standaryzowany	1,0046	0,38008
				wiek	1,0320	0,00225*
<b>A dla czasookresu (&gt;=20lat)</b>	0,60263	1,215%	0,764%	częstoskurcz standaryzowany	0,9940	0,4903
				wiek	0,9786	0,3353

„częstoskurcz standaryzowany” – standaryzowana wartość częstotliwości HR częstoskurczu

\*wynik istotny statystycznie (p<0,05)

Wiek pacjenta nie różnicował grupy ryzyka wystąpienia epizodu omdleń (p=0,199817; Załącznik nr 1). Jednak dodanie informacji o częstotliwości HR częstoskurczu standaryzowanego, skutkowało zwiększeniem roli zmiennej wiek w modelu opisującym omdlenia (model A w Tabeli 16). Mimo, iż cały model nie był istotny (p testu ilorazu wiarygodności: 0,127847), zmienna wiek wykazywała istotny wpływ na zmienną zależną, czyli wystąpienie omdlenia (wartość p testu Walda: 0,043745). Dla takiego samego częstoskurczu, zwiększenie wieku o 1 rok, zwiększa szansę wystąpienia omdlenia o 1,017. W przypadku zwiększenia wieku o 40 lat będzie to już ok. 2-krotnie większa szansa wystąpienia omdlenia.

$$OR = (e^{1,017})^{40} = 1,965$$



W przypadku podgrupy krótszego czasokresu częstoskurczu (czasokres od pierwszego udokumentowanego częstoskurczu poniżej 20 lat), analogiczny model był istotny statystycznie i rola wieku wzrastała (p testu ilorazu wiarygodności wynosiło 0,00888; Tabela 16). Zwiększenie wieku o 40 lat powoduje ok. 3,5-krotnie większą szansę wystąpienia omdlenia.

Tabela 17. Wybrane modele logistyczne opisujące wpływ pojawiających się stanów przedomdleniowych i częstotliwości HR częstoskurczu na omdlenia w zależności od płci oraz współwystępowania innych jawnych chorób.

Model logistyczny	p testu ilorazu wiarygodności	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)
<b>F</b>	0,00019*	częstoskurcz standaryzowany	0,9978	0,62986
		stany przedomdleniowe	6,9754	0,00137*
<b>F dla M</b>	0,08775	częstoskurcz standaryzowany	1,0019	0,7998
		stany przedomdleniowe	4,1052	0,0659
<b>F dla K</b>	0,00117*	częstoskurcz standaryzowany	0,9952	0,4088
		stany przedomdleniowe	12,6904	0,0136*
<b>F dla choroby typu I brak</b>	0,10300	częstoskurcz standaryzowany	0,9939	0,4265
		stany przedomdleniowe	5,5337	0,1033
<b>F dla choroby typu I TAK</b>	0,00094*	częstoskurcz standaryzowany	1,0023	0,6907
		stany przedomdleniowe	7,9734	0,0053*
<b>F1<sup>#</sup></b>	0,00006*	częstoskurcz standaryzowany	0,9992	0,8671
		stany przedomdleniowe	7,1520	0,0012*
		Choroby typ I	1,8762	0,0287*
<b>F2</b>	0,00065*	częstoskurcz standaryzowany	0,9978	0,6220
		stany przedomdleniowe	6,9636	0,0014*
		płeć	1,0360	0,8983

„częstoskurcz standaryzowany” – standaryzowana wartość częstotliwości HR częstoskurczu

<sup>#</sup> istotne statystycznie różnice z modelem zredukowanym (odjęcie ostatniej zmiennej); test ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele;

\*wynik istotny statystycznie (p<0,05)

Tabela 17 pokazuje, iż prawie wszystkie modele zawierające informacje o stanach poprzedzających omdlenia były istotne statystycznie. Model F zawierał dwie zmienne niezależne: HR częstoskurczu standaryzowanego oraz informację o występowaniu stanów poprzedzających omdlenia. Porównanie z modelem zredukowanym, w którym pozostała jedynie jedna zmienna: informacja o występowaniu stanów poprzedzających omdlenia, informuje, iż nie ma istotnych statystycznie różnic pomiędzy tymi modelami (p-testu ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele) wynosiło

0,62953; (Załącznik nr 1; Tab. XIX). Oznacza to, iż dodanie informacji o częstotliwości HR częstoskurczu nie wnosi nic istotnego do modelu. Jednak OR dla stanów przedomdleniowych w modelu prostym (jednowymiarowym) wynosiło 3,75, natomiast w analizowanym modelu F, z dodaną informacją o HR częstoskurczu, to jest aż 6,97. Model ten pokazuje zatem niemal 12-krotnie większą szansę wystąpienia omdlenia dla kobiet nieróżniących się HR częstoskurczu, u których wystąpiły stany poprzedzające omdlenia. W przypadku występowania jakichkolwiek dodatkowych, jawnych chorób współistniejących szanse te oszacowano na: OR=7,9.

Porównując model F1 z jego wersją zredukowaną F, wykazano, iż dodanie informacji o współistniejących chorobach istotnie polepsza wykrywalność omdleń (porównanie modeli za pomocą testu ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele:  $p=0,0253$ ; (Załącznik nr 1; Tab. XIX);  $R^2$  Nagelkerke wzrasta z 7% do 9%). Dodanie do modelu F informacji o płci nie wносило istotnie więcej informacji.

#### ***4.3.2.1 Modele logistyczne po zastosowaniu wstępnej analizy danych***

Wstępne przetwarzanie danych obejmuje czyszczenie i przekształcanie danych w celu ich przygotowania do eksploracji. Motywacje wstępnego przetwarzania (obróbki) danych, to między innymi brakujące wartości (ang. missing values) lub obecność punktów oddalonych (ang. outliers), a także potrzeba wyeliminowania tzw. szumów informacyjnych. W celu dalszego modelowania za pomocą regresji logistycznej zastosowano: wybór predyktorów (rozdział 4.1.2.1), reprezentantów (rozdział 4.1.2.2) oraz dyskretyzację zmiennych (rozdział 4.1.2.3).

##### **4.3.2.1.1 Próba analizy wszystkich zmiennych**

Ze zbioru zmiennych wyłonionych przed dyskretyzacją (w rozdziale 4.1.2.2), wybrano 15 zmiennych (Tabela 18). Są to zmienne związane z danymi zgromadzonymi z wywiadu, przed wykonaniem badania EP.

Test liniowości sprawdza, czy poprawa jakości modelu po dodaniu kwadratu wartości predyktora jest istotna statystycznie. Wyniki testu sprawdzającego liniowy wpływ zmiennych na występowanie epizodów omdleń zawarto w Tabeli 19. Jedynie dla zmiennej TSH, odrzucona została hipoteza o liniowości..

Tabela 18. Zmienne użyte do analizy regresji logistycznej.

	Nazwa zmiennej
1	Wiek
2	Choroby współistniejące typu I
3	Choroby współistniejące typu II
4	K <sup>+</sup> (mmol/l)
5	Na <sup>+</sup> (mmol/l)
6	Ao (mm)
7	LK (mm)
8	LP (mm)
9	PK (mm)
10	TSH (mU/l)
11	grupa CL
12	kategorie BMI 2
13	Pora dnia
14	częstoskurcz standaryzowany
15	różnica należąca HR i z standaryzowanego

Tabela 19. Badanie współliniowości zmiennych wybranych do analizy.

	Nazwa zmiennej	p testu na liniowość
1	LP (mm)	0,858
2	PK (mm)	1,000
3	K <sup>+</sup> (mmol/l)	0,100
4	Wiek	0,470
5	Ao (mm)	0,649
6	LK (mm)	0,343
7	różnica należąca HR i z standaryzowanego	0,133
8	Na <sup>+</sup> (mmol/l)	0,203
9	TSH (mU/l)	0,023*
10	częstoskurcz standaryzowany	0,151

\*wynik istotny statystycznie (p<0,05)

Aby wykryć tzw. efekty interakcji (tzn. wyrażenia iloczynowe dwóch zmiennych), poszukiwano za pomocą testu ilorazu wiarygodności (LR), istotnych interakcji pomiędzy potencjalnymi predyktorami.

Na podstawie wyliczonych interakcji wybrano 11 istotnych statystycznie, dotyczących analizowanych zmiennych – Tabela 20. Wszystkie interakcje przedstawiono w Załączniku nr 1 (Tab.XXII).

Tabela 20. Interakcje istotne statystycznie dla zmiennych wybranych w procesie wstępnej analizy danych.

	Zmienna 1	Zmienna 2	p (LR)
1	Wiek	TSH (mU/l)	0,0002*
2	LP (mm)	TSH (mU/l)	0,0039*
3	Ao (mm)	TSH (mU/l)	0,0051*
4	grupa CL_kat	K+ (mmol/l)	0,0254*
5	grupa CL_kat	LP (mm)	0,0258*
6	grupa CL_kat	LK (mm)	0,0328*
7	kat2_kat	TSH (mU/l)	0,035*
8	Pora dnia	LK (mm)	0,0361*
9	Pora dnia	LP (mm)	0,0388*
10	Pora dnia	PK (mm)	0,0287*
11	LP (mm)	PK (mm)	0,012*

\*wynik istotny statystycznie ( $p < 0,05$ )

„\_kat” oznacza zmienną kategorią, po dyskretyzacji.

Ponieważ w analizowanej bazie występowało wiele braków danych związanych ze zmienną TSH (jedynie 260 osób, spośród tych, u których opisano omdlenia, bądź ich brak, miało zapisaną informację o wysokości poziomu TSH), postanowiono uwzględniać tą zmienną jedynie w wybranych analizach.

Z uwagi na ciągle dużą liczbę zmiennych, proces budowy modelu próbowano przeprowadzać za pomocą regresji postępującej, przy wykorzystaniu mechanizmu walidacji opartego na V-krotnym sprawdzanie krzyżowym. Jednak żadne zmienne w modelu nie okazały się istotne – nie można było zbudować modelu.

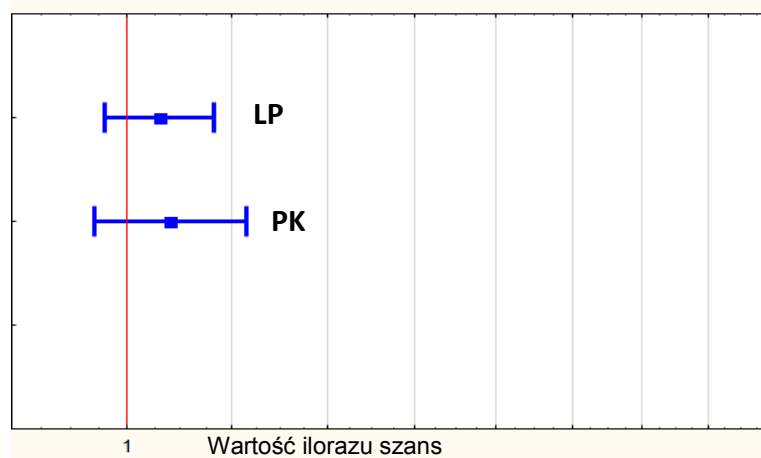
#### 4.3.2.1.2 Analiza zmiennych wybranych przez analizy jednoczynnikowe

Analizując, wybrane w omówionych krokach analizy wstępnej zmienne, wykazano (Załącznik nr 1), iż jedynie zmienne ciągle: LP i PK były istotnie lepsze w modelach jednoczynnikowych od modeli wyłącznie z wyrazem wolnym (test ilorazu wiarygodności LR:  $p = 0,0362$  oraz  $p = 0,0402$  odpowiednio dla zmiennych LP i PK).

Żadna ze zmiennych po dyskretyzacji nie wykazywała istotnego wpływu na analizowane występowanie omdlenia ( $p$  testu LR dla LP\_kat 0,1092), choć najbliższej istotności była zmienna PK po dyskretyzacji, czyli PK\_kat ( $p = 0,0607$ ). Przed przystąpieniem do analizy zmienna mająca wiele podkategorii jest rozbijana na kilka zmiennych fikcyjnych o 2 kategoriach (tak/nie). Dlatego w modelach utworzonych dla zmiennych dyskretyzowanych było tyle

nowopowstałych dodatkowych zmiennych, ile przedziałów/kategorii miała dana zmienna po dyskretyzacji. Niestety, często powodowało to brak możliwości wykonania obliczeń regresji logistycznej, ze względu na zbyt małą liczbę obserwacji, która to powinna być dziesięciokrotnie większa lub równa liczbie szacowanych parametrów modelu.

Mając na uwadze wyniki wstępnej analizy jednoczynnikowej utworzono model regresji logistycznej zawierający zmienne LP i PK (model P; Załącznik nr 1; Tab.XIX: p testu ilorazu wiarygodności równe 0,05047). Następnie przeanalizowano również model zawierający dodatkowo interakcję tych dwóch zmiennych – model P1 (istotność tej interakcji sprawdzona testem ilorazu wiarygodności LR:  $p=0,012$ ; Tabela 19). Proces budowy modelu przeprowadzony został za pomocą regresji rozpatrującej wszystkie efekty, przy wykorzystaniu mechanizmu walidacji opartego na V-krotnym sprawdzanie krzyżowym. Wyniki omawianych modeli znajdują się w Załączniku nr 1 (model P i P1 – Tab. XIX i Tab. XX). Model zawierający interakcję okazał się istotnie lepszy, od tego bez interakcji (porównanie modeli P i P1 za pomocą testu ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele:  $p= 0,012$ ;  $R^2$  Nagelkerke wzrasta z 2,1% do 4,3%) i w przeciwieństwie do modelu mniejszego był istotny statystycznie ( $p = 0,00647$ ).



Rycina 20. Ilorazy szans zmiennych LP i PK z modelu logistycznego P, określające brak wpływu na występowanie epizodów omdleń. Model nieistotny statystycznie ( $p=0,05047$ ).

Rycina 20 obrazuje wyniki uzyskane w modelu P: zmienne są nieistotne statystycznie (przedział ufności dla ilorazu szans zawiera jedynkę), co oznacza, że zmienne te nie zwiększają, ani nie zmniejszają szansy na wystąpienie epizodów omdleń. Natomiast model P1 informuje o wzroście szansy na wystąpienie omdlenia przy wyższych wartościach LP i PK, jednak wpływ ten modyfikowany jest przez istotną interakcję pomiędzy tymi dwoma zmiennymi. Model bez interakcji (model P) niedoszacowuje prawdopodobieństwa omdlenia,

natomiast dodanie efektu interakcji (model P1) pozwala lepiej przewidywać prawdopodobieństwo omdlenia niż same tylko efekty główne (LP i PK). Czułość modelu z interakcją wzrosła z 0% na 4,6%, natomiast swoistość z 99,7% na 100%. Model z interakcją był lepiej dopasowany.

Jeśli między dwoma zmiennymi zachodzi istotna statystycznie interakcja, wartość ilorazu szans dla jednej zmiennej różni się w zależności od konkretnych wartości drugiej zmiennej, wchodzącej z nią w interakcję. Dlatego w sytuacji interakcyjnej wartości ilorazów szans oblicza się zawsze OR dla konkretnej wartości zmiennych wchodzących w interakcje. Nie można interpretować ilorazów szans obliczonych dla modelu P1 (w Załączniku nr 1; Tab. XIX), zamiast tego obliczono ilorazy szans dla dwóch punktów związanych z dwoma grupami LP. Tzn. dla wartości „niższych LP” (10-ty percentyl = 26) i „wyższych LP” (90-ty percentyl = 39). Do obliczenia tych ilorazów szans wykorzystano współczynniki regresji modelu P1 przedstawione w Tabeli 21.

Tabela 21. Współczynniki regresji wybranych modeli

Model	Zmienne w modelu	Wartości $\beta$	wartość p dla zmiennej (test Walda)
P1	LP (mm)	0,422	0,00970*
	PK (mm)	0,546	0,00990*
	LP x PK	-0,015	0,01576*
T1 dla kobiet	Wiek	-0,053	0,00894*
	TSH	-2,314	0,00281*
	TSH x wiek	0,039	0,00249*

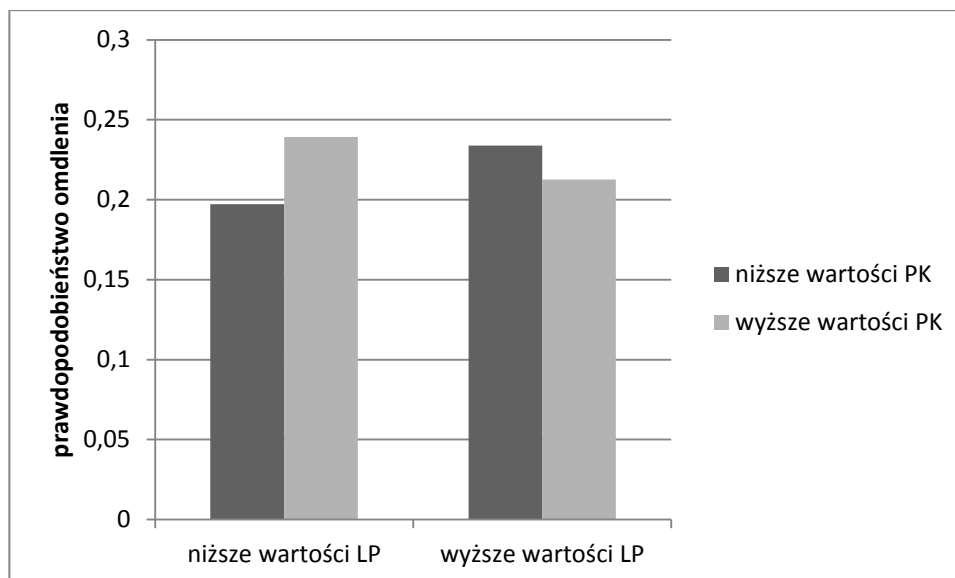
\* wynik istotny statystycznie ( $p < 0,05$ )

Skorygowany iloraz szans dla dwóch punktów odniesienia (LP = 26 i LP = 39) wyliczono z uwzględnieniem wpływu zmiennej moderującej PK.

$$OR_{(26)} = e^{0,5461 + (-0,01554 \cdot 26)} = 1,1526$$

$$OR_{(39)} = e^{0,5461 + (-0,01554 \cdot 39)} = 0,9418$$

Na Rycinie 21 zaprezentowano średnie prawdopodobieństwo wystąpienia omdlenia w dwóch grupach LP w zależności od wartości PK. Grupy LP zostały określone przez „wartości niższe”, tzn. poniżej mediany (Me=32; Załącznik nr 1; Tab. I) oraz „wyższe” od mediany. Zmienną PK również podzielono na wartości „niższe” i „wyższe”, utworzone poprzez podział wg mediany (Me=24; Załącznik nr 1; Tab. I). Dla niższych wartości LP wyraźnie prawdopodobieństwo omdlenia wzrastało wraz ze wzrostem PK, natomiast dla wyższych wartości LP obserwowalna zależność była odwrotna.



Rycina 21. Prawdopodobieństwo wystąpienia omdlenia w podgrupach z wyższymi i niższymi wartościami LP i PK.

Kolejną analizowaną zmienną było TSH, które jednoznacznie różnicowało grupę ryzyka omdlenia w przypadku kobiet (Załącznik nr 1; Tab. V). Co prawda, analizując ten czynnik ryzyka (Załącznik nr 1) testem ilorazu wiarygodności LR, czynnik TSH nie okazał się istotnie lepszy w modelu jednoczynnikowym od modelu wyłącznie z wyrazem wolnym (LR:  $p=0,9375$ ). Jednak, podejrzewając zmienną wiek o pewien rodzaj uwikłania w ten związek TSH i zjawiska omdleń, postanowiono utworzyć model wielowymiarowy (model T) zawierający dwie zmienne TSH i wiek. Przeanalizowano również model zawierający dodatkowo interakcję tych dwóch zmiennych – model T1 (istotność tej interakcji sprawdzona została testem ilorazu wiarygodności LR:  $p=0,0002$ ; Tabela 20). Oba modele: T i T1 zostały utworzone również dla zawężonej grupy – tylko dla kobiet. Proces budowy modelu przeprowadzony został tak, jak w pozostałych modelach, tzn. za pomocą regresji rozpatrującej wszystkie efekty, przy wykorzystaniu mechanizmu walidacji opartego na V-krotnym sprawdzaniu krzyżowym. Model T (czyli bez interakcji) okazał się nieistotny statystycznie zarówno dla całej grupy, jak i dla kobiet ( $p$  testu ilorazu wiarygodności wynosiło 0,67440 i 0,75485 odpowiednio). Modele te nie różniły się także istotnie od modeli zawierających jedynie wiek ( $p$  testu F 0,493614 i 0,930354 odpowiednio). W dodatku model T dla kobiet okazał się być dobrze dopasowany (jakość modelu oceniona testem Hosmera - Lemeshowa  $p=0,03275$ ). Konieczne było więc przebadanie modelu zawierającego interakcję wieku i TSH. Okazał się on istotnie lepszy od tego bez interakcji (porównanie modeli T i T1 za pomocą testu ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele:  $p=0,000126$  dla całej grupy i  $p=0,000434$  dla

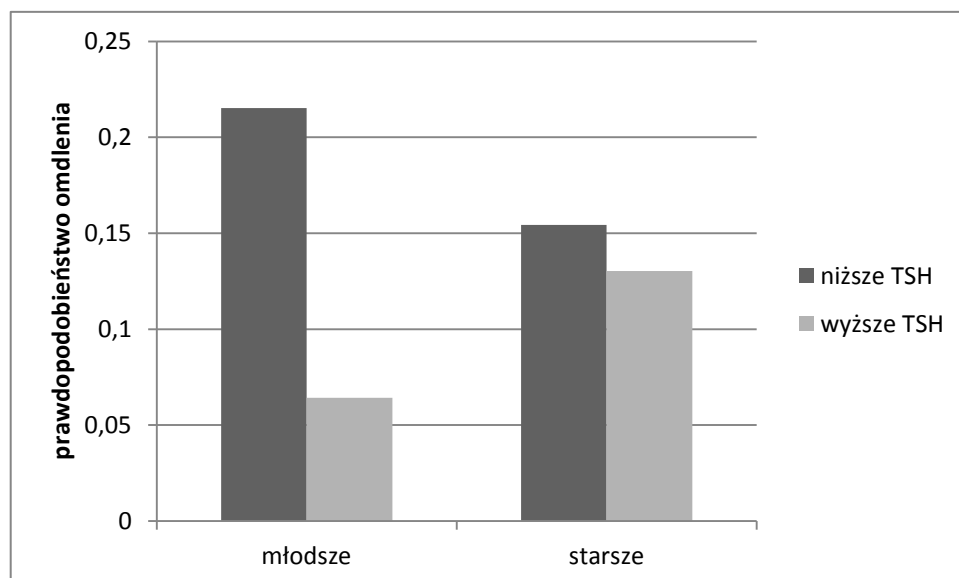
kobiet) oraz istotny statystycznie ( $p = 0,00157$  i  $0,00549$  odpowiednio). Model T1 nie był jednak dobrze dopasowany ( $p$  testu Hosmera - Lemeshowa wynosiło  $0,02398$ ), dlatego najwłaściwszym z modeli logistycznych oceniających związek TSH z występowaniem epizodów omdleń okazał się model T1 z interakcją wieku i TSH, dla kobiet. (Wszystkie wyniki modeli logistycznych znajdują się w Załączniku nr 1; Tab. XIX i Tab. XX).

Z uwagi na brak możliwości zinterpretowania ilorazów szans obliczonych dla modelu T1 z interakcją dla kobiet (w Załączniku nr 1; Tab. XIX), obliczono ilorazy szans dla dwóch punktów związanych z dwoma grupami wiekowymi dla: „starszych” (dolny kwartyl= 29 lat) i „młodszych” pacjentek (górnny kwartyl= 52 lata). Do obliczenia tych ilorazów szans wykorzystano współczynniki regresji modelu T1 dla kobiet przedstawione w Tabeli 21.

$$OR_{(29)} = e^{(-2,314)+0,03939*29} = 0,3098$$

$$OR_{(52)} = e^{(-2,314)+0,03939*52} = 0,7666$$

Iloraz szans dla młodszej grupy wskazuje, iż wraz ze wzrostem wartości TSH 3-krotnie spada szansa na wystąpienie epizodu omdlenia w analizowanej grupie kobiet. Z kolei szansa na wystąpienia epizodu omdlenia u starszych pacjentek również maleje wraz ze zwiększającymi się wartościami TSH, ale już nie tak znacząco. Rycina 22 obrazuje wpływ zmiennej wiek na determinowanie grupy omdleń poprzez wartości TSH.



Rycina 22. Prawdopodobieństwo wystąpienia omdlenia w podgrupach z wyższymi i niższymi wartościami TSH w różnych grupach wiekowych.



#### 4.3.2.2 Zastosowanie składowych głównych

Aby uwzględnić jednocześnie większą liczbę zmiennych niezależnych w analizowanych modelach zastosowano analizę składowych głównych jako metodę redukującą wymiary. Liczbę składowych głównych określano używając kryterium wykresu osypiska oraz kryterium minimalnego zasobu zmienności wspólnej. Wariancja każdego czynnika określa jaki procent całkowitej zmienności wszystkich pierwotnych zmiennych wyjaśnia dany czynnik.

W analizowanej bazie EP wyróżniono trzy grupy zmiennych określonych „tematycznie” ze względu na opisywanie konkretnych aspektów klinicznych. Pierwsza grupa zawierała 5 zmiennych: Wiek, BMI, Na<sup>+</sup>, K<sup>+</sup> oraz częstotliwość HR częstoskurczu standaryzowanego. Druga grupa, związana z hormonami tarczycy, miała 3 zmienne: TSH, FT3 i FT4. Natomiast w trzeciej grupie były zmienne opisujące wymiary jam serca: LK, PK, LP oraz Ao. Przeanalizowano również grupę 18 zmiennych, wybranych bez zamierzonej spójności tematycznej, sugerując się jedynie wstępną analizą struktury korelacyjnej. W grupach tych dokonano „kompresji” stosując analizę czynnikową. Liczbę składowych głównych w każdej grupie określono wg kryteriów wykresu osypiska oraz minimalnego zasobu zmienności wspólnej.

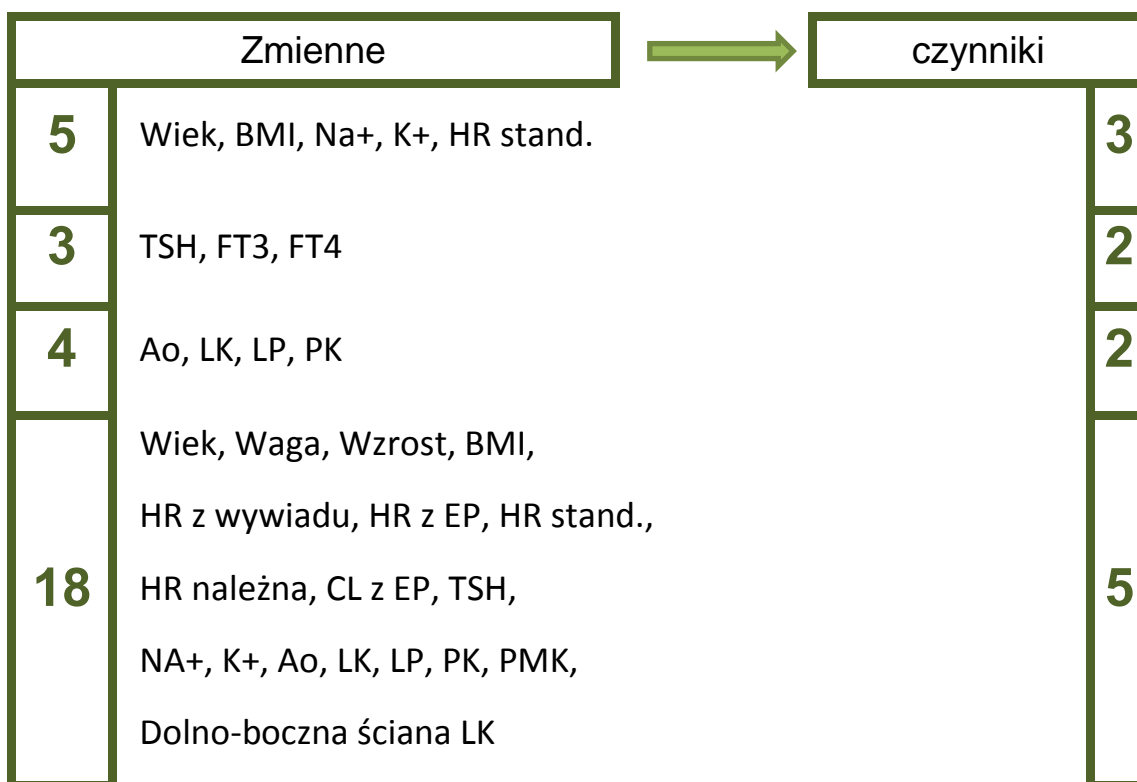
Tabela 22. Wartości własne czynników dla analizowanych grup zmiennych.

grupa	Numer czynnika	Wartość własna	% wariancji	% skumulowany
pierwsza	1	1,64	32,70	32,70
	2	1,05	21,06	53,76
	3	0,95	18,97	72,73
	4	0,85	16,94	89,68
	5	0,52	10,32	100,00
druga	1	1,53	50,94	50,94
	2	1,10	36,81	87,75
	3	0,37	12,25	100,00
trzecia	1	2,33	58,30	58,30
	2	0,64	15,90	74,20
	3	0,60	15,03	89,23
	4	0,43	10,77	100,00
18 zmiennych	1	5,99	33,29	33,29
	2	2,53	14,05	47,34
	3	2,01	11,14	58,48
	4	1,20	6,67	65,15
	5	1,08	6,02	71,17
	6	0,93	5,15	76,31
	7	0,87	4,85	81,17
	8	0,74	4,09	85,25
	9	0,65	3,62	88,87
	10	0,61	3,41	92,28

Przeanalizowane grupy zmiennych oraz wartości własne czynników wyznaczonych w wyodrębnionych podgrupach przedstawiono w Tabeli 22.

Wartość p statystyki Bartletta w każdym przypadku była poniżej przyjętego poziomu istotności, co wskazuje na prawdziwość hipotezy mówiącej o istotnej różnicy pomiędzy uzyskaną macierzą korelacji a macierzą jednostkową, czyli o znacznym skorelowaniu zmiennych (Załącznik nr 1; Tab. XV i Tab. XVI).

W wypadku pierwszej grupy wybrano trzy czynniki do dalszych analiz. Choć w tym wypadku czynnik trzeci ma wartość własną już poniżej 1, to tłumaczy on niemal 19% wariacji, dlatego włączono go również do analizy. Dla drugiej grupy wybrano dwa czynniki wyjaśniające łącznie ponad 87% wariacji. W grupie trzeciej czynnik drugi ma bardzo małą wartość własną (0,64), ale tłumaczy on niemal 16% wariacji, co w sumie z czynnikiem pierwszym daje łącznie 74% wytłumaczonej zmienności. Gdyby pozostawiono jeden czynnik, tłumaczyłby on niewiele ponad połowę zmienności zmiennych w tej grupie, więc byłby w zasadzie nieprzydatny. Następnie dla grupy 18 zmiennych wybrano 5 składowych głównych. W analizowanym przypadku 5 składowych głównych tłumaczy ponad 71% zmienności wśród zmiennych (zasoby zmienności wspólnej określonej przez daną liczbę czynników).



Rycina 23. Omawiane grupy zmiennych i liczba czynników wybranych za pomocą analizy składowych głównych.

Nowopowstałe czynniki są kombinacją liniową zmiennych w poszczególnych grupach i z założenia nie są ze sobą skorelowane. Współczynniki wartości czynnikowych, służące do wyznaczania wartości czynników oraz ładunki czynnikowe, przedstawiono w Załączniku nr 1; (Tab. XV i Tab. XVI). Dzięki znormalizowanej rotacji Varimax, układ ładunków obrazuje możliwie największe zróżnicowanie, pozwalając na łatwą interpretację wyznaczonych czynników. Są one wyróżnione przez wysokie ładunki dla jednych zmiennych i niskie ładunki dla innych.

Redukcja wymiaru we wszystkich omawianych czterech grupach zmiennych została przedstawiona na Rycinie 23.

#### 4.3.2.2.1 Grupa pierwsza

Dla pierwszej grupy wystandaryzowanych zmiennych wyznaczone czynniki wyglądają następująco:

$$\text{Czynnik1/3} = 0,740965 \cdot \text{BMI} + 0,366315 \cdot \text{wiek} + 0,159320 \cdot \text{Na}^+ + (-0,379084) \cdot \text{K}^+ + 0,072802 \cdot \text{częstoskurcz}$$

$$\text{Czynnik2/3} = (-0,022365) \cdot \text{BMI} + 0,054835 \cdot \text{wiek} + 0,790896 \cdot \text{Na}^+ + 0,508612 \cdot \text{K}^+ + 0,190420 \cdot \text{częstoskurcz}$$

$$\text{Czynnik3/3} = (-0,118350) \cdot \text{BMI} + 0,374518 \cdot \text{wiek} + (-0,261384) \cdot \text{Na}^+ + 0,387397 \cdot \text{K}^+ + (-0,679233) \cdot \text{częstoskurcz}$$

Poszczególne zmienne pierwotne w różny sposób oddziałują na pierwszy czynnik (wyznaczony za pomocą składowych głównych):

BMI jest dodatnio skorelowane z pierwszą składową, a zatem im wyższe BMI, tym wyższe wartości tej składowej. Współczynnik wartości dla BMI jest największy w tej składowej i wynosi 0,74. Jego ładunek czynnikowy informuje, że korelacja pomiędzy pierwszą składową główną a BMI jest bardzo wysoka i wynosi 0,88 co daje największy wkład w pierwszą składową. Wiek i Na<sup>+</sup> mają podobny, ale już mniejszy wpływ na pierwszą składową, a wpływ wartości częstotliwości HR częstoskurczu jest znikomy (ładunek <0,15). Wpływ wartości K<sup>+</sup> na pierwszą składową jest umiarkowany i ma ujemny znak korelacji.

Dla drugiego czynnika największe znaczenie ma wartość Na<sup>+</sup> i K<sup>+</sup>, za to wpływ BMI jest do pominięcia. Trzecia składowa oparta głównie na silnej, ujemnej korelacji HR częstoskurczu z rozpatrywanym czynnikiem.

#### 4.3.2.2.2 Grupa druga

Pierwszy czynnik związany jest z wystandaryzowanymi wartościami FT3 i FT4 (dodatnie współczynniki wartości czynnikowych), natomiast drugi odzwierciedla ujemną korelację z TSH (ujemny współczynnik wartości czynnikowej).

$$\text{Czynnik 1/2} = 0,0053 \cdot \text{TSH} + 0,5684 \cdot \text{FT4} + 0,5757 \cdot \text{FT3}$$

$$\text{Czynnik 2/2} = (-0,8814) \cdot \text{TSH} + 0,2533 \cdot \text{FT4} + (-0,2536) \cdot \text{FT3}$$

#### 4.3.2.2.3 Grupa trzecia

Tutaj pierwszy czynnik związany jest z wartościami LK i Ao, natomiast drugi z PK. Zmienne te wpływają poprzez dodatnie skorelowanie z analizowanymi czynnikami.

$$\text{Czynnik 1/2} = 0,4657 \cdot \text{LK} + 0,2067 \cdot \text{LP} + (-0,3716) \cdot \text{PK} + 0,7070 \cdot \text{Ao}$$

$$\text{Czynnik 2/2} = (-0,0453) \cdot \text{LK} + 0,3072 \cdot \text{LP} + 0,9353 \cdot \text{PK} + (-0,3639) \cdot \text{Ao}$$

#### 4.3.2.2.4 Grupa 18 zmiennych

W tej grupie wyznaczono pięć czynników zupełnie analogicznie, jak w poprzednich grupach zmiennych, tzn. wg współczynników wartości czynnikowych (Załącznik nr 1; Tab. XVI). W wypadku tak dużej grupy zmiennych, szczególnego znaczenia nabiera zastosowana rotacja, powodująca możliwie największe zróżnicowanie układu ładunków. Każdy z czynników odzwierciedlony jest w znaczącym stopniu przez inne zmienne:

- Czynnik 1/5: waga, BMI i LP
- Czynnik 2/5: wszystkie odnotowane wartości częstoskurczu (także CL)
- Czynnik 3/5: wiek i wyliczona wg niego należna wartość HR częstoskurczu
- Czynnik 4/5: TSH
- Czynnik 5/5: K+

#### 4.3.2.2.5 Modele regresji logistycznej wyznaczone dla zmiennych „skumulowanych” za pomocą analizy czynnikowej

W pierwszej omawianej grupie zmiennych wyznaczono trzy czynniki, które następnie użyto do utworzenia modelu regresji logistycznej. Niestety, test ilorazu wiarygodności wykazał, że omawiane czynniki nie mogą opisywać ryzyka wystąpienia epizodów omdleń ( $p = 0,19577$ ). Podobnie było w grupie drugiej i trzeciej – modele logistyczne były nieistotne statystycznie. W przypadku grupy drugiej, czyli zawierającej informacje o hormonach tarczycy, niestety było bardzo dużo braków danych i stąd utworzony model bazował jedynie na 30 przypadkach. W takim wypadku uzyskane wyniki mogą nie być wiarygodne. Szczegóły opisano w Załączniku nr 1 (Tab. XIX; modele typu R).

#### 4.3.3 Modele punktacji (scoringowe)

W analizowanym przypadku wybrano regresję logistyczną jako metodę analityczną dla scoringu. Pozwala ona na przedstawienie modelowanych zależności w postaci wzoru matematycznego, który może być następnie przekształcony do formatu karty punktacji (tzw. karty scoringowej). Korzystano również z metod eksploracji danych, takich jak losowy las, w celu odnalezienia najlepiej dyskryminujących reguł.

Przed wykonaniem analizy logistycznej zastosowano narzędzia służące do wstępnego wyboru predyktorów, czyli usunięcia nieistotnych cech oraz wyboru reprezentantów, co związane było z usunięciem nadmiernie skorelowanych zmiennych. Dokonana została także dyskretyzacja danych, po uprzednim wyliczeniu „wagi dowodu” (WoE ang. Weight of Evidence) dla predyktorów. (Analizy związane ze wstępnym wyborem predyktorów opisano w podrozdziale 3.4.1.1.)

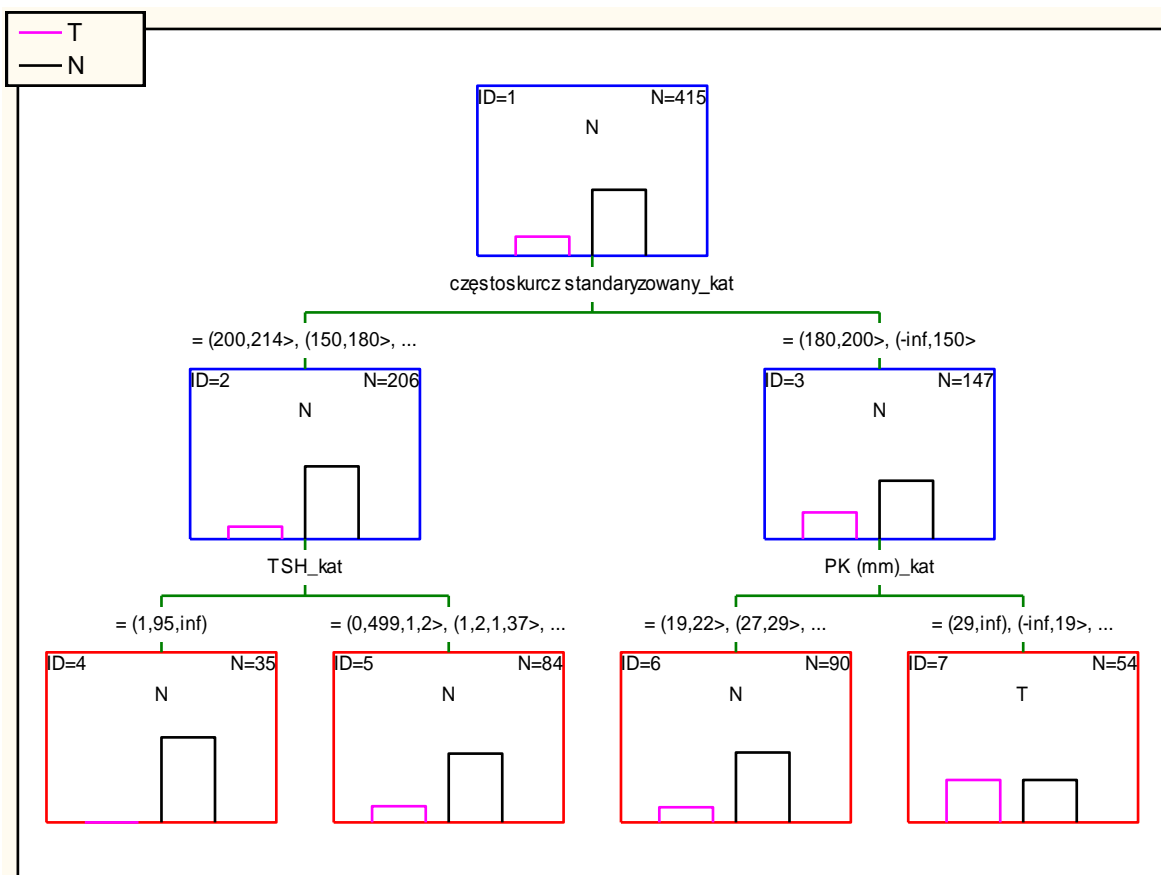
##### 4.3.3.1 *Drzewa losowe*

Poszukiwanie reguł, opisujących analizowane zdarzenie (epizody omdleń), przeprowadzono stosując klasyfikację przy użyciu drzew losowych. Wygenerowano 100 losowych drzew z maksymalną liczbą poziomów równą 2. (Omawiane reguły drzew losowych opisano w Załączniku nr 1; Tab. XXIII i Tab. XXV).

Najlepsze drzewo, dla danych po dyskretyzacji, miało wskaźnik zdarzeń niepożądanych bad rate = 0,76, więc nie nadawało się do prognozy występowania omdleń. Możemy natomiast odnaleźć reguły opisujące zmniejszone ryzyko wystąpienia epizodów omdleń.

Przykładowe drzewo klasyfikujące pacjentów do grupy omdleń, mające bad rate = 0, przedstawiono na Rycinie 24.

Do liścia nr 4 (mającego licznosc 35), który to nie zawiera żadnego przypadku omdleń, prowadzi ścieżka opisana regułą kontrolującą wartość HR częstoskurczu i TSH.



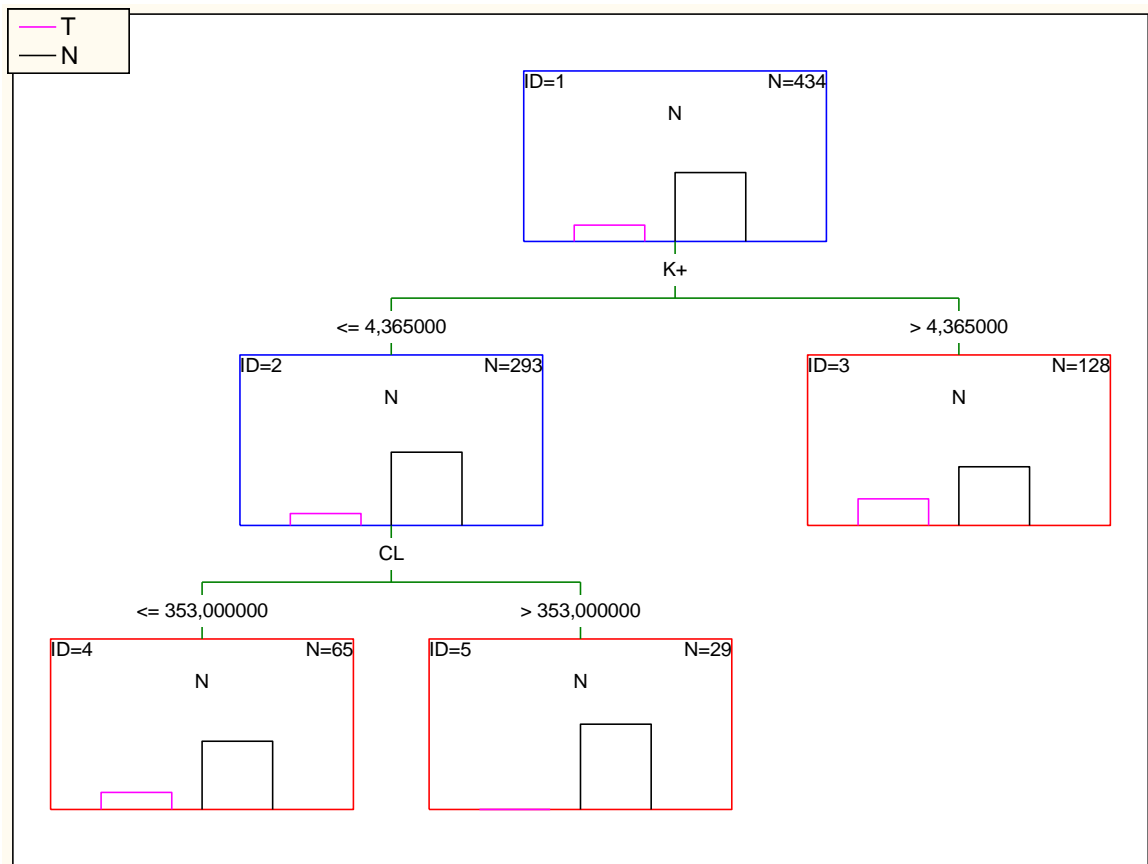
Rycina 24. Drzewo klasyfikacyjne utworzone dzięki analizie losowy las dla zmiennych skategoryzowanych. Bad rate = 0 biorące pod uwagę TSH oraz różnicę należnej i standaryzowanej wartości częstotliwości HR częstoskurczu.

Ponieważ dyskretyzowane zmienne miały wiele kategorii, z uwagi na mało liczne podgrupy jednorodnego ryzyka, o bezwzględnie niewielkich wartościach WoE (na zmianę dodatnich i ujemnych), wygenerowane dla tych zmiennych drzewa losowe są mało przydatne z medycznego punktu widzenia.

Na Rycinie 25 przedstawiono jedno z drzew wygenerowanych dla danych przed dyskretyzacją. Tutaj najwyższy wskaźnik zdarzeń niepożądanych („bad rate”) wynosił 0,7, dlatego również w tym przypadku ważniejsze okazały się reguły informujące o dużym prawdopodobieństwie niewystąpienia omdlenia. Jedna z reguł o wartości bad rate = 0 miała następującą postać:

`('K+'<=4,365) AND ('CL_z _badania_EP'>353)`

co prowadzi do liścia o licznosci 29, gdzie w ogóle nie było przypadków omdleń.



Rycina 25. Drzewo klasyfikacyjne utworzone dzięki analizie losowy las dla danych przed dyskretyzacją. Bad rate = 0 biorące pod uwagę wartość K+ oraz wartość CL z badania EP.

#### 4.3.3.2 Ocena powstałych modeli punktacji

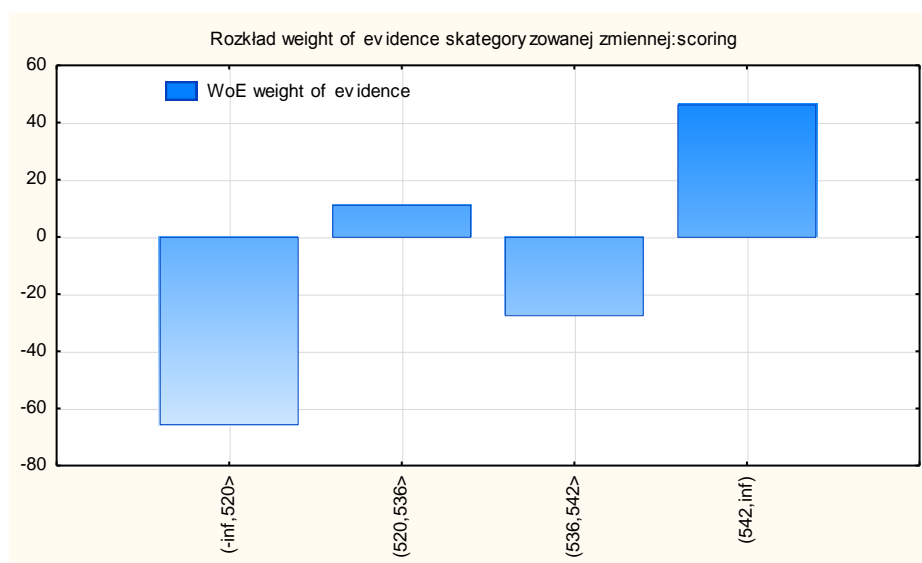
Ostateczny model regresji logistycznej zawierał 6 zmiennych po dyskretyzacji. Analizując liczności w poszczególnych grupach tych skategoryzowanych zmiennych możemy stwierdzić, że niektóre liczności są bardzo małe. Dlatego przed utworzeniem modelu, poszczególne podgrupy zmiennych niezależnych zostały przekodowane, czyli zamienione na odpowiadające im wartości WoE. Dzięki temu mogły być potraktowane przez model jak zmienne interwałowe.

Tabela 23. Współczynniki modelu logistycznego wykorzystywanego w modelu punktacji.

	wsp. b	p-testu Walda	OR (iloraz szans)	-95% CI	+95% CI
Wiek_kat	-0,0186	0,0315*	0,9816	0,9651	0,9984
Na+_kat	-0,0185	0,1068	0,9817	0,9599	1,0040
LP (mm)_kat	0,0146	0,0987	1,0147	0,9973	1,0324
PK (mm)_kat	-0,0164	0,0110*	0,9838	0,9714	0,9963
PMK (mm)_kat	-0,0108	0,1754	0,9892	0,9739	1,0048
częstoskurcz standaryzowany_kat	-0,0060	0,6025	0,9940	0,9719	1,0167

\* wynik istotny statystycznie ( $p < 0.05$ ); „\_kat” oznacza zmienną kategorialną, po dyskretyzacji, tu: po przekodowaniu na wartości WoE..

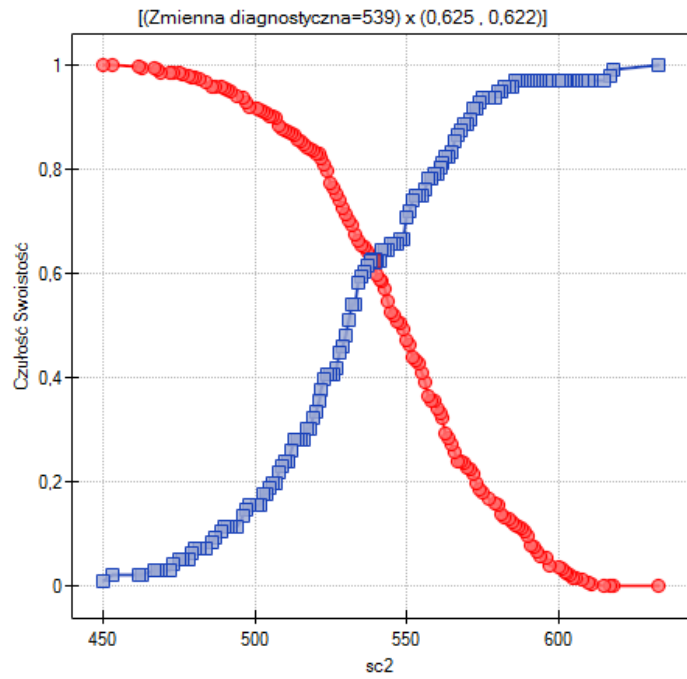
Na podstawie modelu regresji logistycznej utworzono tabelę scoringową. Raport cech zastosowany do wyliczenia punktacji zaprezentowano w Załączniku nr 1 (Tab. XXVI). Model S wyznaczony na podstawie jednej zmiennej niezależnej – obliczonej punktacji, był istotny statystycznie. U osoby z niższą punktacją (mniejszym scoringiem) istotnie częściej zdarzały się epizody omdleń (p testu Manna-Whitneya: 0,00007), choć w procesie dyskretyzacji tej nowopowstałej zmiennej nie było jednoznacznego podziału względem ryzyka występowania omdleń – Rycina 26. Można jednak zauważyć, iż ze wzrostem punktacji modelu rośnie szansa niewystąpienia omdlenia.



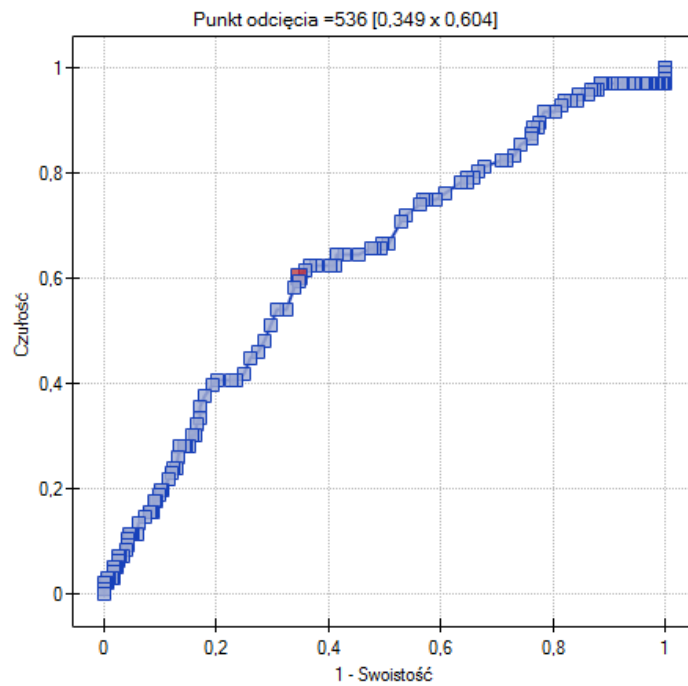
Rycina 26. Wartości WoE dla zmiennej scoring.

Wykonując analizę krzywej ROC, wyznaczono punkt odcięcia równy 539 (Rycina 27); w zaproponowanym punkcie odcięcia wartość czułości i swoistości jest podobna). Punkt ten związany był z czułością równą 78,13% i swoistością 40,27%. Pole pod krzywą ROC (Rycina 28) wyznaczonego modelu S (scoringowego) było istotnie różne od 0,5 ( $p=00007$ ) i wynosiło 0,63.





Rycina 27. Wykres czułości i swoistości dla punktacji wyliczonej za pomocą modelu scoringowego S.

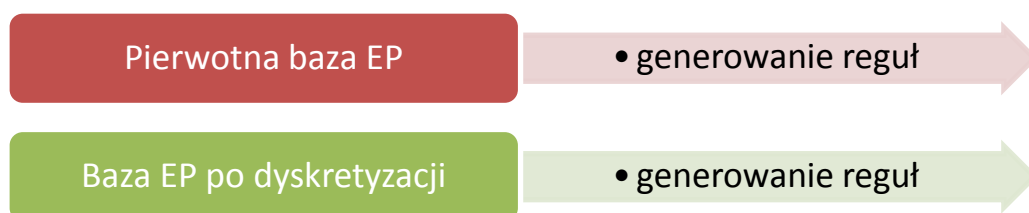


Rycina 28. Krzywa ROC dla punktacji wyliczonej za pomocą modelu punktacji (scoringowego) S.

#### 4.3.4 Zastosowanie zbiorów przybliżonych

Analizę danych, stosującą metodykę opartą na teorii zbiorów przybliżonych, przeprowadzono za pomocą programu ROSE2 (ang. Rough Sets Data Explorer) wersja 2.2. W analizowanym systemie decyzyjnym wzięto pod uwagę dwie klasy decyzyjne: N i T, mówiące o występowaniu, bądź nie, epizodów omdleń. Do klasy N należy 370 obiektów, do T 96.

Przeanalizowano dwa zbiory danych: pierwotny i po dyskretyzacji (Rycina 29). Następnie porównano uzyskane rezultaty.



Rycina 29. Zastosowane w pracy kroki analizy danych metodą zbiorów przybliżonych.

Cała lista wygenerowanych reguł i reduktów znajduje się na stronie Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu (<http://www.ump.edu.pl/index.php?department=70>).

##### 4.3.4.1 Generowanie reguł decyzyjnych

Program ROSE2 generuje reguły decyzyjne używając algorytmu LEM2 [79] [63]. Metoda ta nie działa w przypadku występujących w bazie brakach danych. Dlatego przed wygenerowaniem reguł, zastosowano dostępny w programie ROSE2, jeden z najprostszych sposobów radzenia sobie z brakami danych: zastąpiono brakujące wartości najczęściej występującymi wartościami dla danego atrybutu.

Zarówno dolne, jak i górne przybliżenie obu klas (T i N), zawiera wszystkie obiekty danej klasy. (Tabela 24 przedstawia najważniejsze wyniki przeprowadzanych analiz.) Stosunek wielkości górnego i dolnego przybliżenia, określający dokładność przybliżenia, wynosi zatem 1 dla obu klas. Można by więc przypuszczać, iż w świetle wszystkich predyktorów należących do systemu informacyjnego zakwalifikowanie pacjenta jako zagrożonego wystąpieniem omdlenia, jest bezbłędne. Całkowita jakość klasyfikacji wynosi również 1, co oznacza, iż 100% obiektów w systemie informacyjnym została rozpoznana prawidłowo.

Tabela 24. Podsumowanie wyników analiz przeprowadzonych przy użyciu metodyki opartej na teorii zbiorów przybliżonych w programie ROSE2 (po automatycznym uzupełnieniu braków danych)

	podstawowy zbiór danych	zbiór danych po dyskretyzacji
Klasa N	370	370
Klasa T	96	96
Przybliżenie górne N	370	370
Przybliżenie dolne N	370	370
Przybliżenie górne T	96	96
Przybliżenie dolne T	96	96
Jakość klasyfikacji	1	1
Liczba atomów	465	466
Liczba atrybutów w rdzeniu	0	0
Liczba reduktów	352	>750*
Wygenerowane reguły	147	108

\*przerwano wykonywanie analizy ze względu na duże zapotrzebowanie mocy obliczeniowej

W pierwotnej bazie (bez dyskretyzacji) odnaleziono jeden atom mniej, w porównaniu z bazą po dyskretyzacji, gdzie było 466 atomów. Wygenerowane zostało tam 352 reduktów, z których 3 największe zawierały po 9 atrybutów, a najmniejsze z nich miały 3 atrybuty (szczegółowe listy reguł i reduktów zamieszczono na stronie Uniwersytetu Medycznego). Zatem, najmniejszym zestawem atrybutów warunkowych prowadzących do klasyfikacji o takiej samej jakości, jaką daje klasyfikacja z uwzględnieniem wszystkich dostępnych w systemie informacyjnym predyktorów, jest np.: BMI, wiek pierwszego kołatania oraz częstotliwość HR częstoskurczu. Rdzeń systemu informacyjnego nie zawierał żadnych atrybutów. Oznacza to, iż nie ma takich cech w systemie, których odjęcie spowoduje osłabienie jakości klasyfikacji.

Wszystkie wygenerowane reguły były deterministyczne, tzn. dawały jednoznaczną odpowiedź. Tabela 25 przedstawia dwie najsilniejsze reguły decyzyjne wygenerowane dla bazy podstawowej dające odpowiedź negatywną, co do występowania omdlenia:

Tabela 25. Przykłady reguł decyzyjnych dla klasy omdleń N, program ROSE2

<p><b>reguła 1.</b> (LK = 48) &amp; (Ao = 24) =&gt; (SYNCOPE = N); [5, 5, 1.35%, 100.00%][5, 0][{130, 200, 235, 280, 313}, {}]</p> <p><b>reguła 2.</b> (ARYTMIA = AVNRT) &amp; (LK = 40) &amp; (grupy_przekroczona_nieprzekroczo=przekroczona) =&gt; (SYNCOPE = N); [12, 12, 3.24%, 100.00%][12, 0] [{4, 5, 59, 68, 70, 77, 137, 139, 140, 158, 166, 214}, {}]</p>
--

Pierwszą z reguł zaprezentowanych w Tabeli 25 tłumaczy się w sposób następujący:

**Jeżeli** (LK = 48) i (Ao = 24) **to** decyzja: **brak omdlenia**. Regułą tą wspiera 5 obiektów o numerach identyfikacyjnych {130, 200, 235, 280, 313}, co stanowi raptem 1.35% wszystkich osób, u których nie odnotowano omdleń (klasa N). Nie znalazła się ani jedna osoba, u której wystąpiło omdlenie, a która byłaby opisana takimi samymi wartościami atrybutów warunkowych. (Dokładność reguły to 1, natomiast pewność 0,135)

W Tabeli 26 przedstawiono z kolei kilka z najsilniejszych reguł mówiących o występowaniu omdlenia:

Tabela 26. Przykłady reguł decyzyjnych dla klasy omdleń T, program ROSE2

<b>reguła 89.</b> (BMI = 27.1203) => (SYNCOPE = T); [2, 2, 2.08%, 100.00%][0, 2][{} , {436, 441}]
<b>reguła 90.</b> (BMI = 25.3906) => (SYNCOPE = T); [2, 2, 2.08%, 100.00%][0, 2][{} , {396, 417}]
<b>reguła 91.</b> (LK = 43) & (Choroby_wsp_II = 1) & (naprzemienność_zespołu_QRS = nie) & (kategorie_BMI = otylosc) => (SYNCOPE = T); [3, 3, 3.13%, 100.00%][0, 3] [{} , {373, 378, 379}]
<b>reguła 99.</b> (TSH = 0.88) & (ty_scLK = 8) & (Choroby_wspo_I = 1) & (grupy_przekroczone_nieprzekroczo= przekroczone) => (SYNCOPE = T); [5, 5, 5.21%, 100.00%][0, 5][{} , {371, 402, 438, 439, 442}]
<b>reguła 106.</b> (Plec = M) & (Pora_dnia = dzien_noc) & (ty_scLK = 10) & (kategorie_BMI = nadwaga) => (SYNCOPE = T); [4, 4, 4.17%, 100.00%][0, 4][{} , {405, 412, 414, 418}]
<b>reguła 113.</b> (ARYTMIA = WPW) & (ty_scLK = 11) & (Choroby_wspo_I = 1) & (Choroby_wsp_II = 0) & (kategorie_BMI = nadwaga) => (SYNCOPE = T); [3, 3, 3.13%, 100.00%][0, 3][{} , {450, 458, 464}]

Są to reguły pokryte przez maksymalnie 5 obiektów z klasy opisującej przypadki omdleń.

Podjęto również próbę wygenerowania reguł, bez ingerowania w braki danych, naturalnie występujące w danych medycznych, za pomocą innej implementacji zmodyfikowanego algorytmu LEM2 (umożliwił to program RSES2 - ang. Rough Set Exploration System). Uzyskano tam 146 reguł dla danych dyskretyzowanych. Reguły zawierały od 8 do 12 cech, a największe pokrycie wynosiło 15 – dla klasy N. Klasa T maksymalnie była zdefiniowana przez 5 obiektów.

Najważniejsze reguły uzyskane dla danych po dyskretyzacji zaprezentowano w Tabeli 27:

Tabela 27. Przykłady reguł decyzyjnych dla danych po dyskretyzacji, dla klasy omdleń N, program RSES2

<p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(Wiek_kat=(-inf_41&gt;)&amp;(kat_BMI=norma_lub_niedowaga)&amp;(Choroby_wspo_I=0)&amp;(LP_kat=(-inf_31&gt;)&amp;(kg_waga_kat=(-inf_63&gt;)&amp;(PMK_kat=(-inf_8&gt;)&amp;(Ao_kat=(-inf_25&gt;)=&gt;(Syncope=N[15]) 15</p>
<p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(kat_BMI=nadwaga_lub_otylosc)&amp;(Choroby_wspo_I=1)&amp;(Wiek_kat=(-inf_41&gt;)&amp;(cm_wzrost_kat=(163_170&gt;)=&gt;(Syncope=N[13]) 13</p>
<p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(Wiek_kat=(-inf_41&gt;)&amp;(kat_BMI=norma_lub_niedowaga)&amp;(Choroby_wspo_I=0)&amp;(LP_kat=(-inf_31&gt;)&amp;(Ao_kat=(-inf_25&gt;)&amp;(czestosc_kat=(214_inf))=&gt;(Syncope=N[11]) 11</p>
<p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(Choroby_wspo_I=1)&amp;(kat_BMI=nadwaga_lub_otylosc)&amp;(PMK_kat=(9_5_11&gt;)&amp;(po_ERP_Avnr_wartosci_kat=(210_240&gt;)=&gt;(Syncope=N[11]) 11</p>
<p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(Choroby_wspo_I=1)&amp;(kat_BMI=nadwaga_lub_otylosc)&amp;(kg_waga_kat=(80_96&gt;)&amp;(PMK_kat=(9_5_11&gt;)=&gt;(Syncope=N[10]) 10</p>
<p>(Choroby_wsp_II=0)&amp;(Choroby_wspo_I=0)&amp;(kat_BMI=norma_lub_niedowaga)&amp;(Ao_kat=(27_32&gt;)&amp;(po_ERP_Avna_wartosci_kat=(240_300&gt;)=&gt;(Syncope=N[10]) 10</p>

„\_kat” oznacza zmienną kategoryjną, po dyskretyzacji.

Pierwsza z zaprezentowanych powyżej reguł ma pokrycie 15 obiektów i mogłaby służyć do wstępnego wykluczania przypadków omdleń, gdyż definiuje klasę decyzyjną N. Reguła mówi o zakwalifikowaniu jako osób niezagrażonych omdleniem tych pacjentów, którzy mieli co najwyżej 41 lat, nie byli obciążeni żadną dodatkową, jawną chorobą, kołatania u nich pojawiały się w ciągu dnia (choć nie tylko), ich BMI było w normie lub była to niedowaga (kategoria wagi: <63 kg) i jednocześnie wartość PMK była u nich co najwyżej 8 mm, natomiast Ao poniżej 25 mm.

Przykładowe reguły definiujące klasę T przedstawiono w Tabeli 28. Z pierwszej z reguł zaprezentowanych w Tabeli 28 wynika, iż osoby cięższe i grubsze, obciążone dodatkowymi chorobami i o większym wymiarze Ao, u których również kołatania pojawiają się w ciągu dnia, kwalifikowane są jako zagrożone omdleniem. W analizowanej bazie było 5 takich przypadków.

Tabela 28. Przykłady reguł decyzyjnych dla danych po dyskretyzacji, dla klasy omdleń T, program RSES2

<p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wspo_I=1)&amp;(kat_BMI=nadwaga_lub_otylosc) &amp;(Choroby_wsp_II=1)&amp;(Ao_kat=(25_27&gt;)&amp;(kg_waga_kat=(96_inf))=&gt;(Syncope=T[5]) 5</p> <p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wspo_I=1)&amp;(kat_BMI=nadwaga_lub_otylosc) &amp;(Choroby_wsp_II=1)&amp;(LP_kat=(35_39&gt;)&amp;(po_ERP_Avnr_wartosci_kat=(210_240&gt;))=&gt;(Syncope=T[5]) 5</p> <p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(Wiek_kat=(-inf_41&gt;)&amp;(kat_BMI=norma_lub_niedowaga)&amp;(Choroby_wspo_I=0)&amp;(EP_wywiad_kat=(180_211&gt;)&amp;(roznica_nalezna_HR_stand_kat=(-17_3_10_95&gt;)&amp;(czestosc_kurcz_standaryzowany_kat=(180_200&gt;)&amp;(kg_waga_kat=(71_80&gt;))=&gt;(Syncope=N[4]) 4</p> <p>(Pora_dnia_kat=dzien/noc_dzien)&amp;(Choroby_wsp_II=0)&amp;(Wiek_kat=(-inf_41&gt;)&amp;(kat_BMI=norma_lub_niedowaga)&amp;(Choroby_wspo_I=0)&amp;(LP_kat=(-inf_31&gt;)&amp;(PMK_kat=(-inf_8&gt;)&amp;(cm_wzrost_kat=(163_170&gt;)&amp;(kg_waga_kat=(63_67&gt;))=&gt;(Syncope=N[4]) 4</p>
--

„\_kat” oznacza zmienną kategorialną, po dyskretyzacji.

#### 4.3.4.2 Problem niezrównoważenia danych

W pracy zastosowano algorytm realizujący zmianę poziomu niezrównoważenia danych poprzez zastosowanie tzw. nadpróbkiwania (ang. oversampling). Był to algorytm, dostępny w programie WEKA, z rodziny SMOTE (czyli dogenerowanie nowych przykładów klasy mniejszościowej pomiędzy przykładami pierwotnymi).

Po przekształceniu zdyskretyzowanego uprzednio zbioru danych algorytmem SMOTE, zbiór danych miał 562 obiekty, z czego 34% (192 osoby) stanowiły przypadki omdleń. Na Rycinie 30 przedstawiono zastosowane kroki postępowania z danymi niezrównoważonymi z wykorzystaniem nadlosowania.



Rycina 30. Kroki postępowania z danymi niezrównoważonymi – nadlosowanie.

Analogicznie jak w przypadku danych bez powiększania poprzez losowanie, także i w tym przypadku zdecydowano się na zastosowanie algorytmu generującego reguły decyzyjne, który jest zaimplementowany w programie RSES2. Dla omawianych danych: tzn. po dyskretyzacji i nadlosowaniu algorytmem SMOTE, uzyskano 160 reguł. Najważniejsze z nich zaprezentowano w Tabeli 29:

Tabela 29. Przykłady reguł decyzyjnych dla danych po dyskretyzacji i SMOTE, program RSES2

(Pora_dnia_kat=dzien/noc_dzien)&(Choroby_wsp_II=0)&(Wiek_kat=(-inf_41>) & (kat_BMI=norma_lub_niedowaga)&(pre_ERP_Avna_wartosci_kat=(-inf_220>) &(grupa_CL=ponizej)&(pre_ERP_Avnr_wartosci_kat=(200_220>)=>(Syncope=T[42]) 42
(Pora_dnia_kat=dzien/noc_dzien)&(Choroby_wsp_II=0)&(Choroby_wspo_I=1)& (pre_ERP_Avna_wartosci_kat=(-inf_220>)&(grupa_CL=ponizej)& (Cl_Wen_rete_wartosci_kat=(300_317>)=>(Syncope=T[40]) 40
(Pora_dnia_kat=dzien/noc_dzien)&(Choroby_wsp_II=0)&(Wiek_kat=(-inf_41>)& (kat_BMI=norma_lub_niedowaga)&(Choroby_wspo_I=0)&(LP_kat=(-inf_31>)& (kg_waga_kat=(-inf_63>)&(LK_kat=(-inf_39>)=>(Syncope=N[11]) 11
(Pora_dnia_kat=dzien/noc_dzien)&(Choroby_wspo_I=1)&(kat_BMI=nadwaga_lub_otylosc)& (Choroby_wsp_II=1)&(Ao_kat=(25_27>)&(kg_waga_kat=(96_inf))=>(Syncope=T[7]) 7
(Pora_dnia_kat=dzien/noc_dzien)&(Choroby_wsp_II=0)&(Choroby_wspo_I=1)&(kat_BMI=norma_lub_niedowaga)&(LP_kat=(-inf_31>)&(kg_waga_kat=(-inf_63>)& (pre_ERP_Avna_wartosci_kat=(-inf_220>)&(EP_wywiad_kat=(211_inf))=>(Syncope=T[8]) 8
(Pora_dnia_kat=dzien/noc_dzien)&(Choroby_wsp_II=0)&(Choroby_wspo_I=0)&(kat_BMI=nadwaga_lub_otylosc)&(Na+_kat=(140_142>)&(pre_ERP_Avnr_wartosci_kat=(200_220>)&(grupa_CL=ponizej)=>(Syncope=T[10]) 10

\* „grupa\_CL=ponizej” oznacza CL częstoskurczu z badania EP poniżej 280,5  
„\_kat” oznacza zmienną kategoryjną, po dyskretyzacji.

W Tabeli 29 można zaobserwować dużo większe pokrycie obiektów niż w poprzednich zestawach wygenerowanych reguł (dla niepowiększanego zbioru danych). Wyniki generowania reguł na nowym, powiększonym zbiorze danych zaprezentowano w Tabeli 30.

Tabela 30. Podsumowanie wyników analiz przeprowadzonych przy użyciu metodyki opartej na teorii zbiorów przybliżonych dla bazy powiększonej algorytmem SMOTE.

	Powiększony zbiór danych (algorytm SMOTE)
<b>Klasa N</b>	370
<b>Klasa T</b>	192
<b>Przybliżenie górne N</b>	370
<b>Przybliżenie dolne N</b>	370
<b>Przybliżenie górne T</b>	192
<b>Przybliżenie dolne T</b>	192
<b>Jakość klasyfikacji</b>	1
<b>Liczba atomów</b>	560
<b>Liczba atrybutów w rdzeniu</b>	0
<b>Liczba reduktów</b>	>669*
<b>Wygenerowane reguły</b>	114

\*przerwano wykonywanie analizy ze względu na duże zapotrzebowanie mocy obliczeniowej

#### 4.3.4.3 Klasyfikacja systemu informacyjnego

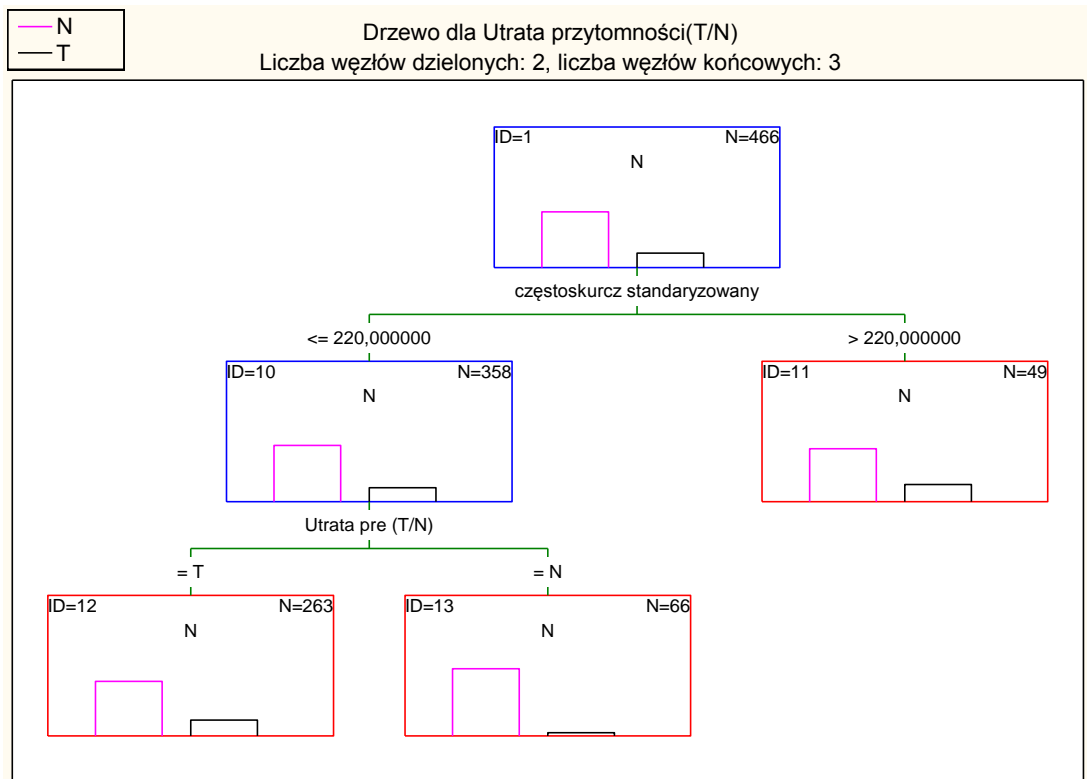
Aby ocenić jakość analizowanych systemów decyzyjnych przeprowadzono tzw. sprawdzian krzyżowy. Jego wyniki przedstawiono w Tabeli 31.

Tabela 31. Wyniki oceny skuteczności klasyfikatorów dla reguł decyzyjnych.

	Zbiór nierozszerzony				Zbiór rozszerzony (SMOTE), po dyskretyzacji	
	Zbiór podstawowy		Zbiór po dyskretyzacji		bez ingerencji w braki danych	po uzupełnieniu braków danych
	bez ingerencji w braki danych	po uzupełnieniu braków danych	bez ingerencji w braki danych	po uzupełnieniu braków danych		
<b>Całkowita średnia dokładność</b>	54.66 +- 8.50	51.54 +- 7.27	63.92 +- 4.58	62.67 +- 7.50	71.51 +- 3.59	53.91 +- 5.89
<b>Średnia dokładność dla klasy decyzyjnej T</b>	40.90 +- 18.97	41.37 +- 11.21	50.34 +- 20.33	36.52 +- 17.36	62.99 +- 5.44	84.19 +- 9.83

#### 4.3.5 Drzewa interakcyjne typu C&RT/CHAID

Aby uwzględnić zmienne ważne dla klinicysty, stworzono szereg drzew, dla których dopasowywano różne punkty podziałów na kategorie. Spośród utworzonych drzew, wybrano te, które najlepiej determinują brak omdleń. Na Rycinie 31 przedstawiono przykładowe drzewo mogące służyć do opisanie przypadków zmniejszonego ryzyka omdleń:



Rycina 31. Przykładowe drzewo interaktywne typu C&RT.



Jeśli pacjent miał wartość częstotliwości HR częstoskurczu nie wyższy niż 220 i jednocześnie nie miał stanów poprzedzających epizody omdleń (było 66 takich przypadków w analizowanej próbie badawczej), to jest bardzo prawdopodobne, że można wykluczyć u niego wystąpienie takiego epizodu. Sugerowana ścieżka postępowania, opisująca pacjentów o mniejszym ryzyku omdlenia, została następnie poddana dalszej analizie. Przeliczając testem Chi<sup>2</sup> dane z Tabela 32 uzyskujemy p= 0,000705, a więc istotną statystycznie różnicę w proporcji występowania epizodów omdleń pomiędzy osobami narażonymi i nienarażonymi na omawiany czynnik ryzyka.

Tabela 32. Porównanie częstości występowania omdleń w grupie narażonej i nienarażonej na omawiane ryzyko\*.

	omdlenia		
	T	N	Razem
Narażenie na czynnik*	3	63	66
Brak narażenia	71	241	312
Ogół	74	304	378

\*Analizowany czynnik ryzyka, to nie spełnianie warunku: HR częstoskurczu <= 220 i jednocześnie brak stanów poprzedzających epizody omdleń

Iloraz szans dla powyższego czynnika ryzyka wynosi 6,187. Mamy tu bardzo szeroki przedział ufności dla obliczonego OR: (1,886;20,297), jednak jest on istotnie różny od 1 (test istotności ilorazu szans: p= 0,002643). Oznacza to, iż pacjenci mający HR częstoskurczu nie wyższy 220, u których nie występowały stany przed omdleniowe, mają prawie sześciokrotnie mniejszą szansę na zaistnienie omdlenia w porównaniu z grupą HR>220 i stanami poprzedzającymi omdlenia oraz HR<=220 bez stanów poprzedzających omdlenia.

#### 4.3.6 Zastosowanie testów dokładnych (exact) i Monte Carlo

Aby umożliwić uzyskiwanie wiarygodnych wyników na podstawie małych prób lub nielicznych obserwacji, część analiz (dla cech ważnych klinicznie, z asymptotycznym wynikiem istotnym lub na krawędzi istotności) przeprowadzono dodatkowo z użyciem testów dokładnych, wyliczając dokładne prawdopodobieństwo oraz metod symulacyjnych – Monte Carlo. Były to testy porównujące dwie grupy dla wybranych zmiennych oraz najważniejsze modele logistyczne.

W Tabeli 34 przedstawiono zestawienie wyników uzyskanych metodami asymptotycznymi, w porównaniu z obliczeniami zwiększonej precyzji oraz wynikami symulacji Monte Carlo (10 000 próbek) dla nieparametrycznego testu Manna-Whitneya. Wybrane wyniki testów Chi<sup>2</sup> (asymptotyczne i dokładne) analizujących zmienne nominalne przedstawiono w Tabeli 33.

Tabela 33. Porównanie wyników testów asymptotycznych i dokładnych porównujących cechy jakościowe.

zmienna	test Chi <sup>2</sup>	
	metoda asymptotyczna	metoda (exact)
arytmia	0,2169	0,2434
dla K	0,0841	0,1016
dla M	0,8990	1,0000
wiek częstoskurczu	0,5414	0,6197
dla M	0,0650	0,0789
Współwystępowanie chorób typu I	0,1954	0,2044
dla K	0,0996	0,1036

Tabela 34. Porównanie grupy z i bez epizodów omdleń (test Manna-Whitneya) za pomocą testów asymptotycznych, dokładnych oraz symulacji Monte Carlo.

zmienna	p-testu Manna-Whitneya		
	metoda asymptotyczna	metoda exact	metoda MCO (99%CI)
częstoskurcz standaryzowany	0,8735	0,8739	0,8755(0,8670-0,8840)
dla K	0,4719	0,4733	0,4794(0,4665-0,4923)
dla M	0,5203	0,5227	0,5222(0,5093-0,5351)
BMI	0,1453	0,1456	0,1493(0,1401-0,1585)
dla K	0,5432	0,5445	0,5489(0,5361-0,5617)
dla M	0,0381*	0,0378*	0,0385(0,0335-0,0435)*
TSH	0,2057	0,2065	0,2077(0,1973-0,2181)
dla K	0,0204*	0,0200*	0,0206(0,0169-0,0243)*
dla M	0,1914	0,1940	0,1935(0,1833-0,2037)
FT3	0,4635	0,4727	0,4766(0,4637-0,4895)
dla K	0,6743	0,6831	0,6879(0,6760-0,6998)
dla M	0,1138	0,1818	0,1777(0,1679-0,1875)
Na+	0,7830	0,7836	0,7864(0,7758-0,7970)
dla K	0,3588	0,3601	0,3598(0,3474-0,3722)
dla M	0,4851	0,4873	0,4830(0,4701-0,4959)
K+	0,0261*	0,0260*	0,0267(0,0226-0,0309)*
dla K	0,1282	0,1286	0,1307(0,1220-0,1394)
dla M	0,0992	0,0995	0,1020(0,0942-0,1098)
LP	0,0618	0,0617	0,0651(0,0588-0,0715)
dla K	0,0265*	0,0263*	0,0292(0,0249-0,0335)*
dla M	0,4997	0,5019	0,5037(0,4908-0,5166)
PK	0,0382*	0,0380*	0,0387(0,0337-0,0437)*
dla K	0,0425*	0,0423*	0,0408(0,0357-0,0459)*
dla M	0,2285	0,2300	0,2280(0,2172-0,2388)
PMK	0,1339	0,1341	0,1323(0,1236-0,1410)
dla K	0,0106*	0,0103*	0,0106(0,0080-0,0132)*
dla M	0,7397	0,7415	0,7476(0,7364-0,7588)

\* wynik istotny statystycznie (p<0,05)

Przeliczenie modeli dokładnych dla regresji logistycznej wymagało bardzo dużej mocy obliczeniowej, dlatego w niniejszej pracy nie podano wartości exact dla modeli regresji logistycznej. W Tabeli 35 przedstawiono niektóre modele regresji logistycznej przeliczone z użyciem symulacji MCO. (Było to możliwe jedynie dla modeli zawierających maksymalnie dwie zmienne niezależne.)

Tabela 35. Porównanie wyników wybranych modeli logistycznych dla omdleń, wyznaczonych za pomocą metod asymptotycznych i symulacji Monte Carlo.

model	zmienne w modelu	metoda asymptotyczna		metoda MCO		
		OR	wartość p dla zmiennej (test Walda)	OR	wartość p dla zmiennej (test Walda)	SE
A	częstoskurcz standaryzowany wiek	1,002	0,5715	1,002	0,5820	0,0091
		1,017	0,0438*	1,017	0,0446*	0,0030
A dla krótszego# czasokresu częstoskurczu	częstoskurcz standaryzowany	1,005	0,3801	1,004	0,3908	0,0079
	wiek	1,032	0,0022*	1,032	0,0012*	0,0005
P	LP	1,036	0,1855	1,035	0,1856	0,0058
	PK	1,044	0,2191	1,043	0,2246	0,0063
F	częstoskurcz standaryzowany	0,998	0,6343	0,998	0,6284	0,0093
	pre syncope	6,039	0,0016*	6,930	0,0000*	0,0006

# analizowano tu czasokres od pierwszego udokumentowanego częstoskurczu

\* wynik istotny statystycznie ( $p < 0,05$ )

#### 4.4 Porównanie modeli

W pracy zastosowano V-krotną walidację rozpatrywanych modeli, aby właściwie ocenić ich trafność prognostyczną (szczegóły definicji wyliczonych modeli przedstawiono w Załączniku nr 1; Tab. XIX i Tab. XIX). Porównanie najważniejszych miar oceniających utworzone modele zaprezentowano na Rycinach: 33 - 36. W Tabeli 36 zestawiono definicje porównywanych modeli logistycznych:

Tabela 36. Definicje porównywanych, wybranych modeli logistycznych.

model	zmiennie niezależne					
A	HR stand.	wiek				
A1	HR stand.	wiek	krótki czasokres częstoskurczu (<20lat)			
E	HR stand.	Choroby typu II				
F	HR stand.	stany przedomdleniowe				
F1	HR stand.	stany przedomdleniowe			Choroby typu I	
F2	HR stand.	stany przedomdleniowe			płeć	
P	LP	PK				
P1	LP	PK	LP x PK			
R	składowe główne (4 modele dla czterech grup zmiennych)					
W	Wiek_kat	Na+_kat	LP_kat	PK_kat	PMK_kat	HR stand._kat
S	scoring					
T	Wiek	TSH				
T1	Wiek	TSH	TSH x wiek			

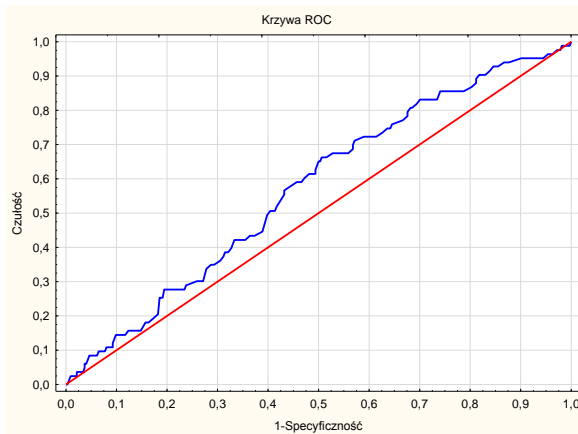
Na przykładzie modelu logistycznego A (analizującego wiek i częstoskurcz standaryzowany) pokazano, iż bez zastosowania krosvalidacji nie można być pewnym, czy model będzie dobrze działał dla danych, które nie były wykorzystywane do jego konstruowania. Porównanie miar, oceniających przydatność modelu logistycznego A do klasyfikacji omdleń (Tabela 37), wykazuje zwiększenie wykrywalności osób zagrożonych omdleniem w przypadku modelu po krosvalidacji (czułość wzrosła do 21,43%). Wartości miar F i G również są bardziej optymistyczne po przeprowadzeniu sprawdzianu krzyżowego. Niestety, precyzja znacznie spadła.

Tabela 37. Porównanie oceny modelu logistycznego (A) jako prognostyka omdleń za pomocą wieku i HR częstoskurczu dla modelu z i bez krosvalidacji.

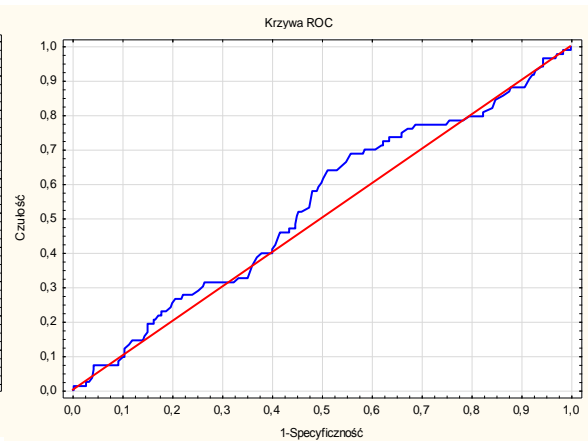
Model logistyczny	Czułość	Swoistość	Precyzja	F-miara			G-miara
				F1	F2	F3	
A (pkt. odcięcia 0,31)	2,41%	98,77%	33,33%	4,49%	2,96%	2,66%	15,43%
A po V-krotnej (pkt. odcięcia 0,27)	21,43%	79,68%	7,23%	10,81%	15,38%	17,91%	41,32%

Również ocena modelu za pomocą krzywej ROC okazała się mniej korzystna dla modelu A po krosvalidacji – pole pod krzywą przestało istotnie różnić się od wartości 0,5 oznaczającej losowy przydział przypadków do klas z omdleniem i bez omdlenia. (Rycina 32)

Krzywa ROC dla modelu A  
tylko próba ucząca  
(AUC=0,57, p=0,0396)

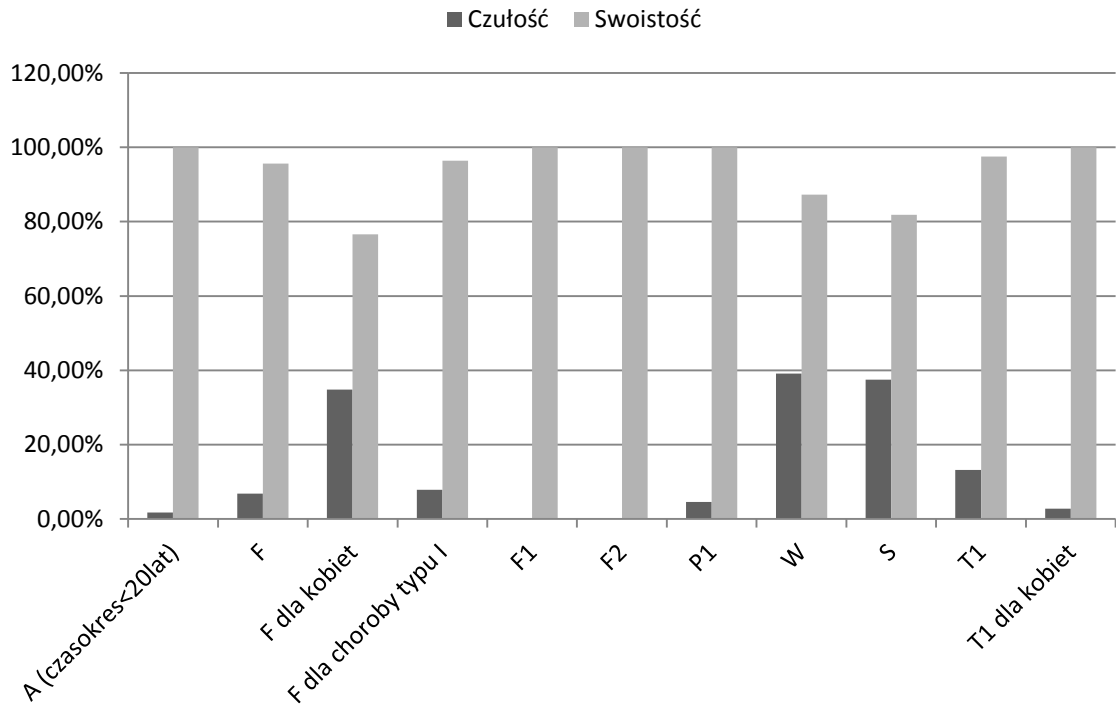


Krzywa ROC dla modelu A  
V-krotna ocena krzyżowa  
(AUC=0,53, p=0,3560)



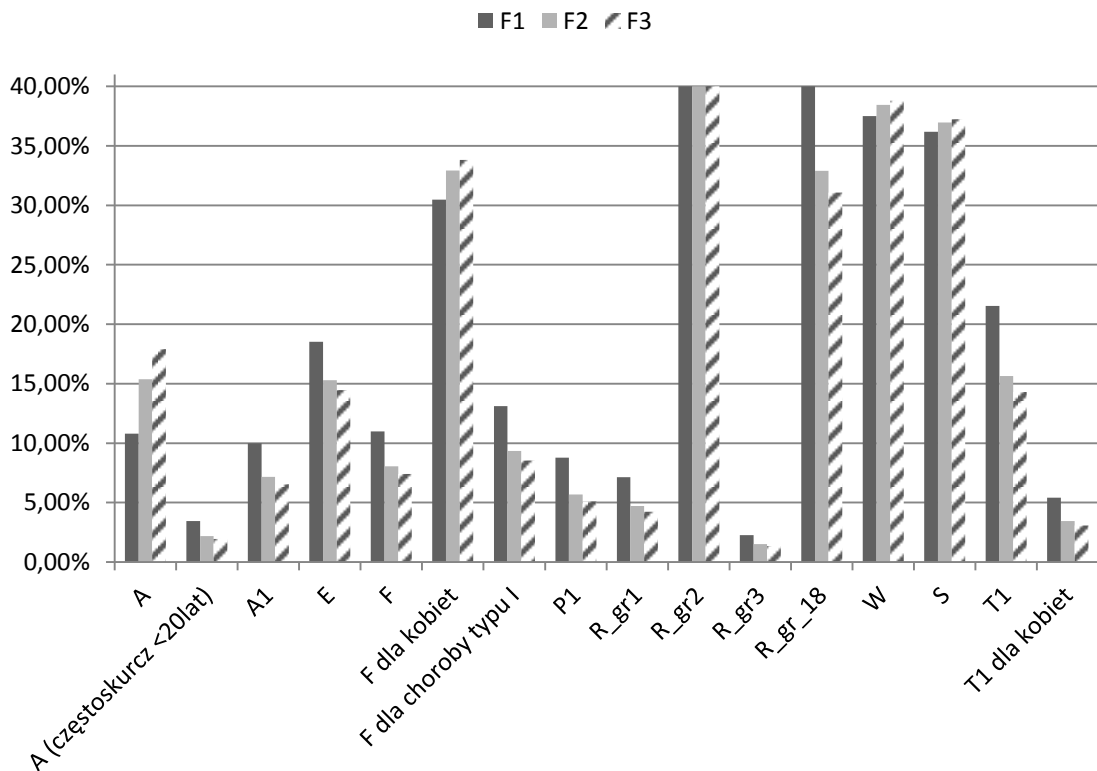
Rycina 32. Porównanie krzywych ROC dla modelu A z i bez walidacji.

Na Rycinie 33 pokazano czułości i swoistości wybranych, istotnych statystycznie modeli logistycznych. W każdym, z przedstawionych modeli klasyfikacja omdleń sprowadzała się do ich wykluczania – świadczą o tym dużo wyższe wartości swoistości. Ocena modeli poprzez pryzmat ich czułości i swoistości, może być jednak znacznie obciążona błędem powodowanym przez niezbalansowanie danych. Dlatego Rycina 34 prezentuje F-miary wybranych modeli logistycznych, czyli średnie harmoniczne czułości i precyzji. (Miara F3 oznacza nadanie 3-razy większej wagi czułości).



Rycina 33. Czułość i swoistość istotnych statystycznie modeli logistycznych.

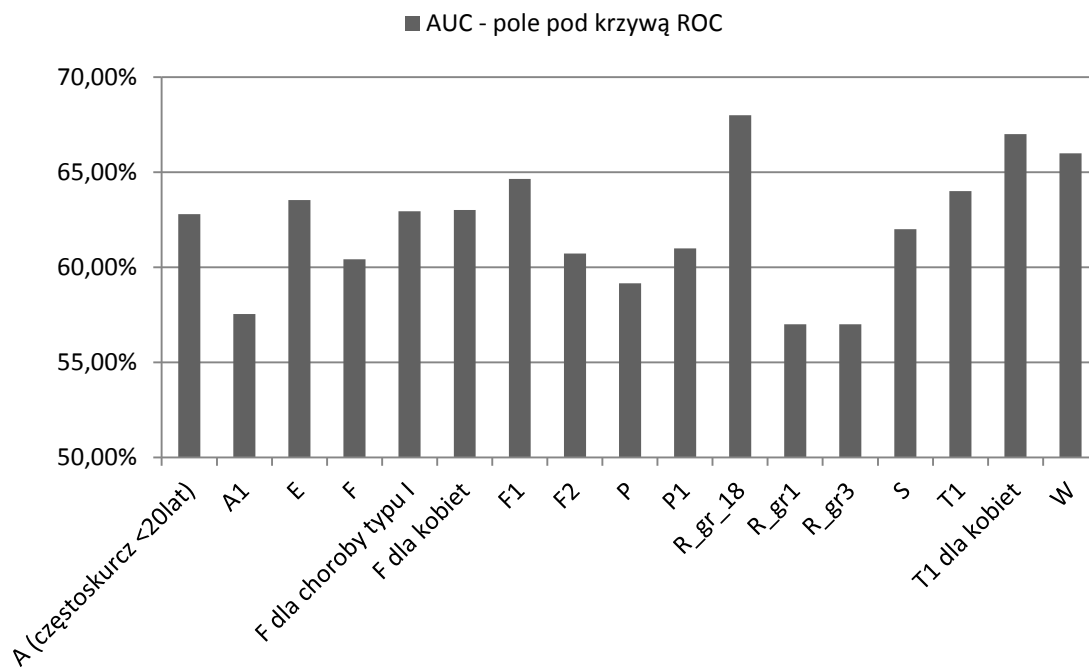
\*definicje prezentowanych modeli logistycznych – Tabela 36 lub szerzej Załącznik nr 1 (Tab. XIX i XX)



Rycina 34. F - miary modeli logistycznych.

\*definicje prezentowanych modeli logistycznych – Tabela 36 lub szerzej Załącznik nr 1 (Tab. XIX i XX)

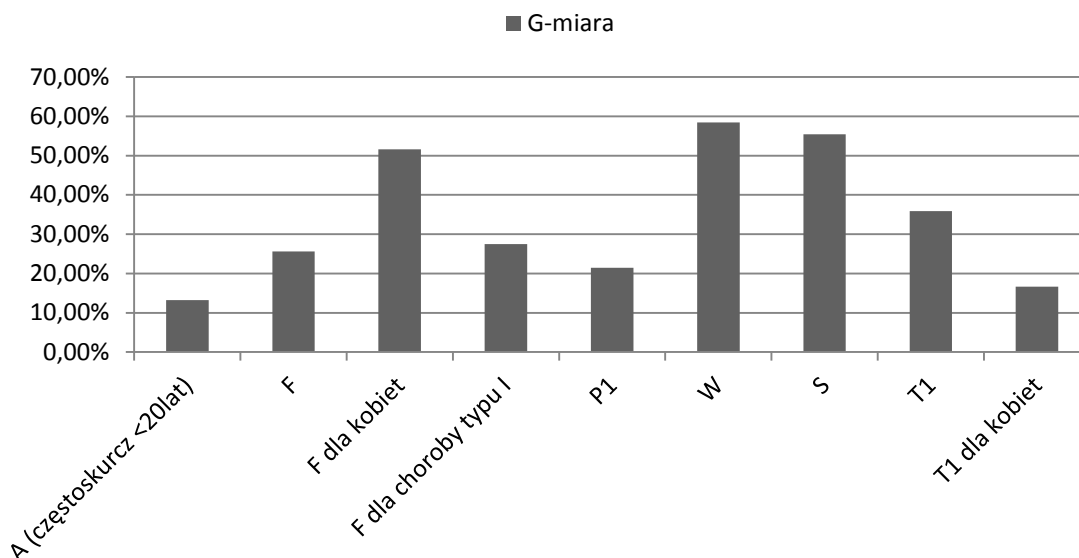
Analizując miary F3 możemy oszacować, iż modele R\_gr2, R\_gr18 (oba nieistotne statystycznie), W, S oraz F dla kobiet dają najlepsze rezultaty. Na Rycinie 35 z kolei, pokazano wielkość pola pod krzywą ROC, istotnie różną od 0,5 – losowego przydziału przypadków do kategorii z i bez omdleń, dla wybranych modeli logistycznych. Analizując miarę AUC zauważono, iż modele punktacji (W i S) bardzo dobrze nadają się do klasyfikacji osób z omdleniami. Bardzo wysoko ocenione zostały również modele T1 i T1 dla kobiet oraz R\_gr18 i F1. Przy czym wszystkie modele typu R nie były istotne statystycznie.



Rycina 35. Istotnie statystycznie AUC - pola pod krzywą ROC dla wybranych modeli logistycznych.

\*definicje prezentowanych modeli logistycznych – Tabela 36 lub szerzej Załącznik nr 1 (Tab. XIX i XX)

G-miary, czyli geometryczne średnie czułości i precyzji (Rycina 36), potwierdzają przewagę klasyfikatorów W i S, a więc związanych z modelem punktacji, budowanym z danych po dyskretyzacji i po przekodowaniu ich ze względu na podgrupy jednakowej wagi dowodu (WoE).



Rycina 36. G-miary istotnych modeli logistycznych.

\*definicje prezentowanych modeli logistycznych – Tabela 36 lub szerzej Załącznik nr 1 (Tab. XIX i XX)

W przypadku metod bazujących na zbiorach przybliżonych, wyniki 10-krotnej walidacji krzyżowej pokazują, iż wygenerowanych reguł dla zbioru nierozszerzonego nie można uogólniać na całą populację, gdyż uzyskujemy najwyżej 50.34+-20.33 procent prawidłowych klasyfikacji omdleń (dla zbioru po dyskretyzacji). (Tabela 31)

Tabela 38. Ocena reguł decyzyjne utworzonych różnymi metodami eksploracji danych (data-minig).

reguły	Dokładność	Czułość	Swoistość	G-miara	F-miara
Drzewa decyzyjne	0,7063	0,2089	0,8351	0,3663	0,2208
Zbiory przybliżone (algorytm LEM2)	0,7642	0,0522	0,9486	0,1396	0,0792
Zbiory przybliżone (algorytm LEM2; po „nadlosowaniu” - SMOTE)	0,4551	0,7622	0,3757	0,5303	0,3651
Zbiory przybliżone (algorytm LEM2; dane po dyskretyzacji; po „nadlosowaniu” - SMOTE)	0,8043	0,9615	0,7743	0,8628	0,5769
BRACID	0,5033	0,4702	0,5119	0,4806	0,2786

Dodatkowe uzupełnienie braków danych powodowało obniżenie wykrywalności omdleń. Dużo wyższe wartości średniej dokładności odnotowano za to dla zbioru z dodanymi dodatkowymi przypadkami klasy mniejszościowej.



Oznacza to, iż na podstawie analizowanych systemów informacyjnych nie możemy utworzyć wiarygodnego modelu oceniającego ryzyko występowania omdlenia. Jednak dokładność nie jest miarą przydatną w przypadku klas niezbalansowanych.

Przedstawione w Tabeli 38 miary, oceniające reguły utworzone technikami eksploracji danych, potwierdzają wyższe zdolności wykrywania omdleń w przypadku analizy danych „nadlosowanych”. Drzewa decyzyjne natomiast, dają rezultaty porównywalne z modelami punktacji. Dla algorytmu BRACID nie uzyskano najlepszej czułości i swoistości, choć wartości G-miary i F-miary były rzeczywiście wysokie.

Zbiór reguł decyzyjnych, uzyskanych z zastosowaniem algorytmu BRACID, szczególnie dedykowanego dla danych niezbalansowanych, zamieszczony został na podstronie Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu [www.ump.edu.pl](http://www.ump.edu.pl), związanej z niniejszą pracą doktorską (<http://www.ump.edu.pl/index.php?department=70>).

---

## 5 Omówienie wyników i dyskusja

---

Omdlenie występuje często jako niegroźny, samoograniczający objaw. Jednak w wielu przypadkach jest zwiastunem realnego zagrożenia życia pacjenta [1]. W ostatnich dekadach podejmowano próby, uściślenia oceny klinicznej pacjenta z omdleniem, w celu utworzenia algorytmu do najbardziej poprawnego rozpoznawania przyczyny omdlenia, a tym samym wyselekcjonowania pacjentów z realnym zagrożeniem życia.

Zaproponowane wcześniej algorytmy rozpoznawania przyczyny omdlenia [91] [3] poprawiają różnicowanie pomiędzy trzema głównymi przyczynami omdleń (neurogenne, ortostatyczne oraz sercowe), jednak nadal nie identyfikują ponad 30% omdleń w populacji. Stwierdzono, że omdlenia powstające w mechanizmie sercowym (sercowo-naczyniowym) są niezależnym czynnikiem nagłego zgonu i śmierci. W 5-letniej obserwacji, 50,5% śmiertelność w grupie omdlenia w mechanizmie sercowym, była wyższa niż 30% śmiertelność wśród przyczyn nie sercowych. [91]

W analizowanej bazie, w prezentowanej pracy, etiologia wszystkich epizodów omdleń nie jest w pełni poznana, jednak ze względu na zdiagnozowany napadowy częstoskurcz przyjęto założenie, iż opisane w pracy przypadki omdleń, to grupa z omdleniami o podłożu arytmicznym, czyli powstające w mechanizmie sercowym. Najważniejszym z ocenianych w niniejszej pracy, potencjalnych czynników ryzyka była maksymalna częstotliwość rytmu serca (HR) w czasie częstoskurczu. Częstoskurcze mogą wywoływać omdlenia, ponieważ szybki rytm komór obniża rzut serca i upośledza dopływ krwi do mózgu. Omdlenie zdarza się najczęściej na początku częstoskurczu lub w końcowej jego fazie. U niektórych chorych mogą pojawić się objawy dławicowe lub niewydolności serca. Omdlenie w mechanizmie sercowym nazwane jest manifestem arytmii zagrażającej życiu. Napadowy częstoskurcz z wąskim zespołem QRS, z bardzo szybkim rytmem serca podczas częstoskurczu, nawet u młodych osób, bez jawnych cech choroby serca może powodować gwałtowne zaburzenie hemodynamiki [50]. Pacjenci z omdleniami w opisanym mechanizmie, winni być możliwie szybko wyselekcjonowani i leczeni, najlepiej ablacją RF substratu arytmii. Dlatego w niniejszej pracy podjęto próbę wyselekcjonowania grupy ryzyka na podstawie przeanalizowanych danych klinicznych oraz wyników badania elektrofizjologicznego serca.

Decyzja o podjęciu właściwego leczenia, szczególnie w przypadkach arytmii pozornie łagodnych, zależy głównie od rodzaju informacji udzielonych lekarzowi pierwszego kontaktu. Jako pomoc przy ocenie ryzyka wystąpienia epizodów omdleń, niezwykle użyteczne mogą okazać się algorytmy wspomagające diagnozę (przedstawione np. w formie graficznej, jako tzw. „schematy blokowego postępowania”). Korzystanie z takiego spisu rozwiązań postawionego problemu medycznego, powinno być możliwie najbardziej intuicyjne i przejrzyste. Dlatego w niniejszej pracy poszukiwano najlepszych, pomocnych dla lekarza reguł, modeli punktacji (tworzących tzw. „check listy”) oraz schematów drzew decyzyjnych.

Ponieważ zwykłe analizy jednowymiarowe nie rozwiązały analizowanego problemu medycznego, zaistniała potrzeba wykorzystania technik wielowymiarowych oraz metod eksploracji danych w celu zbadania bardziej złożonych efektów.

Analizowane dane zostały również przeanalizowane w celu odnalezienia efektów mediacyjnych, interakcyjnych, czy wnikających. Poszukiwanie mediatorów możliwe jest w sytuacji występowania silnej zależności, w celu przeanalizowania procesów pośredniczących, tłumaczących tą zależność. W niniejszej pracy nie odnotowano żadnych istotnych mediatorów, opisano za to zmienne moderujące i wnikające.

## **5.1 Istniejący stan wiedzy w zakresie tematu badań**

Analizując przyczyny omdleń w populacji osób ze zdiagnozowaną arytmia, w sposób naturalny skoncentrowano się na częstotliwości rytmu serca (HR) w czasie częstoskurczu, ponieważ jest to uznany czynnik mogący powodować omdlenie w mechanizmie sercowym [30] [92] [93].

Wysoka częstotliwość rytmu serca ma niekorzystny wpływ na układ sercowo-naczyniowy, gdyż bezpośrednio przyczynia się do hipoperfuzji mózgu, a także innych organów. Przyspieszona czynność serca, ale niekoniecznie częstoskurcz, powoduje skrócenie okresu napełniania komór krwią, a tym samym zmniejszenie rzutu serca oraz zwiększone zużycie tlenu i gorszy przepływ wieńcowy. U pacjentów z organiczną chorobą serca, dodatkowo pogłębia objawy niewydolności serca. Często przyspieszona czynność serca może stanowić niezależny powód hospitalizacji u pacjenta z objawową niewydolnością serca. Istnieje hipoteza, iż wzrost częstotliwości rytmu serca powoduje zmianę prędkości i kierunku przepływu krwi oraz poprzez mechaniczne uszkodzenie śródbłonna naczyniowego sprzyja procesom miażdżycowym i niestabilności blaszki miażdżycowej.

Aktualny stan wiedzy na temat wpływu częstotliwości rytmu serca w częstoskurczu na występowanie epizodów omdleń charakteryzują dwa przeciwstawne stanowiska klinicystów opisane poniżej.

Praca z 1997 roku („Frequency of Disabling Symptoms in Supraventricular Tachycardia”) dowodziła o zależność omawianych cech: im wyższa częstotliwość rytmu serca w częstoskurczu, tym większe ryzyko chwilowej utraty przytomności. [31] W tej pracy analizowano chorych z SVT, czyli z napadowym częstoskurczem „nadkomorowym” w nieco szerszym pojęciu: oprócz grupy AVNRT i AVRT/WPW dodano także częstoskurcz przedsionkowy i trzepotanie przedsionków. Byli to pacjenci, którzy doświadczali epizody napadowego częstoskurczu średnio od 13 lat. Autorzy sugerują, iż taka grupa może mieć poważniejsze objawy kliniczne, w porównaniu do osób, u których rozpoznano i stwierdzano kolejne epizody częstoskurczu w stosunkowo krótszym okresie czasu. Przedstawiony wynik wskazywał na istotnie częstsze występowanie omdleń i zawrotów głowy w grupie z HR częstoskurczu co najmniej 170 uderzeń na minutę. W tej samej pracy powoływano się jednak na inne publikacje, których wyniki były sprzeczne.

Z kolei, badanie starsze o dekadę („Are the results of electrophysiological study different in patients with a pre-excitation syndrome, with and without syncope?”), ale dotyczące jedynie chorych z WPW, dowodziło o braku wpływu częstotliwości częstoskurczu na pojawianie się epizodów omdleń [32]. W publikacji tej wykazano natomiast zwiększoną indukowalność częstoskurczu w grupie z omdleniami.

Okazuje się więc, iż dotychczasowe badania nie rozstrzygają jednoznacznie sprawy związku częstotliwości HR częstoskurczu z występowaniem epizodów omdleń u pacjentów z wąskim zespołem QRS. Sprawa ta wymagała zatem wiarygodnego sprawdzenia oraz podjęcia próby utworzenia modelu charakteryzującego pacjentów z epizodami omdleń w analizowanej populacji.

Już w trakcie przygotowywania niniejszej pracy, w czerwcu 2013 roku zaprezentowano temat „Omdlenia w grupie częstoskurczu z wąskim QRS typu AVNRT oraz AVRT. ” na XXIV Konferencji Sekcji Rytmu Serca Polskiego Towarzystwa Kardiologicznego (POLSTIM 2013; 12-15 czerwca 2013; Wrocław) [90], co także świadczy o aktualności i przydatności klinicznej prowadzonych poszukiwań badawczych.

## **5.2 Czynniki powiązane z częstotliwością HR częstoskurczu**

Zaburzenia rytmu i przewodzenia często są warunkowane genetycznie. Mogą pojawić się także na skutek stosowania różnych leków, a także w

przebiegu urazu serca lub zapalenia mięśnia sercowego. Ryzyko arytmii zwiększa palenie papierosów, zażywanie narkotyków, nadmierne spożywanie alkoholu oraz przewlekły stres. [27] Do występowania arytmii przyczyniają się również zaburzenia w składzie elektrolitowym krwi (zbyt wysokie lub zbyt niskie stężenie potasu, sodu lub magnezu), a także zaburzenia endokrynologiczne. [48]

Powszechnie znany fakt zwolnienia częstotliwości rytmu serca wraz z wiekiem [37] został potwierdzony również w analizowanej populacji, czyli w grupie częstoskurczu z wąskim QRS typu AVNRT oraz AVRT. Układ autonomiczny wydaje się najważniejszym, ale na pewno nie jedynym modulatorem HR. Wzrost HR może być między innymi: wskaźnikiem nadmiernej aktywacji układu współczulnego, obniżonej aktywności fizycznej i utraty rezerwy sercowej. Podczas częstoskurczu, HR wzrasta gwałtownie i w sposób istotny przewyższa należne wartości HR wyliczane wg należnego wieku. W prezentowanej pracy było to widoczne, zarówno dla grupy z epizodami omdleń, jak i bez nich.

Zbadanie jakie czynniki powiązane są z częstotliwością HR częstoskurczu w analizowanej populacji może mieć znaczenie przy próbie odpowiedzi na pytanie o ewentualny związek częstotliwości HR częstoskurczu z występowaniem epizodów omdleń.

U osób starszych, wpływ wieku na HR częstoskurczu widoczny jest znacznie wyraźniej niż u młodszych. Wykazano także, że w niektórych przypadkach pacjenci starsi nie mieli niższych częstotliwości HR częstoskurczu standaryzowanego, jak można by się było spodziewać. Taka sytuacja miała miejsce w przypadku określonego rodzaju arytmii: AVNRT było grupą wiekowo istotnie starszą, ale wcale nie o niższych wartościach częstotliwości HR częstoskurczu. Podobna sytuacja miała miejsce w podgrupie WPW dla chorób typu I i II, a także w podgrupie mężczyzn w przypadku chorób typu I oraz w podgrupach z dłuższym i krótszym czasokresem (w latach) od pierwszego pojawienia się częstoskurczu (czasokres objawów kołatania i częstoskurczu <20 lat). We wszystkich tych przypadkach starsi pacjenci nie mieli niższych wartości częstotliwości HR częstoskurczu. Zaobserwowano także różnice w znaczeniu działania układu współczulnego w zależności od płci i wieku. Mężczyźni w analizowanej populacji, u których częstoskurcze pojawiały się raczej w ciągu dnia (grupa o widocznym działaniu układu współczulnego) byli istotnie młodszy i, jak można było przypuszczać, mieli istotnie szybsze częstoskurcze. Kobiety w grupie „dziennych” częstoskurczów również były młodsze, ale nie odnotowano u nich istotnie szybszych częstoskurczów.

Zestawienie podgrup, u których wpływ wieku na częstotliwość HR częstoskurczu został zniwelowany przedstawiono w Tabeli 10.

Wpływ zarówno płci, jak i rodzaju arytmii, na częstotliwość rytmu serca HR częstoskurczu przejawiał się także poprzez brak istotnych różnic pomiędzy częstotliwością HR częstoskurczu od wyliczonego względem wieku jego wartości należnej HR. Miało to miejsce w grupie WPW oraz dla mężczyzn. W pozostałych przypadkach zawsze obserwowana HR podczas częstoskurczu była istotnie wyższa od wartości należnej.

Analiza modeli wielowymiarowych potwierdziła znaczenie zmiennych wnikających (opisanych przez statystyki jednowymiarowe), takich jak płeć, rodzaj arytmii, czy występowanie stanów przedomdleniowych.

### **5.3 Czynniki powiązane z występowaniem epizodów omdleń**

Próbie wyselekcjonowania grupy ryzyka omdleń w grupie pacjentów z arytmia, z wąskim zespołem QRS (AVNRT i AVNT), rozpoczęto od przeglądu dostępnych danych klinicznych oraz wyników badania elektrofizjologicznego serca.

Po przeanalizowaniu wyników statystyk jednowymiarowych wykazano, iż podwyższony wskaźnik BMI był dla mężczyzn czynnikiem ryzyka wystąpienia epizodów omdleń. W przypadku kobiet nie było takiej zależności. Jednocześnie niedowaga uznana została za istotny czynnik obniżający ryzyko omdlenia.

Grupa zwiększonego ryzyka występowania epizodów omdleń charakteryzowała się także wyższymi poziomami potasu we krwi. Ma on duże znaczenie w organizmie, gdyż jest elektrolitem śródkomórkowym. Potas odpowiedzialny jest za prawidłowe funkcjonowanie układu mięśniowego oraz nerwowego, wpływa na przewodnictwo nerwowo-mięśniowe. Reguluje gospodarkę wodną organizmu i odpowiedzialny jest za pracę niektórych enzymów. Zbyt duże jego stężenie obserwujemy w wielu groźnych sytuacjach klinicznych, a także może skutkować zwolnieniem akcji serca, zaburzeniami w zapisie EKG. [48] [49]

Poziomy jonów potasu (K<sup>+</sup>) w grupie z i bez omdleń różniły się istotnie, choć ich wartości nie były wysokie i nie były istotne klinicznie (dla omdleń mediana wynosiła 4,3 mmol/l, natomiast w grupie bez odnotowanych epizodów omdleń było to 4,2 mmol/l.).

Hormon tyreotropowy, zwany w skrócie TSH to hormon wydzielany przez przysadkę mózgową i regulujący pracę tarczycy wydzielającej hormony mające wpływ na przebieg wielu procesów w organizmie człowieka. Regulacja

wydzielania hormonu TSH odbywa się na zasadzie ujemnego sprzężenia zwrotnego z hormonami tarczycy (FT3 i FT4). [38] W niniejszej pracy wykazano, iż kobiety z omdleniami miały istotnie niższe wartości TSH. Jednak mediany wartości TSH w obu grupach nie przekraczały wartości dopuszczalnych (norma: 0,27 a 4,2 mU/l) i wynosiły 1,11 dla omdleń i 1,41 w pozostałych przypadkach.

Niekiedy obserwuje się istotny wpływ pobudzenia układu współczulnego jako czynnika modulującego w grupie pacjentów z omdleniami. [3] Występowanie omdleń w czasie wysiłku fizycznego (zwiększone napięcie układu współczulnego) przy braku współwystępujących objawów jawnej choroby serca oraz młody wiek pacjenta przemawia za mechanizmem odruchowym (neurogennym). [1] Natomiast w grupie pacjentów z współwystępowaniem kołatania serca, udokumentowanym częstoskurczem (omdlenia sercowe), dodatkowe pobudzenie układu współczulnego zazwyczaj zwiększa częstotliwość rytmu serca w częstoskurczu. W pracy zaobserwowano różnice w znaczeniu działania układu współczulnego w zależności od płci i wieku.

Ponadto wykazano, iż jeśli czasokres od pierwszego udokumentowanego częstoskurczu nie przekraczał 20 lat, zwiększenie wieku o 40 lat spowoduje ok. 3,5-krotnie większą szansę wystąpienia epizodu omdlenia (przy niezmiennym HR częstoskurczu).

Dodanie informacji o współistniejących chorobach, do modelu zawierającego HR częstoskurczu i zmienną określającą występowanie stanów przedomdleniowych, istotnie polepszało wykrywalność omdleń. Współwystępowanie jawnych objawów innej choroby znacząco podwyższało szansę omdlenia.

W niektórych przypadkach dodanie do modelu interakcji analizowanych zmiennych, daje pełniejsze wyjaśnienie wariancji zmiennej zależnej, w porównaniu z modelem bez interakcji. U osób z omdleniami stwierdzono istotnie większe wymiary jam serca w mm – lewy przedsionek (LP) i prawa komora (PK) serca. W pracy wykazano, iż wymiary LP i PK są ze sobą ściśle powiązane i wartość jednej z nich decyduje o kierunku lub sile zależności drugiej zmiennej ze zjawiskiem pojawiania się omdleń (model logistyczny P1 – oceniający zmienne LP, PK oraz ich interakcję). Osoby mające mniejsze wymiary LP związane były ze wzrostem prawdopodobieństwa omdlenia wraz ze wzrostem wymiaru PK. Dla osób o większych wymiarach LP prawdopodobieństwo omdlenia malało wraz ze zwiększającym się wymiarem PK. Fakt ten może świadczyć o istotnym wpływie proporcjonalnej budowy serca na występowanie

epizodów omdleń. [2]

Kolejny moderator odnaleziony przez model z istotną interakcją, to wiek w modelu logistycznym T1 (oceniającym zmienne TSH, wiek oraz ich interakcję) dla kobiet. Wykazano tam, iż TSH może być uznane za czynnik ryzyka w grupie kobiet w analizowanej populacji. Spadek wartości TSH powiązany był ze zwiększeniem prawdopodobieństwa omdlenia, jednak w grupie starszej wiekiem nie było to tak widoczne, jak u młodszych pacjentek, u których jednostkowy wzrost wartości TSH wiązał się z trzykrotnym obniżeniem szansy wystąpienia omdlenia. Omawiany model T1 był dobrze zdeterminowany i istotny statystycznie.

Przeanalizowanie modelu regresji logistycznej zawierającego wiek i HR częstoskurczu (model logistyczny A) sugeruje, iż dla takiego samego częstoskurczu zwiększenie wieku o 1 rok zwiększa szansę wystąpienia omdlenia o 1,017. W przypadku zwiększenia wieku o 40 lat szansa wystąpienia omdlenia wzrasta 2-krotnie.

Ponadto wykazano, iż pomimo tego, że częstotliwość rytmu serca podczas częstoskurczu nie determinuje grupy omdleń w sposób bezpośredni, może być uznana za czynnik ryzyka wystąpienia omdlenia w grupie po podaniu leku (Isoprenaliny) podczas badania EP, w przypadku osób nieobarczonych dodatkową, jawną chorobą współwystępującą. [90]

Również podwyższone szanse wystąpienia omdlenia w grupie mężczyzn, w przypadku krótszego cyklu częstoskurczu (CL) mierzonego (w ms) podczas badania EP, może świadczyć o ewentualnym działaniu negatywnym wysokich częstotliwości HR częstoskurczu. Jednak trzeba mieć na uwadze, iż wartości CL częstoskurczu mierzone podczas badania EP, które mogą być znacząco skrócone po doraźnym zastosowaniu leku iv. (Isoprenalina), nie koniecznie oddają faktyczny HR częstoskurczu danego pacjenta. Można jednak przypuszczać, iż grupa mężczyzn ma większe predyspozycje do omdleń w przypadku wyższych częstoskurczów.

Zastosowanie drzew decyzyjnych, generowanych za pomocą algorytmu losowy las, przyniosło korzyść w postaci odnalezienia reguł opisujących przypadki zmniejszonego prawdopodobieństwa wystąpienia omdlenia. Jedną z nich dotyczyła poziomu potasu i długości cyklu częstoskurczu CL z badania EP: ('K+'<=4,365) AND ('CL'>353)

Również niespełnianie warunku: częstotliwość HR częstoskurczu <= 220 i jednocześnie brak stanów poprzedzających epizody omdleń – uznano za istotny czynnik ryzyka omdleń (OR=6).



Poszukiwanie przyczyn omdleń w analizowanej populacji za pomocą algorytmów stosujących teorię zbiorów przybliżonych, dawało bardzo długie reguły. Np. zakwalifikowanie pacjenta jako niezagrażonego omdleniem: co najwyżej 41 lat, bez obciążenia żadną dodatkową, jawną chorobą, kołatania serca pojawiające się w ciągu dnia (choć nie tylko), BMI w normie lub niedowaga, i jednocześnie wartość echokardiograficzna grubości przegrody międzykomorowej (PMK) co najwyżej 8 mm, natomiast szerokość aorty (Ao) poniżej 25 mm. Z kolei przykładowa klasyfikacja osób zagrożonych omdleniem, to np.: osoby cięższe i grubsze, obciążone dodatkowymi chorobami i z większym wymiarem aorty Ao, u których również kołatania pojawiają się w ciągu dnia.

#### **5.4 Porównanie zastosowanych metod analizy danych**

Problem predykcji ryzyka zachorowania w medycynie najczęściej jest formułowany jako problem klasyfikacji. Dotyczy to także, analizowanej w tej pracy, potrzeby wyjaśnienia czynników powiązanych z występowaniem epizodów omdleń. Proces klasyfikacji jest podstawową czynnością poznawczą wykonywaną przez człowieka, polegającą na podziale obiektów na grupy i kategorie [23].

Wynikiem procesu klasyfikacji, jest zbudowanie klasyfikatora (np. w formie drzewa decyzyjnego, czy zestawu reguł klasyfikacyjnych), potrafiącego samodzielnie rozpoznawać (w oparciu o wyuczone wcześniej wzorce) przynależność nowych przypadków do odpowiednich klas.

Jakość predykcji klasyfikatora często stawiana jest na równi z dobrym uzasadnieniem podjętej decyzji. To właśnie generowanie czytelnej wiedzy stało się inspiracją przy poszukiwaniu odpowiednich metod analizy danych. W opisie różnych metod klasyfikacji pominięto systemy oparte na logice rozmytej i sieci neuronowe (również stosowane do klasyfikacji). Uznano, iż nie ma potrzeby opisu tych technik, ponieważ są to typowe metody tzw. „czarnej skrzynki”, utrudniające swobodną interpretację oraz ingerencję w stworzoną bazę wiedzy. Brak jasno zdefiniowanych wzorców uniemożliwia wnikliwą weryfikację ocenianego zdarzenia i w wypadku modeli mających wspierać diagnozę kliniczną, całkowicie deklasuje takie rozwiązania

W celu odnalezienia potencjalnych czynników ryzyka wystąpienia epizodów omdleń, zastosowano całą serię analiz jedno- i wielowymiarowych. W badaniach zajmujących się prognozowaniem wystąpienia lub braku wystąpienia, pewnego zdarzenia, do grupy kanonu stosowanych metod statystycznych należy regresja logistyczna. Jest ona jedną z najczęściej

stosowanych metod analizy danych medycznych, w której zmienną objaśnianą jest zmienna dychotomiczna (określająca tu występowanie, bądź brak występowania omdlenia). W sposób naturalny i całkowicie intuicyjny wiąże ona oceniane parametry modelu z szansą zaistnienia analizowanego epizodu. Na tle metod tzw. nowej generacji (jak chociażby sieci neuronowe), modele logistyczne uznawane są za mniej kosztowne i bardziej przejrzyste. Ocenione parametry modelu mają naturalną, przyjazną interpretację.

Zastosowane w pierwotnych obliczeniach podziały klasy przykładów i dokonanie porównań w podgrupach płci, rodzaju arytmii, pory dnia występowania częstoskurczu itp. spowodowała dodatkowe zmniejszenie liczności grupy przykładów mniejszościowych. Stąd powstałe skupienia przypadków z epizodami omdleń mają charakterystykę zbliżoną do „szumu informacyjnego”. Nachodzenie się obszarów występowania przykładów z różnych klas (z omdleniami i bez omdleń), może powodować trudności w określeniu granicy pomiędzy klasami, przyczyniając się do złej klasyfikacji przykładów przygranicznych. Powoduje to, iż rozpatrywany problem nie jest liniowo-separowalny.

Jeśli algorytm klasyfikacyjny ma służyć do wspierania diagnozy rzadko występującej choroby, często dane uczące, będące historycznymi danymi pacjentów, zawierają bardzo dużo opisów przypadków, w których nie wykryto danej choroby oraz kilka opisów dotyczących pacjentów z jej rozpoznaniem. Ponadto, często dane medyczne charakteryzują się licznymi brakami danych. W pracy celowo nie zastosowano metod uzupełniania braków danych np. wielokrotnego przypisywania (ang. multiple imputation). Analizowany zbiór danych, z racji specyficznych własności analizowanej populacji, nie jest duży. Jest za to odzwierciedleniem nieidealnej rzeczywistości – braki danych są uzasadnione. Skutkuje to błędami lub sprzecznościami w analizowanym zbiorze danych. Jednak „sztuczne” uzupełnianie danych mogłoby prowadzić do niewłaściwych wniosków, co w przypadku poszukiwania wsparcia dla decyzji klinicznej, mogłoby mieć zbyt poważne skutki.

Uczenie z nie zrównoważonych danych (przykłady reprezentujące jedną klasę są dużo mniej liczne niż przykłady z pozostałych klas), jest istotnym problemem badawczym, ponieważ klasyczne algorytmy uczące, które były tworzone przy założeniu, że klasy są w przybliżeniu równomiernie reprezentowane w danych, są nadmiernie ukierunkowane na rozpoznawanie klasy większościowej, pomijając klasę mniejszościową. W skrajnych przypadkach, klasyfikator przydziela wszystkie przykłady uczące do klasy większościowej a jednocześnie prawidłowe rozpoznanie przykładów z tej mało

licznej klasy jest kluczowe. W analizowanym przypadku klasą mniejszościową są przypadki wystąpienia omdleń. Jednocześnie wiadomo, iż niepoprawne zdiagnozowanie osoby chorej i niepodjęcie leczenia, jest bardziej krytyczne niż rozpoznanie choroby u osoby zdrowej, u której diagnoza może być skorygowana w trakcie kolejnych badań.

Problem uczenia z takich niezbalansowanych danych nie wynika wyłącznie ze stopnia niezrównoważenia liczebności klas [94]. Wielu badaczy uważa, iż jeśli klasy są łatwo separowalne, standardowe metody uczenia mimo wszystko potrafią poprawnie nauczyć się definicji obu klas, bez względu na ich stopień niezrównoważenia.

Badania w tym zakresie prowadzone były głównie na sztucznie generowanych zbiorach danych. W zderzeniu z rzeczywistymi danymi medycznymi, obarczonymi brakami danych i niedokładnymi pomiarami, istniejące rozwiązania dedykowane danym niezrównoważonym, mogą być niewystarczające.

Jednym z proponowanych rozwiązań jest wykorzystanie wiedzy eksperckiej do klasyfikowania krytycznych, tzn. wątpliwych przykładów uczących. Podejście to może w przyszłości poprawić rozpoznawanie klasy mniejszościowej dla danych niezbalansowanych, nie powodując jednocześnie pogorszenia prawidłowego rozpoznawania przypadków z klasy większościowej [95].

Popularna regresja logistyczna nie jest najwłaściwszym rozwiązaniem w przypadku danych niezbalansowanych. Modele uzyskane za pomocą tej metody nie nadają się zazwyczaj do poprawnego rozpoznawania przypadków rzadkich. Mogą natomiast służyć jako pomoc przy ewentualnym wykluczaniu rzadko występujących przypadków klasy mniejszościowej, czyli epizodów omdleń.

W małych próbach estymatory parametrów modelu regresji logistycznej uzyskiwane metodą klasyczną, tzn. metodą największej wiarygodności (MNW), charakteryzują się znacznym obciążeniem [96]. W niniejszej pracy dołożono szczególnych starań, aby dane zostały wykorzystane maksymalnie efektywnie, a wnioskowanie było obarczone jak najmniejszym błędem systematycznym. Dlatego zastosowano model regresji logistycznej Firtha, który jest niewielką modyfikacją klasycznego modelu regresji logistycznej, natomiast estymowane parametry modelu charakteryzują się lepszymi właściwościami probabilistycznymi. Zastosowanie modelu regresji logistycznej Firtha [96] rozwiązuje także problem całkowitego rozdzielenia (ang. complete separation). Występuje ono wtedy, gdy sukcesy i porażki mogą być całkowicie rozdzielone za pomocą jednej zmiennej niezależnej lub liniowej kombinacji kilku zmiennych

niezależnych. W modelu Firtha zmienne powodujące problem całkowitego rozdzielenia mogą pozostać w modelu, bez szkody dla jego predykcji.

W pracy poruszony został również problem obliczania asymptotycznej wartości testu. Każdy z powszechnie stosowanych asymptotycznych testów statystycznych wymaga spełnienia wielu warunków, aby wyznaczone przez test prawdopodobieństwo  $p$  było wiarygodne. W przypadku wystarczająco dużych i dobrze zrównoważonych zbiorów danych różnice między wartością asymptotyczną przeprowadzonego testu a rzeczywistą są zazwyczaj zaniedbywane. Istnieją jednak pewne układy danych, dla których różnice te mają znaczenie. Ze względu na małocenne próby oraz niezrównoważenie danych, aby uniknąć ewentualnego ryzyka nieprawidłowego oszacowania asymptotycznej wartości prawdopodobieństwa testów, dla części analiz, zdecydowano o wyznaczeniu dokładnej wartości prawdopodobieństwa. Niestety, metody exact wymagają wielkiej mocy obliczeniowej komputera, są z reguły bardzo czasochłonne oraz rzadko implementowane w dostępnych pakietach statystycznych. Utrudnia to, a w niektórych przypadkach całkowicie uniemożliwia, uzyskanie testu exact odpowiadającego wykorzystywanej metodzie statystycznej. Z tych przyczyn obliczenia dokładne zastępowane są często metodą symulacyjną – Monte Carlo. W metodzie tej uzyskiwany jest nieobciążony estymator prawdopodobieństwa wraz z przedziałem ufności na określonym poziomie istotności [59]. Obliczone w niniejszej pracy wyniki testów dokładnych oraz symulacji MC, dla wybranych analiz, nie różniły się znacząco od wyników oszacowanych przez klasyczne testy asymptotyczne (rozdział 4.3.6), co świadczy o wystarczającym dopasowaniu asymptotycznych rozkładów. Dlatego podczas dalszego porównywania metod klasyfikacyjnych, wyniki testów dokładnych i symulacji Monte Carlo nie były uwzględniane.

Ze względu na dużą liczbę zmiennych objaśniających, podczas analizowania danych EP w niniejszej pracy, zastosowano redukcję wymiaru. Jest to sposób zapobiegający przeuczeniu modelu, czyli nadmiernemu dopasowaniu się (ang. overfitting) modelu do przypadkowych błędów w danych uczących oraz zanikaniu jego zdolności interpretacyjnych. Niestety modele logistyczne utworzone z użyciem metod redukcji wymiaru nie były istotne statystycznie.

Również modele scoringowe, z udziałem regresji logistycznej, stosują redukcję zmiennych nadmiernie skorelowanych (jest to tzw. wybór reprezentantów). Tworzone przez taką redukcję wiązki zmiennych, w sposób analogiczny do składowych głównych, przyczyniają się do redukcji wymiaru analizowanych danych, a co za tym idzie, do łatwiejszej interpretacji powstałych

modeli. Powstały model punktacji (scoringowy) był istotny statystycznie, co może świadczyć o dużym znaczeniu przejrzystości wstępnej analizy danych przed analizą logistyczną w procesie tworzenia modeli. W przypadku modeli scoringowych proces ten jest szczególnie przejrzysty. Odpowiednie opracowanie predyktorów poprzez dyskretyzację zmiennych pozwala wychwycić wiele błędów i sprzeczności występujących w danych. W każdej ze zmiennych można wyróżnić grupy jednorodne ze względu na ryzyko wystąpienia omdlenia, co powoduje wyeliminowanie negatywnego wpływu obserwacji odstających oraz wygładzanie szumów informacyjnych, jakie występują w danych. Przygotowane na tej podstawie zmienne pochodne, kategoriale, używane są następnie do finalnej analizy, co skutkuje naturalnym obsłużeniem braków danych. Pod warunkiem, że nie utworzono zbyt wielu kategorii na skutek dyskretyzacji, modele zawierające dyskretyzowane zmienne są bardziej stabilne i wykazują mniejszą skłonność do przeuczenia. Niestety, występujące podgrupy jednorodnego ryzyka omdlenia, wyznaczone dla analizowanych cech, miały zazwyczaj małą liczebność. Świadczy to o słabym związku analizowanych cech ze zmienną zależną (czyli występowaniem omdleń).

Wstępna analiza predyktorów może przyczynić się do identyfikacji ciekawych podzbiorów danych wykorzystywanych w dalszej analizie. Może także dać nowe spojrzenie na dane i zweryfikować wstępne założenia dotyczące możliwych powiązań pomiędzy atrybutami i zmienną celu. Wszystko to jest niezwykle istotne, szczególnie w przypadku analizowania dużych baz danych, dlatego wstępny dobór predyktorów ma tak wielkie znaczenie w przypadku metod eksploracji danych (data mining). W przypadku tworzenia punktacji (scoringu), wyliczona zostaje ocena, która wyraża prawdopodobieństwo wystąpienia analizowanego zdarzenia. Z pewnością przygotowanie skali punktacji jako wyniku modelu jest jedną z podstawowych zalet metod scoringowych. Jest to wynik zrozumiały i czytelny dla klinicysty, gdyż utworzona punktacja, pozwala w przejrzysty sposób oszacować ryzyko występowania omdlenia.

Prostota stosowania modelu punktacji (scoringu) jest jedną z jego największych zalet. Związana jest z ograniczeniem kosztów analizy, skróceniem czasu weryfikacji nowych przypadków i jednolitości procesu oceny, a co za tym idzie, z coraz szerszym jego zastosowaniem w różnych dziedzinach badań. Jako wady scoringu najczęściej wymieniane są: zbyt duża ogólnikowość i mała elastyczność takich modeli. Należy również podkreślić, iż scoring, jako jedna z metod eksploracji danych, to metoda tzw. "czarnej skrzynki". Otrzymujemy kwantyfikację ryzyka, ale w żadnym razie analiza nie wyjaśnia bezpośrednich

związków przyczynowo skutkowych związanych z obliczonym ryzykiem omdlenia. W niniejszej pracy zmienne: LP, PK, PMK, Na+, wiek oraz częstotliwość HR częstoskurczu, po dyskretyzacji i wyliczeniu ich „wagi dowodu” (WoE), tworzyły istotny model punktacji (scoringowy). Nie jest jednak możliwe pełne wyjaśnienie znaczenia wszystkich tych zmiennych i interakcji zachodzących między nimi.

Modele punktacji stają się coraz bardziej potrzebne w świecie zautomatyzowanych decyzji, szczególnie w przypadku analizowania bardzo dużych zbiorów danych. Należy jednak pamiętać (szczególnie, gdy mówimy o systemach wspierających diagnozę pacjentów), iż systemy scoringowe powinny być przygotowane, i często aktualizowane, w oparciu o niezastąpioną wiedzę ekspercką. Obecnie istniejące algorytmy klasyfikacyjne, wywodzące się z dziedziny uczenia maszynowego, zyskują coraz szersze praktyczne zastosowania. Wciąż pracuje się nad ich ulepszeniem podnoszącym jakość i wydajność klasyfikacji. Opracowywane są również nowe algorytmy działające lepiej dla pewnej ściśle określonej klasy problemów (np. dla danych niezbalansowanych).

Aby uzyskać możliwie najwyższą jakość systemu decyzyjnego oraz zwiększyć czytelność generowanej wiedzy, w niniejszej pracy postawiono duży nacisk na graficzne, zrozumiałe przedstawienie wyników. W tym celu utworzono reguły decyzyjne.

Zbiór losowych reguł decyzyjnych uzyskano za pomocą modułu „losowy las”. (Podział w węzle realizowany jest tam na podstawie najlepszego atrybutu z wylosowanej każdorazowo puli atrybutów.) Jednak cechy charakteryzujące badaną grupę badawczą (głównie: mała liczność grupy z odnotowanymi epizodami omdleń), spowodowały, iż nie uzyskano nigdzie wysokiego wskaźnika „bad rate” (max.0,682). Uzyskano za to duże prawdopodobieństwo wykluczenia wystąpienia epizodów omdleń („bad rate” = 0), a więc dobre reguły opisujące zmniejszone ryzyko omdlenia. Podział zmiennych na przedziały, potrzebne do tworzenia kolejnych węzłów drzewa decyzyjnego, powodował utratę precyzji, skutkowałam jednak dobrym uogólnieniem w modelu.

Największą wadą wygenerowanych losowych drzew decyzyjnych był brak możliwości ich modyfikacji. Dlatego reguły powstałe z losowych drzew, najczęściej nie są do przyjęcia dla lekarza, ze względu na nieistotne klinicznie punkty podziału i kolejność zmiennych w węzłach. Kolejnym problemem jest fakt, iż algorytmy losowe determinują możliwość uzyskania różnych drzew dla tych samych danych. Dlatego stabilność otrzymanych w ten sposób wyników jest często dyskusyjna. W utworzonych drzewa odnotowano także stosunkowo

małe licznosci w końcowych liściach, więc przetestowanie modeli regresji logistycznej dla podziałów zaproponowanych przez te drzewa nie dawało istotności statystycznej. „Losowy las” nie dał zadawalających reguł spełniających oczekiwania w predykcji omdleń, jednak w przejrzysty sposób uwidocznili istnienie czynników zmniejszających prawdopodobieństwo omdlenia w analizowanej populacji.

Generowanie interakcyjnych drzew decyzyjnych w niniejszej pracy, uznano za jedną z najbardziej intuicyjnych form prezentowania wiedzy. Idea konstrukcji drzew klasyfikacyjnych jest bardzo podobna do sytuacji, gdy lekarz zadaje pacjentowi szereg pytań, w celu jego zdiagnozowania. Przy czym kolejność tych pytań nie jest przypadkowa. Każde kolejne, w danej serii, pytanie uzależnione jest od uzyskanej odpowiedzi na pytanie poprzednie. Dzięki drzewom interakcyjnym typu C&RT/CHAID w niniejszej pracy odnaleziono pewne reguły opisujące zmniejszone ryzyko wystąpienia epizodu omdlenia, bardziej czytelne niż te powstałe z losowych drzew, właśnie ze względu na interpretowalność dla klinicysty punkty podziału kategorii. Zauważono, iż kolejnym krokiem analizy mogłoby być włączenie do modelu nowej zmiennej, będącej wynikiem którejś z reguł, uznanej za szczególnie istotną np. z klinicznego punktu widzenia. Np. przetestowanie nowej zmiennej, powstałej dzięki wykrytej regule mówiącej, iż osoby mające częstotliwość HR częstoskurczu nie wyższą niż 220, u których nie występowały stany poprzedzające omdlenia, pokazało, iż tak zdefiniowana grupa ryzyka ma prawie sześciokrotnie mniejszą szansę na zaistnienie omdlenia.

Kolejną metodą prezentowania poszukiwanych zależności przyczynowo-skutkowych pomiędzy atrybutami warunkowymi i decyzyjnymi w postaci reguł decyzyjnych, a więc w bardzo intuicyjny sposób, jest wykorzystanie teorii zbiorów przybliżonych. Za bardzo pomocne w analizie konkretnego problemu medycznego uznano określanie minimalnych podzbiorów badanych atrybutów warunkowych takich, które pozwalają na określenie przynależności obiektów do klas decyzyjnych (z omdleniem lub bez omdleń) z zadawalającą jakością klasyfikacji, czyli tzw. zbiorów minimalnych. Dzięki takiej analizie można stworzyć zbiory najbardziej reprezentatywnych obiektów dla poszczególnych klas decyzyjnych.

Niestety, wyniki przeprowadzonej analizy zbioru zawierającego dane dotyczące grupy badawczej z wykorzystaniem teorii zbiorów przybliżonych nie były zadowalające. Jakość wygenerowanych reguł decyzyjnych potwierdziła istniejący kliniczny dylemat w zdefiniowaniu cech charakteryzujących osoby z omdleniami. We wcześniej publikowanych pracach klinicznych analizujących

omdlenie w grupach napadowego częstoskurczu nadkomorowego, także nie udało się wskazać pojedynczego objawu lub cechy klinicznej selekcyjnej najbardziej zagrożonych pacjentów. Co najwyżej, podkreślano współwystępowanie kołatania serca w grupie omdleń lub częstsze występowanie omdleń na przestrzeni kolejnych lat obserwacji tych pacjentów.

W prezentowanej pracy całkowita jakość klasyfikacji wynosiła 1, co oznaczało, iż wszystkie obiekty w systemie informacyjnym zostały prawidłowo rozpoznane. Jednocześnie wykazano, iż nie ma takich cech w systemie, których odjęcie spowodowałoby osłabienie jakości klasyfikacji (brak rdzenia). Dodatkowo dolne i górne przybliżenie klas definiujących omdlenia i brak omdleń, zawierały wszystkie obiekty danej klasy, natomiast liczba odnalezionych atomów była równa, bądź bliska liczbie wszystkich elementów w bazie. Dotyczyło to zarówno pierwotnej bazy danych, jak i jej wersji zdyskretyzowanej oraz sztucznie powiększonej poprzez nadlosowanie. Analizując reduktę stwierdzono, iż najmniejszym wykrytym zestawem atrybutów warunkowych prowadzących do klasyfikacji o takiej samej jakości, jaką daje klasyfikacja z uwzględnieniem wszystkich dostępnych w systemie informacyjnym predyktorów, jest np.: BMI, wiek pierwszego kołatania oraz częstotliwość HR częstoskurczu.

Dla algorytmów generujących reguły z wykorzystaniem teorii zbiorów przybliżonych, szczególną trudność sprawia fakt, iż analizowane dane charakteryzują się dużym niezbalansowaniem. Niestety, wszystkie klasyczne algorytmy zakładają równy rozkład w klasach i przez to dyskryminują klasę mniejszościową, koncentrując się wyłącznie na poprawnym rozpoznawaniu klasy większościowej.

W niniejszej pracy podjęta została próba rozwiązania problemu danych niezbalansowanych, poprzez zastosowanie jednego z dostępnych algorytmów nadlosowujących (ang. oversampling) obiektów klasy mniejszościowej – algorytmu SMOTE. (Można również stosować metody hybrydowe [94], czyli połączenie over- i undersamplingu, np. metoda SPIDER, jednak to wykracza poza zakres niniejszej pracy.) Zauważono, iż dodatkowe uzupełnienie braków danych powodowało obniżenie wykrywalności omdleń. Dużo wyższe wartości średniej dokładności odnotowano natomiast dla zbioru z dodanymi dodatkowymi przypadkami klasy mniejszościowej. Co prawda zestawy wygenerowanych reguł dla zbiorów po zastosowaniu algorytmu nadlosowującego (SMOTE) wykazywały znacznie większe pokrycie obiektów niż reguły wygenerowane dla pierwotnego zbioru danych, jednak ingerencja w przetwarzany zbiór danych, zarówno usuwanie danych tzw. nadmiarowych (z



klasy większościowej), jak i nadlosowywanie przypadków klasy mniejszościowej, wydaje się być ingerencją zbyt daleko zmieniającą naturalny charakter analizowanego zjawiska. Należy pamiętać, iż reguły, które zostały wygenerowane z danych sztucznie powiększonych, mogą prowadzić do niewłaściwych wniosków, a prognozowanie na podstawie takich sztucznie utworzonych przypadków, tj. wymyślonych pacjentów, może wiązać się ze zbyt dużym obciążeniem wyników.

Poszukując najlepszych rozwiązań radzenia sobie z danymi niezbalansowanymi, bez ingerencji w dane, przeprowadzono analizę danych EP za pomocą algorytmu BRACID [88]. Jest to oryginalna metoda uczenia reguł dla zbiorów niezbalansowanych, której innowacyjność polega na zintegrowaniu uczenia reguł i przypadków. Jednak wygenerowane, bardzo długie reguły, dla danych zawężonych po dyskretyzacji i wyborze reprezentantów, znów potwierdziły trudność w zdefiniowaniu cech charakteryzujących osoby z omdleniami. Niestety, reguły wygenerowane za pomocą algorytmu BRACID nie przyniosły znacząco lepszych rezultatów, w porównaniu z poprzednimi metodami (Rozdział 4.4) i potwierdziły wcześniejsze wnioski.

Należy zwrócić szczególną uwagę na fakt, iż standardowe miary oceny modeli ukierunkowane są na klasę większościową. Nie powinny więc być stosowane dla danych niezbalansowanych. [82] Z tego względu w pracy zastosowano specjalne miary oceny metod, ukierunkowanych na klasę mniejszościową, np.: F-miara, czy G-miara.

Porównanie modeli potwierdziło, iż wykrywalność omdleń wzrasta po zastosowaniu technik nadlosowania. Najlepsze wyniki osiągnęły modele punktacji utworzone za pomocą regresji logistycznej oraz reguły drzew decyzyjnych. Równie dobrze ocenione zostały reguły utworzone za pomocą algorytmu BRACID. Jednak ich mała przejrzystość i słaba możliwość interpretacji sprawiają, iż nie są dobrym rozwiązaniem analizowanego problemu medycznego.

W niniejszej pracy zastosowano techniki porównawczej oceny, polegające na stosowaniu poszczególnych metod dla analizowanej bazy danych, a następnie próby wybrania najlepszej z nich. Analizując dalsze, możliwe ścieżki działania w temacie wykrywania grupy ryzyka występowania epizodów omdleń w analizowanej populacji, wydaje się, iż następnym krokiem mógłby być model złożony [71]. Najważniejsze techniki łączenia modeli [97], stanowiące kluczową część tzw. predykcyjnego data mining, to m.in.:

a) **agregacja modeli** (głosowanie i uśrednianie) (ang. bagging)

Polega na łączeniu przewidywań wielu modeli tego samego typu uzyskanych dla różnych zbiorów uczących lub wielu modeli różnego typu uzyskanych dla tego samego zbioru danych. Jeśli modelujemy zmienną ciągłą (problemy regresyjne) to procedurę taką nazywamy uśrednianiem, a w przypadku zmiennych jakościowych (zagadnień klasyfikacyjnych) stosujemy termin głosowanie (voting). Proste głosowanie polega na wybieraniu klasy, którą wskazało najwięcej modeli. (W prezentowanej pracy idea baggingu wykorzystywana była w module „losowy las”)

b) **wzmacnianie** (nazywane też losowaniem adaptacyjnym i łączeniem modeli, ang. boosting)

Jest to głosowanie ważone lub uśrednianie. Technika ta polega na zbudowaniu sekwencji modeli dla danych i wyznaczeniu wag dla modelu zbiorczego. Pierwszy model budowany jest przy równych wagach wszystkich przypadków, a w kolejnych krokach wagi przypadków modyfikowane są tak, aby uzyskać dokładniejsze przewidywania dla tych przypadków, dla których wcześniejsze modele dawały błędne przewidywania. Większe wagi przypisujemy tym obserwacjom, które są trudne do klasyfikacji (które zostały błędnie zaklasyfikowane przez dotychczasowy model). Wzmacnianie powoduje utworzenie sekwencji modeli. Każdy kolejny model jest tzw. "ekspertem" w przewidywaniu dla przypadków, z którymi niedostatecznie radziły sobie poprzedzające go modele. [73] Przy stosowaniu modelu (do przewidywania wartości dla nowych przypadków) wyniki działania wszystkich modeli są łączone w celu uzyskania najlepszej prognozy.

c) **kontaminacja modeli** (ang. stacking, stacked generalizations) i metauczenie (ang. meta-learning).

Przy kontaminacji modeli wyniki działania poszczególnych modeli są danymi wejściowymi dla metauczenia. Wyniki drzew klasyfikacyjnych, czy modelu logistycznego, mogą np. stanowić dane wejściowe dla sieci neuronowej, która "nauczy się", jak połączyć wyniki różnych modeli tak, aby uzyskać najlepszy finalny model prognostyczny.

Łączenie różnych modeli wydaje się właściwym kierunkiem dalszych prac związanych z próbą najlepszego rozwiązania analizowanego problemu klinicznego. Dzięki agregacji modeli możemy uzyskiwać dokładniejsze i pewniejsze przewidywania dla skomplikowanych zależności [73], gdyż wyniki działania wszystkich modeli są łączone w celu uzyskania najlepszej prognozy.

Może to rozwiązać problem tzw. "wrodzonej" niestabilności wyników uzyskiwanych w przypadku stosowania skomplikowanej metody dla stosunkowo małego zbioru niezbalansowanych danych. Niestety, taki kierunek badań komplikuje znacznie proces badawczy i może prowadzić do spadku czytelności generowanych rozwiązań.

## 5.5 Podsumowanie

Analizowana baza pacjentów z wąskim zespołem QRS (AVNTR i AVRT) charakteryzuje się znacznym niezbalansowaniem danych. Jest to naturalna sytuacja wynikająca z faktu, iż omdlenia występujące w tej populacji zdarzają się rzadko [31] [32] (w analizowanej bazie: 20%), stanowią natomiast realne zagrożenie dla zdrowia i życia pacjenta z udokumentowaną arytmia [4] [5].

W niniejszej pracy wykazano, iż częstotliwość rytmu serca w częstoskurczu (dla chorych z częstoskurczem z wąskim zespołem QRS) nie determinuje grupy omdleń w sposób bezpośredni. Również indukowalność częstoskurczu nie była istotnie wyższa w grupie z omdleniami, bez względu na rodzaj arytmii.

Zastosowano różne metody klasyfikacyjne w celu odnalezienia czynników ryzyka występowania epizodów omdleń. Były to klasyczne metody regresji logistycznej oraz techniki eksploracji danych. Oczekiwano, iż metody eksploracji danych dedykowane dla danych niezrównoważonych okażą się dużo lepszym podejściem, służącym do rozwiązania analizowanego problemu. Niestety, nie stwierdzono wyraźnych różnic w zdolności prognostycznej omawianych typów modeli. Również wyniki testów dokładnych oraz symulacji MC, nie różniły się znacząco od wyników oszacowanych przez klasyczne testy asymptotyczne.

Nieintuicyjny mechanizm, związany z budowaniem rozwiązania problemu klasyfikacji przy wykorzystaniu technik eksploracyjnych, stawia pod znakiem zapytania możliwość zastosowania go na gruncie medycyny. Jednak tworzenie punktacji dla modelu scoringowego w niniejszej pracy wykonano z użyciem regresji logistycznej. Można więc wyjaśnić utworzoną punktację odpowiednim wpływem poszczególnych elementów modelu: wartości wymiarów jam serca: PMK, PK i LP oraz sodu ( $\text{Na}^+$ ), wieku i częstotliwości HR częstoskurczu standaryzowanego. (Przy czym istotne współczynniki regresji odnotowano jedynie dla PK i wieku pacjenta.) Przyjęty próg odcięcia analizowanego modelu punktacji, równy 539, wyznacza w prosty i czytelny sposób grupy ryzyka występowania epizodów omdleń. Osoby o punktacji poniżej progu 539 uznawane zostały za zagrożone omdleniem.

Występujące w danych medycznych niepełne informacje zgromadzone dla poszczególnych pacjentów mogą mieć różne przyczyny, natomiast najczęściej mają charakter naturalny. Zdecydowano się nie podążać w kierunku metod uzupełniania braków danych, czyli sztucznego zmieniania analizowanej próby badawczej. Jednym ze źródeł niepowodzeń w znalezieniu dobrego modelu może być, wspomniane już, znaczne niezbalansowanie analizowanych danych. Uznano, iż podobnie jak metody uzupełniania braków danych, metody przetwarzające analizowany zbiór danych (np. poprzez „nadlosowanie” klasy mniejszościowej) mogłyby spowodować faktyczną utratę informacji. W przypadku tak ważnego zagadnienia medycznego, jakim jest ocena czynników ryzyka występowania epizodów omdleń sercowych (sercowo-naczyniowych), skutki ewentualnej niepoprawnej klasyfikacji mogłyby być zbyt poważne.

Największą wadą technik eksploracji danych (technik typu data-mining) jest fakt, iż produktem takiej eksploracji danych są zazwyczaj ogromne zbiory ogólnych opisów. Dzięki zastosowaniu teorii zbiorów przybliżonych uzyskano wiedzę w postaci reguł oraz tzw. minimalne zestawy atrybutów warunkowych, które mogą prowadzić do klasyfikacji o takiej samej jakości, jaką można by uzyskać uwzględniając wszystkie dostępne zmienne. Jeden z takich zbiorów minimalnych to: BMI, wiek pierwszego kołatania oraz częstotliwość rytmu serca podczas częstoskurczu. Ponieważ spośród ogromnych zbiorów reguł należy wybrać tylko te, które najbardziej odpowiadają oczekiwaniom badacza, bardzo ważnym tematem jest wybór optymalnego sposobu wizualizacji odkrytej wiedzy. Dlatego właśnie w pracy wykorzystano najbardziej intuicyjną metodę technik wizualizacji uzyskanych informacji: drzewa decyzyjne. Uznano, iż ten sposób prezentowania wiedzy (ścieżek / algorytmów klasyfikacji) mógłby być realnym wsparciem diagnozy pacjentów zagrożonych omdleniem sercowym. Odnalezione dzięki drzewom decyzyjnym reguły mogą pomóc w definiowaniu czynników zmniejszających prawdopodobieństwo omdlenia: np.: niedowaga lub częstotliwość rytmu serca podczas częstoskurczu  $\leq 220$  i jednocześnie brak stanów przedomdleniowych.

Zastosowane metody statystyczne i eksploracji danych nie wyłoniły jednoznacznych czynników ryzyka występowania omdleń w analizowanej populacji, co potwierdza zawikłanie problemu klinicznej klasyfikacji pacjentów zagrożonych omdleniem sercowym. W niniejszej pracy zaprezentowano jednak możliwe czynniki zmniejszające prawdopodobieństwo omdlenia w analizowanej populacji oraz opisano liczne zmienne interakcyjne i wikłające. Zaprezentowane wyniki wzbogacają wiedzę dotyczącą omdleń w grupie chorych z wąskim zespołem QRS (AVNRT i AVRT).

---

## 6 Wnioski

---

1. W pracy, której celem była statystyczna ocena wpływu częstotliwości rytmu serca podczas częstoskurczu oraz innych potencjalnych czynników ryzyka na występowanie epizodów omdleń w grupie chorych z wąskim zespołem QRS (AVNRT i AVRT), nie wskazano jednoznacznie określonych czynników ryzyka występowania epizodów omdleń w analizowanej grupie chorych. Obliczone miary jakości klasyfikacji dla zastosowanych metod statystycznych i technik eksploracji danych najwyżej oceniły model punktacji. W pracy wykazano także wartość wstępnego przetwarzania danych. Najlepszymi technikami wizualizacji klasyfikacji omdleń okazały się: metoda drzew losowych i interakcyjnych. Stwierdzono, iż reguły decyzyjne uzyskane przy okazji budowy drzew oraz z zastosowaniem teorii zbiorów przybliżonych mogą być wykorzystane do lepszego opisanie przypadków ze zmniejszonym prawdopodobieństwem omdlenia.
2. Opisane zostały zmienne wnikające, które mogą osłabiać wpływ wieku na częstotliwość HR podczas częstoskurczu. Są to: rodzaj arytmii, płeć, współwystępowanie objawów dodatkowych, jawnych chorób, czasokres od momentu pojawienia się pierwszego częstoskurczu oraz działanie układu współczulnego i przywspółczulnego.
3. Analizując potencjalne czynniki ryzyka dla występowania epizodów omdleń w analizowanej populacji wykazano, iż przyspieszona częstotliwość HR podczas częstoskurczu nie warunkuje wystąpienia omdlenia w sposób bezpośredni. Jedynie po podaniu leku podczas badania EP, w przypadku pacjentów nieobarczonych dodatkową, jawną chorobą współwystępującą, wyższe wartości częstotliwości HR częstoskurczu mogły być uznane za czynnik ryzyka wystąpienia omdlenia. Stwierdzono, iż czynnikami ryzyka omdleń w badanej grupie chorych mogą być również: występowanie stanów przedomdleniowych, wymiary jam serca: LP i PK (wraz z ich istotną interakcją), obniżony poziom TSH u kobiet (zależność moderowana wiekiem), podwyższone BMI w przypadku mężczyzn oraz wyższe poziomy potasu dla całej populacji. Inne zmienne uwikłane, to: czasokres od pierwszego udokumentowanego częstoskurczu wraz z wiekiem oraz obecność jawnych objawów innej choroby współwystępującej.
4. Wyznaczono model punktacji wspierający rozpoznawanie grupy ryzyka omdlenia oraz utworzono zbiór reguł mogących identyfikować pacjentów o znacznie zmniejszonym prawdopodobieństwie wystąpienia omdleń sercowych w analizowanej populacji chorych z wąskim zespołem QRS (AVNRT i AVRT).

---

## 7 Streszczenie

---

W pracy dokonano porównania różnych metod analizy statystycznej i eksploracji danych (data mining), służących analizie przyczyn omdleń u pacjentów ze zdiagnozowanym częstoskurczem z wąskim zespołem QRS typu AVNRT i AVRT. Czynniki ryzyka wystąpienia epizodów omdleń poszukiwano wśród danych klinicznych oraz wyników badania elektrofizjologicznego serca (EP). Szczególną uwagę zwrócono na częstotliwość rytmu serca (HR) w trakcie częstoskurczu. Wartość ta staje się niezwykle ważna w czasie omdlenia o podłożu sercowym, związanym z podwyższonym zagrożeniem nagłym zgonem sercowym. Dotychczasowe badania nie rozstrzygają jednoznacznie związku HR częstoskurczu z występowaniem epizodów omdleń, dlatego niezbędne było sprawdzenie tego zagadnienia oraz utworzenie modelu charakteryzującego osoby z epizodami omdleń.

W niniejszej pracy wykazano, iż częstotliwość HR częstoskurczu nie determinuje pacjentów z omdleniami, w grupie chorych z częstoskurczem z wąskim zespołem QRS, w sposób bezpośredni. Jednocześnie stwierdzono brak jednoznacznie określonych czynników ryzyka występowania epizodów omdleń w analizowanej grupie. Dokonano oceny czynników powiązanych z częstotliwością HR częstoskurczu oraz hipotetycznych czynników ryzyka omdlenia w analizowanej grupie.

Wstępne charakterystyki dostępnych podgrup wykonano z wykorzystaniem testów parametrycznych i nieparametrycznych dla porównania dwóch i więcej grup. W celu zbadania bardziej złożonych efektów, wykorzystano techniki wielowymiarowe oraz metody eksploracji danych (data mining). Do budowy modeli wykorzystano m.in. regresję wieloraką i logistyczną, także z użyciem narzędzi scoringowych oraz metod analizy czynnikowej. Wykorzystano również obliczenia zwiększonej precyzji, teorię zbiorów przybliżonych oraz drzewa decyzyjne – m.in. typu C&RT/CHAID.

Wyniki, uzyskane za pomocą poszczególnych metod statystycznych i eksploracji danych, porównano i oceniono. Przedstawiono wady i zalety tychże rozwiązań ze szczególnym zwróceniem uwagi na ich zdolności w modelowaniu omawianych cech. Następnie opisano najlepsze modele i ścieżki decyzyjne ułatwiające klasyfikację osób z podwyższonym ryzykiem wystąpienia epizodów omdleń.

---

## 8 Abstract

---

The dissertation presents a comparison of different methods of statistical analysis and data mining for analysing the causes of syncope in patients diagnosed with tachycardia with narrow QRS complex (AVNRT and AVRT). Risk factors for episodes of syncope were sought among clinical data and the results of a cardiac electrophysiological study (EP). Particular attention was paid to the heart rate (HR) during tachycardia. This value becomes extremely important at the time of a cardiogenic syncope, because the heart rate (HR) is associated with an increased threat of sudden cardiac death. The results of previous studies are not clear due to the influence of tachycardia HR on the occurrence of syncopes, so it was necessary to identify the problem and create a model that describes persons with episodes of syncope.

In the present study it is demonstrated that HR tachycardia does not directly determine the occurrence of syncopes in patients with tachycardia with a narrow QRS complex. Also, there were no clearly defined risk factors of episodes of syncope in the study group. An evaluation of factors associated with the incidence of tachycardia and the risks of loss of consciousness in the analysed group was made.

Subgroup characterization was performed using comparative tests with two or more groups. Multivariate analysis and the technique of data mining were used to investigate the more complex effects. The models were built with the use of, among other methods, multiple regression and logistic regression, scoring tools, factor analysis, increased precision calculations, the theory of rough sets, and decision trees – including C&RT/CHAID algorithms.

The results, obtained using the different methods of statistical and data mining, were compared and evaluated. This paper presents the advantages and disadvantages of these solutions with a particular focus on their ability to model the discussed features. Next, the best models and decision paths to facilitate the classification of persons with an increased risk of episodes of syncope are described.

---

## Literatura

---

- [1] Grupa Robocza Europejskiego Towarzystwa Kardiologicznego (ESC) do spraw diagnostyki i postępowania w omdleniach, "Wytyczne Europejskiego Towarzystwa Kardiologicznego dotyczące diagnostyki i postępowania w omdleniach (wersja 2009)," *Kardiol. Pol.*, 2009.
- [2] M. Brignole, P. Alboni, D. Benditt, L. Bergfeldt, i wsp., "Guidelines on management (diagnosis and treatment) of syncope – Update 2004 The task force on Syncope, European Society of Cardiology," *Eur. Heart J.*, vol. 25, no. 22, s. 2054–2072, 2004.
- [3] A. Moya, R. Sutton, F. Ammirati, J.-J. Blanc, i wsp., "Guidelines for the diagnosis and management of syncope (version 2009) The Task Force for the Diagnosis and Management of Syncope of the European Society of Cardiology (ESC)," *Eur. Heart J.*, vol. 30, no. 21, s. 2631–2671, 2009.
- [4] S. C. Day, E. F. Cook, H. Funkenstein, L. Goldman, "Evaluation and outcome of emergency room patients with transient loss of consciousness.," *Am. J. Med.*, vol. 73, no. 1, s. 15–23, 1982.
- [5] M. D. Silverstein, D. E. Singer, A. G. Mulley, G. E. Thibault, i wsp., "Patients with syncope admitted to medical intensive care units," *JAMA J. Am. Med. Assoc.*, vol. 248, no. 10, s. 1185–1189, 1982.
- [6] D. D. Savage, L. Corwin, D. L. McGee, W. B. Kannel, i wsp., "Epidemiologic features of isolated syncope: the Framingham Study," *Stroke J. Cereb. Circ.*, vol. 16, no. 4, s. 626–629, 1985.
- [7] K. S. Ganzeboom, N. Colman, J. B. Reitsma, W. K. Shen, i wsp., "Prevalence and triggers of syncope in medical students.," *Am J Cardiol*, vol. 91, no. 8, s. 1006–1008, 2003.
- [8] A. Serletis, S. Rose, A. G. Sheldon, R. S. Sheldon, "Vasovagal syncope in medical students and their first-degree relatives," *Eur. Heart J.*, vol. 16, no. 27, s. 1965–1970, 2006.
- [9] E. S. Soteriades, J. C. Evans, M. G. Larson, M. H. Chen, i wsp., "Incidence and prognosis of syncope," *N. Engl. J. Med.*, vol. 347, no. 12, s. 878–885, 2002.
- [10] N. Colman, K. Nahm, K. S. Ganzeboom, W. K. Shen, i wsp., "Epidemiology of reflex syncope," *Clin. Auton. Res. Off. J. Clin. Auton. Res. Soc.*, vol. 14 Suppl 1, s. 9–17, 2004.
- [11] M. H. Ruwald, M. L. Hansen, M. Lamberts, C. M. Hansen, i wsp., "The relation between age, sex, comorbidity, and pharmacotherapy and the risk of syncope: a Danish nationwide study," *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.*, vol. 14, no. 10, s. 1506–1514, 2012.



- [12] F. Walczak and Ł. Szumowski, "Pozornie łagodne napady miarowego częstoskurczu z 'wąskim' zespołem QRS, których podłoże trzeba usunąć metodą ablacji.," *Forum Med. Rodz.*, vol. 1, no. 2, s. 165–176, 2007.
- [13] D. H. Spodick, "Normal sinus heart rate: sinus tachycardia and sinus bradycardia redefined," *Am. Heart J.*, vol. 124, no. 4, s. 1119–1121, 1992.
- [14] P. R. David H. Spodick, "Operational definition of normal sinus heart rate.," *Am. J. Cardiol.*, vol. 69, no. 14, s. 1245–6, 1992.
- [15] D. Esberger, "ABC of clinical electrocardiography: Junctional tachycardias," *BMJ*, vol. 324, no. 7338, s. 662–665, 2002.
- [16] M. A. McGuire, M. J. Janse, and D. L. Ross, "'AV nodal' reentry: Part II: AV nodal, AV junctional, or atrionodal reentry?," *J Cardiovasc Electrophysiol*, vol. 4(5), s. 573–86, 1993.
- [17] P. Denes, D. Wu, and i wsp, "Demonstration of dual A -V nodal pathways in patients with paroxysmal supraventricular tachycardia.," *Circulation*, s. 549 –555, 1973.
- [18] K. Małaczyńska-Rajpold, K. Błaszcyk, W. Seniuk, E. Koźluk, i wsp., "Częstoskurcz nawrotny w węźle przedsionkowo-komorowym.," *Pol. Przegląd Kardiologiczny*, vol. 14, no. 3, s. 196–203, 2012.
- [19] H. R. Middlekauff, W. G. Stevenson, L. W. Stevenson, L. A. Saxon, "Syncope in advanced heart failure: High risk of sudden death regardless of origin of syncope," *J. Am. Coll. Cardiol.*, vol. 21, no. 1, s. 110–116, 1993.
- [20] D. G. Kleinbaum, M. Klein, *Logistic Regression: A Self-Learning Text*. Springer, 2010.
- [21] J. Koronacki, J. Ćwik, *Statystyczne systemy uczące sie.*, WNT. Warszawa, 2005.
- [22] T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [23] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and Regression Trees*. Taylor & Francis, 1984.
- [24] A. H. A. W. G. on M. Segmentation, R. for C. Imaging, M. D. Cerqueira, N. J. Weissman, i wsp., "Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart: A Statement for Healthcare Professionals From the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association," *Circulation*, vol. 105, no. 4, s. 539–542, 2002.
- [25] R. M. Lang, M. Bierig, R. B. Devereux, F. A. Flachskampf, i wsp., "Recommendations for chamber quantification: a report from the American Society of Echocardiography's Guidelines and Standards Committee and the Chamber Quantification Writing Group, developed in conjunction with the European Association of Echocardiography, a branch of the European Society of Cardiology.," *J. Am. Soc. Echocardiogr. Off. Publ. Am. Soc. Echocardiogr.*, vol. 18, no. 12, s. 1440–1463, 2005.

- [26] H. C. Bazett, "An analysis of the time-relations of electrocardiograms.," *Heart*, no. 7, s. 35–70, 1920.
- [27] P. M. Rautaharju, B. Surawicz, and L. S. Gettes, "AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram Part IV: The ST Segment, T and U Waves, and the QT Interval A Scientific Statement From the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology," *J. Am. Coll. Cardiol.*, vol. 53, no. 11, s. 982–991, 2009.
- [28] A. J. Moss, "Measurement of the QT interval and the risk associated with QTc interval prolongation: a review.," *Am. J. Cardiol.*, vol. 72, no. 6, 1993.
- [29] M. M. Scheinman and S. Huang, "The 1998 NASPE Prospective Catheter Ablation Registry," *Pacing Clin. Electrophysiol.*, vol. 23, no. 6, s. 1020–1028, 2000.
- [30] P. Alboni, M. Brignole, C. Menozzi, A. Raviele, i wsp., "Diagnostic value of history in patients with syncope with or without heart disease," *J. Am. Coll. Cardiol.*, vol. 37, no. 7, s. 1921–1928, 2001.
- [31] K. A. Wood, B. J. Drew, M. M. Scheinman, "Frequency of Disabling Symptoms in Supraventricular Tachycardia," *Am. J. Cardiol.*, vol. 79, no. 2, s. 145–149, 1997.
- [32] B. Brembilla-Perrot, F. Chometon, L. Groben, C. Tatar, i wsp., "Are the results of electrophysiological study different in patients with a pre-excitation syndrome, with and without syncope?," *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.*, vol. 10, no. 2, s. 175–180, 2008.
- [33] F. Giada, I. Silvestri, A. Rossillo, P. G. Nicotera, i wsp., "Psychiatric profile, quality of life and risk of syncopal recurrence in patients with tilt-induced vasovagal syncope.," *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.*, vol. 7, no. 5, s. 465–471, 2005.
- [34] European Society of Cardiology, European Heart Rhythm Association, P. Vardas, A. Auricchio, J. Blanc, J. Daubert, i wsp., "Guidelines for cardiac pacing and cardiac resynchronization therapy.," *Eur Heart J*, no. 28, s. 2256–2295, 2007.
- [35] S. M. 3rd Fox, J. P. Naughton, and W. L. Haskell, "Physical activity and the prevention of coronary heart disease.," *Ann. Clin. Res.*, vol. 3, no. 6, s. 404–432, 1971.
- [36] M. H. Whaley, L. A. Kaminsky, G. B. Dwyer, L. H. Getchell, i wsp., "Predictors of over- and underachievement of age-predicted maximal heart rate.," *Med. Sci. Sports Exerc.*, vol. 24, no. 10, s. 1173–1179, 1992.

- [37] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *J. Am. Coll. Cardiol.*, vol. 37, no. 1, s. 153–156, 2001.
- [38] A. Fater-Dębska, P. Gworys, J. Brzeziński, and Z. Gawor, "Zaburzenia tyreometaboliczne a niewydolność serca.," *Endokrynol. Pol. Pol. J. Endocrinol.*, vol. T. 58 nr 3, p. s. 228–235, 2007.
- [39] A. S. Balcioglu, H. Müderrisoğlu, "Diabetes and cardiac autonomic neuropathy: Clinical manifestations, cardiovascular consequences, diagnosis and treatment," *World J Diabetes*, no. 6(1), s. 80–91, 2005.
- [40] E. Kozek, "Choroba wieńcowa u chorych na cukrzycę — odrębności kliniczne," *Diabetol. Prakt.*, vol. 3, no. 4, 2002.
- [41] S. Bakth, J. Arena, W. Lee, R. Torres, i wsp., "Arrhythmia susceptibility and myocardial composition in diabetes. Influence of physical conditioning.," *J. Clin. Invest.*, vol. 77, no. 2, s. 382–395, 1986.
- [42] A. Aidietis, A. Laucevicus, G. Marinskis, "Hypertension and cardiac arrhythmias.," *Curr. Pharm. Des.*, vol. 13, no. 25, s. 2545–2555, 2007.
- [43] F. Lombardi and P. Terranova, "Hypertension and concurrent arrhythmias.," *Curr. Pharm. Des.*, vol. 9, no. 21, s. 1703–1713, 2003.
- [44] C. J. Berger, J. M. Murabito, J. C. Evans, K. M. Anderson, i wsp., "Prognosis after first myocardial infarction. Comparison of Q-wave and non-Q-wave myocardial infarction in the Framingham Heart Study.," *JAMA*, vol. 268, no. 12, s. 1545–1551, 1992.
- [45] A. P. Maggioni, G. Zuanetti, M. G. Franzosi, F. Rovelli, i wsp., "Prevalence and prognostic significance of ventricular arrhythmias after acute myocardial infarction in the fibrinolytic era. GISSI-2 results.," *Circulation*, vol. 87, no. 2, s. 312–322, 1993.
- [46] C. J. Nalliah, S. Zaman, A. Narayan, J. Sullivan, i wsp., "Coronary artery reperfusion for ST elevation myocardial infarction is associated with shorter cycle length ventricular tachycardia and fewer spontaneous arrhythmias.," *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.*, vol. 16, no. 7, s. 1053–1060, 2014.
- [47] S. M. Al-Khatib, C. B. Granger, Y. Huang, K. L. Lee, i wsp., "Sustained ventricular arrhythmias among patients with acute coronary syndromes with no ST-segment elevation: incidence, predictors, and outcomes.," *Circulation*, vol. 106, no. 3, s. 309–312, 2002.
- [48] J. N. Cohn, P. R. Kowey, P. K. Whelton, L. M. Prisant, "New guidelines for potassium replacement in clinical practice: a contemporary review by the National Council on Potassium in Clinical Practice.," *Arch. Intern. Med.*, vol. 160, no. 16, s. 2429–2436, 2000.
- [49] G.-H. Kim, J. S. Han, "Therapeutic approach to hypokalemia.," *Nephron*, vol. 92 Suppl 1, 2002.

- [50] S. I. Gass, *Decision making, models and algorithms: a first course*. Wiley, 1985.
- [51] Y. Lu, L. Tian, J. Fang, *Advanced Medical Statistics*. World Scientific Publishing Company Incorporated, 2015.
- [52] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.
- [53] A. Petrie, C. Sabin, *Statystyka medyczna w zarysie*, 1st ed. Wydawnictwo Lekarskie PZWL, 2006.
- [54] R. A. Fisher, *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1925.
- [55] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, C. E. McCulloch, *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer, 2012.
- [56] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression*. Wiley, 2004.
- [57] M. H. Zweig, G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.," *Clin. Chem.*, vol. 39, no. 4, s. 561–77, 1993.
- [58] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.
- [59] J. A. Moczko, "Metody dokładnego wnioskowania nieparametrycznego w przypadku nietypowych rozkładów danych eksperymentalnych," *Przegląd Lek.*, vol. 64, no. 10, s. 874–878, 2007.
- [60] R. Zieliński, *Metody Monte Carlo*. Warszawa: WNT, 1970.
- [61] G. Migut, "Przykład wykorzystania modeli scoringowych w medycynie," [www.StatSoft.pl/czytelnia.html](http://www.StatSoft.pl/czytelnia.html), 2010.
- [62] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley, 2012.
- [63] B. Predki, R. Słowiński, J. Stefanowski, R. Susmaga, i wsp., "ROSE - Software Implementation of the Rough Set Theory," in *Rough Sets and Current Trends in Computing*, L. Polkowski and A. Skowron, Eds. Springer Berlin Heidelberg, 1998, s. 605–608.
- [64] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2005, s. 853–867.
- [65] D. E. Holmes, L. C. Jain, *Data Mining: Foundations and Intelligent Paradigms: Volume 3: Medical, Health, Social, Biological and other Applications*. Springer, 2012.
- [66] A. Stanis, *Przystępny kurs statystyki z wykorzystaniem programu STATISICA PL na przykładach z medycyny*, vol. II. Kraków, 2000.
- [67] A. Stanis, "Regresja krokowa," *Med. Prakt.*, no. 7, 2001.
- [68] A. Stanis, "Regresja logistyczna," *Med. Prakt.*, no. 10, 2001.

- [69] L. L. Thurstone, "The Measurement of Social Attitudes.," *J. Abnorm. Soc. Psychol.*, no. 27, s. 249–269, 1931.
- [70] StatSoft, "Internetowy Podręcznik Statystyki." [Online]. Available: <http://www.statsoft.pl/textbook/stathome.html>.
- [71] D. T. Larose, *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN. 2012.
- [72] P. Armitage, T. Colton, *Encyclopedia of Biostatistics*, 2nd Edition., 8-Volume Set vols. United Kingdom: John Wiley & Sons Australia, 2005.
- [73] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.
- [74] R. Tadeusiewicz, "Data mining jako szansa na relatywnie tanie dokonywanie odkryć naukowych przez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych.," *Statsoft Polska*, 2006.
- [75] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 2, s. 123–140, 1996.
- [76] J. A. Moczko, "Wykorzystanie teorii zbiorów przybliżonych do analizy danych jakościowych w badaniach toksykologicznych.," *Przegląd Lek.*, vol. 61, no. 10, s. 1189–1193, 2004.
- [77] Z. Pawlak, "Rough sets.," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, s. 341–356, 1982.
- [78] J. W. Grzymala-Busse, "A new version of the rule induction system LERS," *Fundam. Informaticae*, vol. 31, no. 1, s. 27–39, 1997.
- [79] "Rough Set Data Explorer User's Guide," Prosoft, 2002.
- [80] V. García, J. S. Sánchez, R. A. Mollineda, "Exploring the Performance of Resampling Strategies for the Class Imbalance Problem," in *Trends in Applied Intelligent Systems*, N. García-Pedrajas, F. Herrera, C. Fyfe, J. M. Benítez, and M. Ali, Eds. Springer Berlin Heidelberg, 2010, s. 541–549.
- [81] H. He, Member, E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. DATA ENG.*, vol. 21, no. 9, 2009.
- [82] M. Buckland, F. Gey, "The relationship between Recall and Precision," *J. Am. Soc. Inf. Sci.*, vol. 45, no. 1, s. 12–19, 1994.
- [83] C. J. "Keith" van Rijsbergen, *Information retrieval, Second edition*. Butterworths, 1979.
- [84] H. He and Y. Ma, *Imbalanced Learning : Foundations, Algorithms, and Applications*. Piscataway, NJ : IEEE Press ; Hoboken, New Jersey : Wiley, 2013.
- [85] J. R. Edwards, L. S. Lambert, "Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis.," *Psychol. Methods*, vol. 12, no. 1, p. 1, 2007.
- [86] R. Baron, D. Kenny, "The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations," *J. Pers. Soc. Psychol.*, vol. 51, no. 6, p. 1173—1182, 1986.

- [87] P. E. Shrout, N. Bolger, "Mediation in experimental and nonexperimental studies: new procedures and recommendations.," *Psychol. Methods*, vol. 7, no. 4, p. 422, 2002.
- [88] K. Napierała, J. Stefanowski, "BRACID: a comprehensive approach to learning rules from imbalanced data," *J. Intell. Inf. Syst.*, vol. 39, no. 2, s. 335–373, 2012.
- [89] J. W. Mason, D. J. Ramseth, D. O. Chanter, T. E. Moon, i wsp., "Electrocardiographic reference ranges derived from 79,743 ambulatory subjects," *J. Electrocardiol.*, vol. 40, no. 3, s. 228–234, 2007.
- [90] K. Błaszcyk, J. Marcinkowska, W. Seniuk, A. Gwizdała, i wsp., "Czy omdlenie koreluje z częstością częstoskurczu z wąskim zespołem QRS typu AVNRT i AVRT?," presented at the XXIV Konferencja Sekcji Rytmu Serca Polskiego Towarzystwa Kardiologicznego POLSTIM 2013, POLSTIM, Wrocław, 2013, vol. 8.
- [91] W. N. Kapoor, "Evaluation and outcome of patients with syncope," *Medicine (Baltimore)*, vol. 69, no. 3, s. 160–175, 1990.
- [92] F. P. Sarasin, M. Louis-Simonet, D. Carballo, S. Slama, A. Rajeswaran, J. T. Metzger, C. Lovis, P. F. Unger, A. F. Junod, "Prospective evaluation of patients with syncope: a population-based study," *Am. J. Med.*, vol. 111, no. 3, s. 177–184, 2001.
- [93] F. Croci, M. Brignole, P. Alboni, C. Menozzi, i wsp., "The application of a standardized strategy of evaluation in patients with syncope referred to three syncope units," *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.*, vol. 4, no. 4, s. 351–355, 2002.
- [94] K. Napierała, J. Stefanowski, S. Wilk, "Learning from Imbalanced Data in Presence of Noisy and Borderline Examples," in *Rough Sets and Current Trends in Computing*, M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen, and Q. Hu, Eds. Springer Berlin Heidelberg, 2010, s. 158–167.
- [95] M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen, i wsp., *Rough Sets and Current Trends in Computing: 7th International Conference, RSCTC 2010, Warsaw, Poland, June 28-30, 2010 Proceedings*. Springer Science & Business Media, 2010.
- [96] D. Firth, "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, vol. 80, no. 1, p. 27, 1993.
- [97] D. T. Larose, A. Wilbik, *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji danych*. Wydawnictwo Naukowe PWN, 2006.

---

# Załącznik nr 1

---

## Tabele dodatkowe

Tab. I Zgodność z rozkładem normalnym .....	160
Tab. II Charakterystyka wieku i HR częstoskurczu w podgrupach – cz.1. ....	162
Tab. III Charakterystyka wieku i HR częstoskurczu w wybranych podgrupach – cz.2. ....	164
Tab. IV Zmienne jakościowe dla omdleń.....	166
Tab. V Zmienne ilościowe dla omdleń.....	170
Tab. VI Porównania częstotliwości HR częstoskurczu standaryzowanego w podgrupach ze zdiagnozowaną i niezdiagnozowaną dodatkową chorobą współistniejącą .....	175
Tab. VII Porównania częstotliwości HR częstoskurczu podwyższonego podaniem leku w podgrupach ze zdiagnozowaną i niezdiagnozowaną dodatkową chorobą współistniejącą.....	176
Tab. VIII Korelacje.....	177
Tab. IX Korelacje w podgrupach cz.1.....	178
Tab. X Korelacje w podgrupach cz.2.....	179
Tab. XI Wartość należna HR częstoskurczu wobec wartości faktycznie zaobserwowanej (standardowe HR częstoskurczu).....	180
Tab. XII Porównanie obu arytmii: zmienne ilościowe .....	181
Tab. XIII Wartości własne czynników wyodrębnionych podczas wstępnego przetwarzania analizowanej bazy.....	182
Tab. XIV Wykresy wartości WoE dla wybranych zmiennych po dyskretyzacji	183
Tab. XV Charakterystyka czynników składowych głównych dla wybranych trzech grup zmiennych. ....	185
Tab. XVI Charakterystyka czynników składowych głównych dla grupy 18 wybranych zmiennych. ....	186
Tab. XVII Wykresy osypiska dla omawianych grup.....	187
Tab. XVIII Modele regresji wielorakiej.....	188
Tab. XIX Modele logistyczne.....	191
Tab. XX Miary oceniające przydatność modeli logistycznych do klasyfikacji (v- krotna walidacja) .....	197
Tab. XXI Wyniki analizy jednoczynnikowej – test ilorazu wiarygodności LR. .	199
Tab. XXII Interakcje wykryte testem LR .....	200
Tab. XXIII Reguły związane z drzewami wygenerowanymi w analizie losowy las (zmienne po dyskretyzacji).....	201
Tab. XXIV Reguły związane z drzewami wygenerowanymi w analizie losowy las (zmienne przed dyskretyzacją).....	203
Tab. XXV Tablica scoringowa – raport cech .....	204
Tab. XXVI Wybrane reguły uzyskane za pomocą algorytmu BRACID .....	206

**Tab. I Zgodność z rozkładem normalnym**

	N ważnych	średnia	Ufność -95%	Ufność 95%	Mediana	Min	Max	Odch.std	Skośność	Bł. std.	Kurtoza	Bł. std.	p testu Lilliefors'a
wiek (lata)	516	41,41	40,08	42,74	41,00	15,00	81,00	15,36	0,30	0,11	-0,82	0,21	p <0,01
waga (kg)	502	72,99	71,61	74,36	71,00	42,00	125,00	15,67	0,52	0,11	-0,24	0,22	p <0,01
wzrost (cm)	501	170,20	169,43	170,98	170,00	138,00	198,00	8,84	0,34	0,11	0,00	0,22	p <0,01
BMI (kg/m <sup>2</sup> )	501	25,11	24,70	25,53	24,90	16,61	43,75	4,70	0,74	0,11	0,90	0,22	p <0,05
HR z wywiadu	352	185,75	182,71	188,79	185,00	117,00	300,00	29,01	0,35	0,13	0,77	0,26	p <0,01
CL z wywiadu	214	333,74	326,70	340,78	324,50	200,00	500,00	52,23	0,62	0,17	0,44	0,33	p <0,01
TSH (mU/l)	269	1,64	1,49	1,80	1,38	0,00	14,80	1,29	4,48	0,15	<b>39,79</b>	0,30	p <0,01
FT4 (pmol/l)	107	6,74	5,21	8,28	1,32	0,55	40,98	8,01	1,36	0,23	1,99	0,46	p <0,01
FT3 (pmol/l)	59	3,31	2,91	3,71	2,96	0,75	9,26	1,52	1,28	0,31	<b>2,84</b>	0,61	p <0,01
Na+ (mmol/l)	488	141,54	141,27	141,82	142,00	131,00	150,00	3,08	-0,40	0,11	0,22	0,22	p <0,01
K+ (mmol/l)	488	4,26	4,23	4,28	4,23	3,50	5,30	0,30	0,33	0,11	0,04	0,22	p <0,01
LK (mm)	465	45,60	45,11	46,08	45,00	31,00	62,00	5,33	0,30	0,11	0,00	0,23	p <0,01
LP (mm)	465	32,12	31,66	32,59	32,00	17,00	50,00	5,14	0,32	0,11	0,20	0,23	p <0,01
PK (mm)	464	24,23	23,87	24,58	24,00	16,00	39,00	3,87	0,41	0,11	0,56	0,23	p <0,01
dolno-boczna ściana LK (mm)	465	9,47	9,34	9,60	9,00	4,00	15,00	1,43	0,30	0,11	0,64	0,23	p <0,01
PMK (mm)	465	9,63	9,47	9,80	9,50	5,00	23,00	1,82	1,28	0,11	<b>6,54</b>	0,23	p <0,01
Ao (mm)	465	28,68	28,28	29,08	28,00	19,00	44,00	4,43	0,46	0,11	0,12	0,23	p <0,01
EP: pre-ERP-Avna	269	242,83	237,44	248,21	240,00	160,00	430,00	44,88	1,15	0,15	1,59	0,30	p <0,01
EP: po-ERP-Avna	421	269,98	263,85	276,10	260,00	160,00	560,00	63,90	0,99	0,12	1,08	0,24	p <0,01
EP: pre-ERP-Avnr	258	247,74	239,29	256,18	220,00	146,00	550,00	68,91	1,80	0,15	<b>3,07</b>	0,30	p <0,01
EP: po-ERP-Avnr	360	264,91	255,80	274,01	220,00	20,00	640,00	87,81	1,55	0,13	<b>2,30</b>	0,26	p <0,01
EP: HR-Wenck-ante	190	180,45	175,77	185,13	178,50	107,00	260,00	32,70	0,14	0,18	-0,40	0,35	p >0,20
EP: CI-Wen-ante	190	343,23	333,85	352,62	333,50	230,00	560,00	65,59	0,77	0,18	0,45	0,35	p <0,01
EP: po-HR-Wen-ante	398	177,27	174,30	180,24	178,00	95,00	294,00	30,16	0,25	0,12	0,20	0,24	p <0,15



**Tab. I Zgodność z rozkładem normalnym c.d.**

	N ważnych	średnia	Ufność -95%	Ufność 95%	Mediana	Min	Max	Odch.std	Skośność	Bł. std.	Kurtoza	Bł. std.	p testu Lilliefors'a
EP: po-CI-Wen-ante	396	349,69	342,87	356,50	336,00	196,00	933,00	68,97	2,07	0,12	<b>12,85</b>	0,24	p <0,01
EP: HR-Wenck-retro	247	187,23	182,73	191,74	189,00	83,00	276,00	35,97	-0,20	0,15	-0,14	0,31	p <0,05
EP: CI-Wen-rete	247	333,24	323,63	342,85	318,00	210,00	716,00	76,66	1,73	0,15	<b>5,06</b>	0,31	p <0,01
EP: po-HR-Wenck-retro	335	182,38	178,42	186,33	188,00	86,00	267,00	36,77	-0,24	0,13	-0,55	0,27	p <0,01
EP: po-CI-Wen-rete	335	343,68	335,07	352,29	320,00	224,00	700,00	80,12	1,26	0,13	1,69	0,27	p <0,01
częstoskurcz standaryzowany	425	187,88	184,99	190,77	188,00	109,00	300,00	30,29	0,18	0,12	0,19	0,24	p <0,01
CL standaryzowane	392	350,68	345,04	356,32	341,50	194,00	560,00	56,83	0,82	0,12	1,10	0,25	p <0,01
HR z EP	432	182,67	179,73	185,62	180,00	107,00	277,00	31,17	0,13	0,12	-0,30	0,23	p <0,01
CL z EP	431	338,00	332,20	343,81	332,00	194,00	560,00	61,36	0,83	0,12	0,98	0,23	p <0,01
należna HR max	517	183,14	182,27	184,01	183,35	157,35	210,00	10,04	-0,27	0,11	-0,77	0,21	p <0,01
Czasokres objawów kołatania i częstoskurczu	509	14,98	13,85	16,11	11,00	0,00	66,00	12,98	1,25	0,11	1,39	0,22	p <0,01

\*Wysokie współczynniki kurtozy wyróżniono pogrubioną czcionką.

**Tab. II Charakterystyka wieku i HR częstoskurczu w podgrupach – cz.1.**

podgrupy		N	Średnia	Ufność - 95%	Ufność 95%	Mediana	Min	Max	Odch.std	p testu #	
Wiek	arytmia	AVNRT	288	44,302	42,526	46,078	45	17	81	15,3133	0,000001*
		WPW	228	37,750	35,838	39,662	36	15	79	14,6535	
	Choroby współistniejące I	NIE	232	34,961	33,217	36,705	32,5	16	81	13,4844	<0,000001*
		TAK	284	46,673	44,944	48,401	48	15	79	14,8003	
	Choroby współistniejące typu II	NIE	397	37,627	36,221	39,033	35	15	81	14,2491	<0,000001*
		TAK	119	54,017	51,870	56,164	53	19	79	11,8286	
	pleć	K	298	40,799	39,122	42,475	39	16	81	14,7049	0,397061
		M	218	42,239	40,075	44,402	41	15	79	16,2105	
	naprzemienność zespołu QRS	NIE	343	44,131	42,490	45,772	45	16	81	15,4486	<0,000001*
		TAK	87	34,759	32,008	37,509	32	17	68	12,9048	
	Czasokres objawów kołatania i częstoskurczu	krótszy	356	37,003	35,489	38,517	34	16	79	14,5237	<0,000001*
		dłuższy	153	52,007	50,187	53,826	51	28	81	11,3917	
	Działanie układu współczulnego	TAK	366	39,366	37,823	40,909	37	16	81	15,0124	<0,000001*
		NIE	88	49,341	46,328	52,354	51	20	79	14,2191	

**Tab. II Charakterystyka wieku i HR częstoskurczu w podgrupach – cz.1. c.d.**

podgrupy		N	Średnia	Ufność - 95%	Ufność 95%	Mediana	Min	Max	Odch.std	p testu #	
częstoskurcz standaryzowany	arytmia	AVNRT	269	188,089	184,444	191,735	189	120	300	30,3669	0,830719
		WPW	156	187,526	182,742	192,309	185	109	300	30,2458	
	Choroby współistniejące I	NIE	183	192,126	187,559	196,692	192	118	300	31,3085	0,010022*
		TAK	242	184,674	180,982	188,365	184	109	255	29,1494	
	Choroby współistniejące typu II	NIE	322	190,773	187,388	194,158	190	118	300	30,8735	0,000963*
		TAK	103	178,845	173,656	184,033	180	109	250	26,5481	
	płeć	K	256	189,754	186,077	193,431	189	109	300	29,8724	0,239445
		M	169	185,047	180,373	189,721	185	118	300	30,7791	
	naprzemienność zespołu QRS	NIE	321	182,769	179,645	185,894	180	109	300	28,4509	<0,000001*
		TAK	84	208,071	202,035	214,108	211	150	300	27,8173	
	Czasokres objawów kołatania i częstoskurczu	krótszy	285	189,130	185,567	192,693	190	118	300	30,5597	0,2936106
		dłuższy	139	185,345	180,351	190,340	185	109	258	29,7805	
	Działanie układu współczulnego	TAK	319	189,903	186,667	193,139	189	118	300	29,3754	0,04889154*
		NIE	79	182,823	175,181	190,464	180	120	300	34,1166	

# p testu Manna-Whitneya; zarówno wiek, jak i HR częstoskurczu w badanych podgrupach miały wynik testu Lilliefors'a na poziomie  $p < 0,05$ ;

\* wynik istotny statystycznie ( $p < 0,05$ )

**Tab. III Charakterystyka wieku i HR częstoskurczu w wybranych podgrupach – cz.2.**

Podgrupy szczegółowe			N	Średnia	Ufność - 95%	Ufność 95%	Mediana	Min	Max	Odch.std	p testu #	
Wiek	AVNRT	Choroby współistniejące typu I	TAK	169	50,355	48,403	52,307	50	20	78	12,8514	<0,000001*
			NIE	119	35,706	33,087	38,325	32	17	81	14,4295	
		Choroby współistniejące typu II	TAK	80	55,925	53,625	58,225	56	32	78	10,3334	<0,000001*
			NIE	208	39,832	37,841	41,823	37	17	81	14,56306	
		Działanie układu współczulnego	TAK	204	41,490	39,415	43,566	40	17	81	15,0341	0,000004*
			NIE	80	50,763	47,641	53,884	52	20	79	14,0282	
	WPW	Choroby współistniejące typu I	TAK	115	41,261	38,336	44,186	40	15	79	15,8347	0,000997*
			NIE	113	34,177	31,861	36,493	35	16	64	12,4278	
		Choroby współistniejące typu II	TAK	39	50,103	45,645	54,560	51	19	79	13,7512	<0,000001*
			NIE	189	35,201	33,261	37,141	34	15	79	13,5227	
		Działanie układu współczulnego	TAK	162	36,691	34,427	38,956	34	16	79	14,5957	0,891663
			NIE	8	35,125	29,982	40,268	36	21	42	6,1514	
	mężczyźni	Choroby współistniejące typu I	TAK	131	46,863	44,089	49,636	49	15	79	16,0436	<0,000001*
			NIE	87	35,276	32,320	38,232	33	16	79	13,8696	
Działanie układu współczulnego		TAK	156	40,128	37,610	42,647	39	17	79	15,9235	0,000143*	
		NIE	31	52,516	46,962	58,070	53	20	79	15,1413		
kobiety	Działanie układu współczulnego	TAK	210	38,800	36,853	40,747	36	16	81	14,3106	0,000057*	
		NIE	57	47,614	44,028	51,200	49	21	75	13,5155		

**Tab. III Charakterystyka wieku i HR częstoskurczu w wybranych podgrupach – cz.2. c.d.**

Podgrupy szczegółowe			N	Średnia	Ufność - 95%	Ufność 95%	Mediana	Min	Max	Odch.std	p testu #	
częstoskurcz standaryzowany	AVNRT	Choroby współistniejące typu I	TAK	158	185,298	180,735	189,860	185	120	250	29,0380	0,045524*
			NIE	111	192,063	186,067	198,059	200	120	300	31,8760	
		Choroby współistniejące typu II	TAK	75	176,027	169,848	182,206	175	120	250	26,85646	0,000051*
			NIE	194	192,753	188,445	197,061	198	120	300	30,42277	
		Działanie układu współczulnego	TAK	191	191,079	187,025	195,132	192	126	250	28,3974	0,027830*
			NIE	74	182,068	174,166	189,969	181	120	300	34,1055	
	WPW	Choroby współistniejące typu I	TAK	84	183,500	177,099	189,901	180	109	255	29,4968	0,096892
			NIE	72	192,222	185,024	199,421	189,5	118	300	30,6342	
		Choroby współistniejące typu II	TAK	28	186,393	176,858	195,928	186,5	109	221	24,5890	0,777717
			NIE	128	187,773	182,277	193,270	183	118	300	31,4263	
		Działanie układu współczulnego	TAK	128	188,148	182,760	193,537	185	118	300	30,8068	0,812631
			NIE	5	194,000	149,197	238,803	180	170	258	36,0832	
	mężczyźni	Choroby współistniejące typu I	TAK	105	182,781	177,356	188,206	185	120	255	28,0340	0,198947
			NIE	64	188,766	180,090	197,441	190	118	300	34,7320	
Działanie układu współczulnego		TAK	131	188,130	182,893	193,366	188	118	300	30,2941	0,031705*	
		NIE	27	172,778	160,338	185,218	180	120	222	31,4463		
kobiety	Działanie układu współczulnego	TAK	188	191,138	187,004	195,273	190	126	260	28,7347	0,324464	
		NIE	52	188,039	178,415	197,662	184	133	300	34,5668		

# p testu Manna-Whitneya; zarówno wiek, jak i HR częstoskurczu w badanych podgrupach miały wynik testu Lilliefors'a na poziomie  $p < 0,05$ ;

\* wynik istotny statystycznie ( $p < 0,05$ )

**Tab. IV Zmienne jakościowe dla omdleń**

potencjalne czynniki ryzyka	podgrupy	grupa z omdleniem (syncope T)	grupa bez omdleń (syncope N)	p testu Chi <sup>2</sup> lub Fishera-Freemana-Haltona				
		(liczność i %N)	(liczność i %N)	Cała grupa	Dla kobiet	Dla mężczyzn		
Rodzaj arytmii	całość	96	370	p=0,21851	p=0,08312	p=0,94928		
	AVNRT	52(54,17%)	226(61,08%)					
	WPW	44(45,83%)	144(38,92%)					
Płeć	całość	96	370	p=0,85979				
	K	57(59,38%)	216(58,38%)					
	M	39(40,63%)	154(41,62%)					
Pora dnia	całość	91	349	p=0,607637	p=0,64859	p=0,34401		
	Widoczne działania układu współczulnego	dzień	73(80,22%)				285(81,66%)	po zgrupowaniu: grupa „dzień” vs. pozostali: p=0,75304      p=0,59423      p=0,20568
	Brak jednoznacznego działania układu współczulnego	dzień/noc	15(16,48%)				44(12,61%)	
		Noc/dzień	2(2,20%)				16(4,58%)	
		noc	1(1,10%)				4(1,15%)	
Występowanie stanów poprzedzających omdlenia(T/N)	całość	84	343	p=0,00149*	p=0,00363*	p=0,13479		
	T	78(92,86%)	266(77,55%)	OR=3,75	OR=6,65			
	N	6(7,14%)	77(22,45%)	(1,56-10,9)	(1,56-28,44)			

**Tab. IV Zmienne jakościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	podgrupy	grupa z omdleniem (syncope T)	grupa bez omdleń (syncope N)	p testu Chi <sup>2</sup> lub Fishera-Freemana-Haltona		
		(liczność i %N)	(liczność i %N)	Cała grupa	Dla kobiet	Dla mężczyzn
Naprzemienność zespołu QRS	całość	85	328	p=0,29287	p=0,10536	p=0,77568
	NIE	71(83,53%)	257(78,35%)			
	TAK	14(16,47%)	71(21,65%)			
grupy 4 (wgprocentów należnej HR)	całość	96	370	p= 0,8742	p= 0,9043	p= 0,1745
	poniżej 80%	6(7,23%)	19(5,86%)			
	poniżej 100%	29(34,94%)	120(37,04%)			
	przekroczona do 20%	35(42,17%)	141(43,52%)			
	przekroczona ponad 20%	13(15,66%)	44(13,58%)			
kategorie BMI (4 grupy)	całość	96	370	p=0,32037	p=0,65051	p=0,16769
	niedowaga	2(2,08%)	22(5,95%)			
	OK	41(42,71%)	173(46,76%)			
	nadwaga	34(35,42%)	115(31,08%)			
	otyłość	19(19,79%)	60(16,22%)			

**Tab. IV Zmienne jakościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	podgrupy	grupa z omdleniem (syncope T)	grupa bez omdleń (syncope N)	p testu Chi <sup>2</sup> lub Fishera-Freemana-Haltona		
		(liczność i %N)	(liczność i %N)	Cała grupa	Dla kobiet	Dla mężczyzn
grupa CL z EP	całość	83	329	p=0,12663	p=0,55908	p=0,00059* OR=4,66 (1,83;11,84)
	poniżej (<=280,5)	19(22,89%)	52(15,81%)			
	powyżej (>280,5)	64(77,11%)	277(84,19%)			
grupa CL standaryzowana	całość	76	301	p=0,90724	p=0,56942	p=0,58373
	poniżej (<=280,5)	70(92,11%)	276(91,69%)			
	powyżej (>280,5)	6(7,89%)	25(8,31%)			
grupa CL z wywiadu	całość	38	173	p=0,76800	p=0,86018	p=0,25522
	poniżej (<=280,5)	33(86,84%)	147(84,97%)			
	powyżej (>280,5)	5(13,16%)	26(15,03%)			



**Tab. IV Zmienne jakościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	podgrupy	grupa z omdleniem (syncope T)	grupa bez omdleń (syncope N)	p testu Chi <sup>2</sup> lub Fishera-Freemana-Haltona		
		(liczność i %N)	(liczność i %N)	Cała grupa	Dla kobiet	Dla mężczyzn
Choroby współistniejące I	całość	96	370	p=0,19544	p=0,09958	p=0,95686
	NIE	35(36,46%)	162(43,78%)			
	TAK	61(63,54%)	208(56,22%)			
Choroby współistniejące II	całość	96	370	p=0,12628	p=0,15673	p=0,45387
	NIE	67(69,79%)	286(77,30%)			
	TAK	29(30,21%)	84(22,70%)			
HR>=170	całość	83	324	p=0,61341	p=0,52712	p=0,96096
	NIE	23(27,71%)	81(25,00%)			
	TAK	60(72,29%)	243(75,00%)			
Grupy wg czasokresu objawów kołatania i częstoskurczu	całość	96	370	p=0,54136	p=0,46879	p=0,06496
	Dłuższy(>=20lat)	27(28,13%)	116(31,35%)			
	Krótszy(<20lat)	69(71,88%)	254(68,65%)			
Indukowalność częstoskurczu	całość	96	370	p=0,50212	p=0,11599	p=0,59922
	Nie	13	41			
	Tak	83	329			

\* wynik istotny statystycznie (p<0,05)

**Tab. V Zmienne ilościowe dla omdleń**

potencjalne czynniki ryzyka	płeć	cała grupa			grupa z omdleniem			grupa bez omdleń			Test Manna-Whitneya
		średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	
wiek	całość	41,41(15,36)	41,00	516	43,29(14,86)	44,00	96	41,06(15,22)	39,00	370	0,199817
	K	40,80(14,70)	39,00	298	42,47(14,14)	44,00	57	40,74(14,62)	39,00	216	0,389635
	M	42,24(16,21)	41,00	218	44,49(15,96)	47,00	39	41,51(16,05)	41,00	154	0,333911
waga (kg)	całość	72,99(15,67)	71,00	502	74,59(17,15)	72,00	93	72,44(15,13)	70,00	361	0,408875
	K	66,26(14,11)	64,00	291	66,71(14,94)	65,00	56	65,86(13,47)	64,00	210	0,756532
	M	82,26(12,72)	80,00	211	86,53(12,95)	82,00	37	81,59(12,29)	81,00	151	0,057738
wzrost (cm)	całość	170,20(8,84)	170,00	501	169,61(9,37)	168,00	93	170,26(8,71)	170,00	360	0,615149
	K	165,29(6,39)	165,00	291	164,66(6,87)	164,00	56	165,29(5,85)	165,00	210	0,659658
	M	177,02(7,07)	176,00	210	177,11(7,52)	176,00	37	177,22(7,17)	177,00	150	0,838405
BMI (kg/m <sup>2</sup> )	całość	25,11(4,70)	24,90	501	25,79(4,88)	25,39	93	24,92(4,58)	24,59	360	0,145434
	K	24,28(5,07)	23,34	291	24,63(5,31)	23,53	56	24,13(4,93)	23,14	210	0,543842
	M	26,27(3,84)	26,04	210	27,55(3,52)	26,87	37	26,02(3,81)	25,95	150	0,038257*
Wywiad-udokumentowany (HR)	całość	185,75(29,01)	185,00	352	188,03(29,48)	187,50	66	185,30(28,35)	185,00	277	0,603104
	K	185,80(28,65)	184,00	219	186,07(29,91)	184,50	42	185,32(27,89)	181,00	171	0,907239
	M	185,66(29,70)	187,00	133	191,46(29,02)	189,00	24	185,26(29,22)	187,00	106	0,299059
TSH (mU/l)	całość	1,64(1,29)	1,38	269	1,66(2,03)	1,25	54	1,64(1,05)	1,39	206	0,20603
	K	1,64(1,45)	1,36	172	1,51(2,33)	1,11	37	1,69(1,12)	1,41	129	0,020533
	M	1,63(0,96)	1,48	97	1,96(1,16)	1,55	17	1,56(0,92)	1,37	77	0,193029
FT4 (pmol/l)	całość	6,74(8,01)	1,32	107	6,02(7,03)	1,24	17	6,84(8,22)	1,33	88	0,500166
	K	6,89(7,48)	1,38	80	6,14(7,26)	1,18	14	7,14(7,59)	1,39	65	0,365338
	M	6,32(9,56)	1,25	27	5,47(7,21)	1,74	3	6,00(9,93)	1,17	23	0,688109

**Tab. V Zmienne ilościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	płeć	cała grupa			grupa z omdleniem			grupa bez omdleń			Test Manna-Whitneya
		średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	
FT3 (pmol/l)	całość	3,31(1,52)	2,96	59	3,02(1,50)	2,77	9	3,33(1,52)	3,02	48	0,470167
	K	3,33(1,60)	3,01	47	3,17(1,54)	2,87	8	3,38(1,65)	3,03	38	0,684935
	M	3,23(1,23)	2,86	12	1,89(0,00)	1,89	1	3,10(0,95)	2,86	10	1
Na+ (mmol/l)	całość	141,54(3,08)	142,00	488	141,64(3,32)	142,00	91	141,59(2,99)	142,00	363	0,783381
	K	141,66(2,93)	142,00	285	142,00(3,12)	142,00	57	141,62(2,78)	142,00	211	0,359353
	M	141,38(3,27)	142,00	203	141,03(3,58)	141,00	34	141,55(3,26)	142,00	152	0,486195
K+ (mmol/l)	całość	4,26(0,30)	4,23	488	4,30(0,27)	4,30	91	4,25(0,30)	4,20	363	0,026134*
	K	4,24(0,29)	4,20	285	4,28(0,29)	4,30	57	4,24(0,29)	4,20	211	0,128444
	M	4,28(0,30)	4,30	203	4,34(0,24)	4,35	34	4,26(0,31)	4,27	152	0,099567
LK (mm)	całość	45,60(5,33)	45,00	465	46,18(5,01)	46,00	87	45,48(5,28)	45,00	357	0,188997
	K	43,53(4,56)	43,00	276	44,13(4,22)	44,00	53	43,54(4,55)	43,00	210	0,289109
	M	48,61(4,93)	48,00	189	49,38(4,48)	50,50	34	48,26(5,00)	48,00	147	0,136941
LP (mm)	całość	32,12(5,14)	32,00	465	33,10(5,26)	33,00	87	31,83(4,96)	32,00	357	0,045787*
	K	30,50(4,74)	30,00	276	31,64(4,77)	32,00	53	30,22(4,60)	30,00	210	0,039189*
	M	34,50(4,77)	35,00	189	35,38(5,25)	35,50	34	34,14(4,55)	34,00	147	0,19374
PK (mm)	całość	24,23(3,87)	24,00	464	24,92(4,00)	25,00	87	23,97(3,79)	24,00	356	0,041396*
	K	23,09(3,42)	23,00	275	23,96(3,66)	24,00	53	22,80(3,30)	23,00	209	0,043363*
	M	25,88(3,90)	26,00	189	26,41(4,09)	26,50	34	25,64(3,82)	25,00	147	0,220517
dolno-boczna ściana LK (mm)	całość	9,47(1,43)	9,00	465	9,61(1,52)	9,50	87	9,42(1,42)	9,00	357	0,388672
	K	9,08(1,38)	9,00	276	9,23(1,35)	9,00	53	9,00(1,38)	9,00	210	0,235872
	M	10,03(1,32)	10,00	189	10,21(1,61)	10,00	34	10,02(1,24)	10,00	147	0,880622
PMK (mm)	całość	9,63(1,82)	9,50	465	9,85(1,65)	10,00	87	9,59(1,88)	9,00	357	0,134004
	K	9,11(1,59)	9,00	276	9,55(1,50)	9,00	53	9,00(1,61)	9,00	210	0,010579*
	M	10,40(1,88)	10,00	189	10,31(1,79)	10,00	34	10,42(1,94)	10,00	147	0,741104

**Tab. V Zmienne ilościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	płeć	cała grupa			grupa z omdleniem			grupa bez omdleń			Test Manna-Whitneya
		średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	
Ao (mm)	całość	28,68(4,43)	28,00	465	29,21(4,65)	28,00	87	28,56(4,32)	28,00	357	0,359379
	K	27,10(3,51)	27,00	276	27,83(3,83)	27,00	53	26,95(3,41)	27,00	210	0,180334
	M	30,99(4,62)	31,00	189	31,35(5,04)	32,50	34	30,84(4,47)	31,00	147	0,688713
CL z wywiadu	całość	333,74(52,23)	325	214	332,21(53,13)	324	38	333,32(52,05)	324	173	0,835279
	K	329,76(50,53)	324	146	326,89(53,28)	324	27	330,06(50,40)	322	117	0,979483
	M	342,28(55,12)	329	68	345,27(52,90)	325	11	340,13(55,19)	329	56	0,615319
CL z EP	całość	338,00(61,36)	332	431	339,20(63,83)	337	83	337,26(58,89)	332	329	0,882723
	K	333,28(60,35)	328	260	343,51(59,82)	341	49	330,24(58,64)	326	200	0,129200
	M	345,19(62,36)	340	171	333,00(69,67)	329	34	348,14(57,84)	344	129	0,093976
CL standaryzowane	całość	350,68(56,83)	342	392	355,26(57,03)	343	76	348,20(55,31)	341	301	0,411073
	K	346,49(54,35)	340	239	354,17(48,80)	345	48	343,33(53,86)	339	182	0,164306
	M	357,22(60,09)	348	153	357,14(69,88)	341	28	355,66(56,88)	352	119	0,805121
EP: pre-ERP-Avna - wartości	całość	242,83(44,88)	240,00	269	240,98(46,96)	230,00	51	244,07(44,31)	240,00	209	0,479274
	K	233,47(36,91)	220,00	176	231,29(34,81)	220,00	31	234,58(37,41)	225,00	142	0,71773
	M	260,54(52,84)	260,00	93	256,00(59,15)	255,00	20	264,18(50,94)	260,00	67	0,298862
EP: po-ERP-Avna - wartości	całość	269,98(63,90)	260,00	421	273,93(58,73)	260,00	89	268,18(65,14)	260,00	313	0,212991
	K	264,14(58,17)	260,00	249	276,42(59,75)	270,00	53	260,92(57,70)	250,00	185	0,052345
	M	278,43(70,72)	260,00	172	270,28(57,84)	260,00	36	278,67(73,59)	260,00	128	0,787973
EP: pre-ERP-Avnr - wartości	całość	247,74(68,91)	220,00	258	250,43(72,05)	220,00	47	247,41(69,28)	220,00	202	0,851029
	K	243,86(65,30)	220,00	175	246,45(65,96)	240,00	31	244,11(66,07)	220,00	140	0,757657
	M	255,90(75,72)	220,00	83	258,13(84,40)	220,00	16	254,84(76,07)	220,00	62	0,850217
EP: po-ERP-Avnr - wartości	całość	264,91(87,81)	220,00	360	260,41(80,02)	220,00	74	264,06(89,36)	220,00	270	0,925997
	K	263,09(87,19)	220,00	222	262,39(81,71)	225,00	46	262,42(87,98)	220,00	168	0,619271
	M	267,83(89,05)	220,00	138	257,14(78,54)	220,00	28	266,76(91,96)	220,00	102	0,617746

**Tab. V Zmienne ilościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	płeć	cała grupa			grupa z omdleniem			grupa bez uomdleń			Test Manna-Whitneya
		średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	
EP: HR-Wenck-ante -wartości	całość	180,45(32,70)	178,50	190	179,88(29,56)	183,50	40	180,29(33,13)	178,00	143	0,987874
	K	183,24(32,49)	184,00	125	178,78(33,77)	180,00	27	184,06(32,12)	184,00	95	0,410216
	M	175,09(32,68)	174,00	65	182,15(18,97)	184,00	13	172,83(34,17)	167,00	48	0,110854
EP: CI-Wen-ante -wartości	całość	343,23(65,59)	333,50	190	342,50(58,32)	326,50	40	343,78(67,13)	336,00	143	0,935402
	K	337,85(65,81)	326,00	125	347,48(66,22)	333,00	27	335,99(66,03)	326,00	95	0,345298
	M	353,58(64,39)	348,00	65	332,15(37,17)	326,00	13	359,19(67,31)	360,00	48	0,125354
EP: po-HR-Wen-ante- wartości	całość	177,27(30,16)	178,00	398	176,88(32,08)	174,00	81	176,92(29,33)	178,00	300	0,625671
	K	176,37(27,12)	177,00	237	173,81(28,82)	174,00	48	176,64(26,52)	178,00	180	0,353093
	M	178,59(34,20)	179,00	161	181,33(36,31)	174,00	33	177,35(33,20)	179,00	120	0,734293
EP: po-CI-Wen-ante - wartości	całość	349,69(68,97)	336,00	396	346,73(64,16)	334,00	79	351,28(70,53)	340,00	298	0,855323
	K	343,86(61,65)	335,00	239	336,41(59,17)	327,00	49	346,89(62,72)	338,00	179	0,417738
	M	358,55(78,20)	340,00	157	363,60(69,30)	353,50	30	357,87(80,69)	340,00	119	0,434701
EP: HR-Wenck-retro - wartości	całość	187,23(35,97)	189,00	247	187,85(34,59)	191,00	47	187,53(35,78)	188,50	192	0,91468
	K	192,23(35,92)	195,00	167	188,06(37,12)	191,00	31	192,84(35,70)	200,00	133	0,372997
	M	176,81(33,99)	176,50	80	187,44(30,23)	190,50	16	175,54(33,25)	174,00	59	0,186991
EP: CI-Wen-rete - wartości	całość	333,24(76,66)	318,00	247	330,28(62,76)	316,00	47	332,32(76,22)	317,50	192	0,826665
	K	324,57(74,16)	307,00	167	331,23(66,02)	314,00	31	323,65(76,37)	300,00	133	0,347632
	M	351,33(79,07)	337,50	80	328,44(57,94)	318,00	16	351,86(72,79)	344,00	59	0,226496
EP: po-HR-Wenck-retro - wartości	całość	182,38(36,77)	188,00	335	182,30(35,10)	180,00	67	182,75(37,34)	190,00	253	0,604176
	K	183,63(37,77)	190,00	214	179,51(34,67)	178,00	43	184,65(38,70)	192,50	162	0,209855
	M	180,16(34,96)	184,00	121	187,29(36,06)	194,00	24	179,36(34,74)	184,00	91	0,387607
EP: po-CI-Wen-rete - wartości	całość	343,68(80,12)	320,00	335	342,01(75,46)	332,00	67	343,88(82,28)	316,00	253	0,638811
	K	342,66(83,48)	316,00	214	347,51(78,01)	336,00	43	341,69(85,71)	311,00	162	0,235103
	M	345,49(74,10)	330,00	121	332,17(71,20)	317,00	24	347,78(76,10)	330,00	91	0,385742

**Tab. V Zmienne ilościowe dla omdleń c.d.**

potencjalne czynniki ryzyka	płeć	cała grupa			grupa z omdleniem			grupa bez omdleń			Test Manna-Whitneya
		średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	średnia (sd)	mediana	liczność	
częstoskurcz standaryzowany	całość	187,88(30,29)	188,00	425	188,13(33,07)	188,00	83	188,35(29,04)	188,00	324	0,87389
	K	189,75(29,87)	189,00	256	188,27(31,25)	188,00	51	190,36(28,79)	190,00	194	0,472545
	M	185,05(30,78)	185,00	169	187,91(36,31)	187,50	32	185,35(29,27)	187,00	130	0,521629
HR częstoskurczu z badania EP	całość	182,67(31,17)	180,00	432	182,72(33,10)	178,00	83	182,36(29,58)	180,00	329	0,932584
	K	185,13(31,44)	183,00	261	179,67(31,81)	175,00	49	186,19(30,50)	184,00	200	0,145579
	M	178,92(30,46)	176,00	171	187,12(34,90)	183,00	34	176,44(27,16)	174,00	129	0,093959
należna HR max	całość	183,14(10,04)	183,35	517	181,86(9,66)	181,40	96	183,31(9,89)	184,65	370	0,199817
	K	183,48(9,56)	184,65	298	182,39(9,19)	181,40	57	183,52(9,50)	184,65	216	0,389635
	M	182,67(10,68)	183,35	219	181,08(10,38)	179,45	39	183,02(10,44)	183,35	154	0,333911
Czasokres częstoskurczu	całość	14,98(12,98)	11,00	509	14,77(12,70)	11,00	96	15,40(13,31)	11,00	370	0,717223
	K	15,30(12,29)	11,00	295	16,84(13,16)	12,00	57	15,15(12,30)	11,00	216	0,396180
	M	14,54(13,89)	10,00	214	11,74(11,51)	9,00	39	15,75(14,65)	11,00	154	0,166739

\* wynik istotny statystycznie (p<0,05)

**Tab. VI Porównania częstotliwości HR częstoskurczu standaryzowanego w podgrupach ze zdiagnozowaną i niezdiagnozowaną dodatkową chorobą współistniejącą**

Współistniejąca choroba (TAK) lub brak jakiegokolwiek dodatkowej choroby jawnej - norma (NIE)	AVNRT				WPW				Obie bazy			
	T	N	Bez względu na syncope	Porównanie grupy syncope i brak syncope (p testu)	T	N	Bez względu na syncope	Porównanie grupy syncope i brak syncope (p testu)	T	N	Bez względu na syncope	Porównanie grupy syncope i brak syncope (p testu)
<b>NIE</b>												
Liczność	15	93	111		14	51	72		29	144	183	
Średnia z częstoskurcz standaryzowany	192,80	192,76	192,06	p=0,679	182,36	195,14	192,22	p=0,236	187,76	193,60	192,13	p=0,259
<b>TAK</b>												
Liczność	33	119	158		21	61	84		54	180	242	
Średnia z częstoskurcz standaryzowany	189,88	184,29	185,30	p=0,401	185,90	183,87	183,50	p=0,753	188,33	184,14	184,67	p=0,374
<b>Bez podziału na chorobę</b>												
Liczność	<b>48</b>	<b>212</b>	<b>269</b>		<b>35</b>	<b>112</b>	<b>156</b>		<b>83</b>	<b>324</b>	<b>425</b>	
Średnia z częstoskurcz standaryzowany	<b>190,79</b>	<b>188,00</b>	<b>188,09</b>	p=0,845	<b>184,49</b>	<b>189,00</b>	<b>187,53</b>	p=0,595	<b>188,13</b>	<b>188,35</b>	<b>187,88</b>	p=0,874
<b>porównanie częstoskurczu zdrowych do tych ze współistniejącą chorobą (p testu)</b>			p=0,045*				p=0,096		p=0,837	p=0,002*	p=0,010*	

\* wynik istotny statystycznie (p<0,05)

**Tab. VII Porównania częstotliwości HR częstoskurczu podwyższonego podaniem leku w podgrupach ze zdiagnozowaną i niezdiagnozowaną dodatkową chorobą współistniejącą**

Współistniejąca choroba (TAK) lub brak jakiegokolwiek dodatkowej choroby jawnej - norma (NIE)	AVNRT				WPW				Obie bazy			
	T	N	Bez względu na syncope	Porównanie grupy syncope i brak syncope (p testu)	T	N	Bez względu na syncope	Porównanie grupy syncope i brak syncope (p testu)	T	N	Bez względu na syncope	Porównanie grupy syncope i brak syncope (p testu)
<b>NIE</b>												
Liczność	7	30	39		1		1		8	30	40	
Średni częstoskurcz -po Isopren (osoby, które miały lek przy EP)	232,57	204,17	211,85	p=0,011*	225,00		225,00	-	231,63	204,17	212,18	p=0,010*
<b>TAK</b>												
Liczność	13	45	61			5	5		13	50	66	
Średni częstoskurcz -po Isopren (osoby, które miały lek przy EP)	181,85	190,18	190,10	p=0,305		184,20	184,20	-	181,85	189,58	189,65	p=0,329
<b>Bez podziału na chorobę</b>												
Liczność	<b>20</b>	<b>75</b>	<b>100</b>		<b>1</b>	<b>5</b>	<b>6</b>		<b>21</b>	<b>80</b>	<b>106</b>	
Średni częstoskurcz -po Isopren (osoby, które miały lek przy EP)	<b>199,60</b>	<b>195,77</b>	<b>198,58</b>	p=0,759	<b>225,00</b>	<b>184,20</b>	<b>191,00</b>	-	<b>200,81</b>	<b>195,05</b>	<b>198,15</b>	p=0,555
<b>porównanie częstoskurczu zdrowych do tych ze współistniejącą chorobą (p testu)</b>			p=0,001*				-				p<0,001*	

\* wynik istotny statystycznie (p<0,05)



**Tab. VIII Korelacje**

	Wiek	BMI	HR z wywiadu	CL z wywiadu	TSH	Na+	K+	CL w EP	HR w EP	LK	LP	PK	d.-b. śc. LK	PMK	Ao	HR stand.	CL stand.	należna HR max
Wiek	<b>1</b>	0,42*	-0,34*	0,32*	-0,12	0,15*	0,17*	-0,25*	0,26*	0,23*	0,35*	0,26*	0,37*	0,36*	0,26*	-0,32*	0,33*	<b>-1*</b>
BMI	0,42*	<b>1</b>	-0,12*	0,11	0,09	0,06	0,09*	-0,09	0,09	0,48*	0,55*	0,43*	0,53*	0,48*	0,3*	-0,08	0,1*	-0,42*
HR z wywiadu	-0,34*	-0,12*	<b>1</b>	<b>-1*</b>	0,1	-0,11*	-0,05	0,53*	-0,52*	-0,04	-0,08	-0,15*	-0,1	-0,08	-0,09	<b>0,89*</b>	<b>-0,77*</b>	0,34*
CL z wywiadu	0,32*	0,11	<b>-1*</b>	<b>1</b>	-0,09	0,13	0,11	-0,45*	0,45*	0,06	0,15*	0,2*	0,18*	0,15*	0,09	<b>-0,87*</b>	<b>0,82*</b>	-0,32*
TSH	-0,12	0,09	0,1	-0,09		-0,17*	0	0,14*	-0,14*	-0,05	-0,03	-0,03	0,03	0	0	0,08	-0,13*	0,12
Na+	0,15*	0,06	-0,11*	0,13	-0,17*	<b>1</b>	0,15*	-0,08	0,08	0,02	0,01	0,04	0,13*	0,21*	-0,04	-0,06	0,14*	-0,15*
K+	0,17*	0,09*	-0,05	0,11	0	0,15*	<b>1</b>	-0,08	0,09	0,11*	0,09*	0,13*	0,18*	0,15*	0,08	-0,09	0,15*	-0,17*
CL w EP	-0,25*	-0,09	0,53*	-0,45*	0,14*	-0,08	-0,08	<b>1</b>	<b>-1*</b>	-0,19*	-0,05	-0,08	-0,11*	-0,11*	-0,14*	<b>0,73*</b>	<b>-0,83*</b>	0,26*
HR w EP	0,26*	0,09	-0,52*	0,45*	-0,14*	0,08	0,09	<b>-1*</b>	<b>1</b>	0,19*	0,06	0,09	0,12*	0,12*	0,15*	<b>-0,73*</b>	<b>0,84*</b>	-0,26*
LK	0,23*	0,48*	-0,04	0,06	-0,05	0,02	0,11*	-0,19*	0,19*	<b>1</b>	0,53*	0,37*	0,29*	0,28*	0,43*	-0,11*	0,18*	-0,23*
LP	0,35*	0,55*	-0,08	0,15*	-0,03	0,01	0,09*	-0,05	0,06	0,53*	<b>1</b>	0,52*	0,54*	0,5*	0,47*	-0,06	0,11*	-0,35*
PK	0,26*	0,43*	-0,15*	0,2*	-0,03*	0,04*	0,13*	-0,08*	0,09*	0,37*	0,52*	<b>1*</b>	0,38*	0,39*	0,38*	-0,13*	0,15*	-0,26*
d.-b.śc.LK	0,37*	0,53*	-0,1*	0,18*	0,03*	0,13*	0,18*	-0,11*	0,12*	0,29*	0,54*	0,38*	<b>1*</b>	<b>0,73*</b>	0,33*	-0,07*	0,11*	-0,37*
PMK	0,36*	0,48*	-0,08*	0,15*	0*	0,21*	0,15*	-0,11*	0,12*	0,28*	0,5*	0,39*	<b>0,73*</b>	<b>1*</b>	0,32*	-0,06*	0,12*	-0,36*
Ao	0,26*	0,3*	-0,09*	0,09*	0*	-0,04*	0,08*	-0,14*	0,15*	0,43*	0,47*	0,38*	0,33*	0,32*	<b>1*</b>	-0,14*	0,16*	-0,26*
HR stand.	-0,32*	-0,08*	<b>0,89*</b>	<b>-0,87*</b>	0,08*	-0,06*	-0,09*	<b>0,73*</b>	<b>-0,73*</b>	-0,11*	-0,06*	-0,13*	-0,07*	-0,06*	-0,14*	<b>1*</b>	<b>-0,82*</b>	0,32*
CL stand.	0,33*	0,1*	<b>-0,77*</b>	<b>0,82*</b>	-0,13*	0,14*	0,15*	<b>-0,83*</b>	<b>0,84*</b>	0,18*	0,11*	0,15*	0,11*	0,12*	0,16*	<b>-0,82*</b>	<b>1*</b>	-0,33*
należna HR max	<b>-1*</b>	-0,42*	0,34*	-0,32*	0,12*	-0,15*	-0,17*	0,26*	-0,26*	-0,23*	-0,35*	-0,26*	-0,37*	-0,36*	-0,26*	0,32*	-0,33*	<b>1*</b>

\* wynik istotny statystycznie (p<0,05) - istotne współczynniki korelacji Spearmana; „d.-b. śc.” –dolno-boczna ściana; pogrubienie czcionki – współczynniki korelacji Spearmana przekraczający wartość 0,7;

**Tab. IX Korelacje w podgrupach cz.1**

	Cały zbiór danych		dla K		dla M		dla AVNRT		dla WPW	
	HR stand.	Wiek	HR stand.	Wiek	HR stand.	Wiek	HR stand.	Wiek	HR stand.	Wiek
HR stand.	<b>1,00</b>	-0,32*	<b>1,00</b>	-0,28*	<b>1,00</b>	-0,37*	<b>1,00</b>	-0,37*	<b>1,00</b>	-0,25*
Wiek	-0,32*	<b>1,00</b>	-0,28*	<b>1,00</b>	-0,37*	<b>1,00</b>	-0,37*	<b>1,00</b>	-0,25*	<b>1,00</b>
kg-waga	-0,07	0,24*	-0,11	0,33*	0,04	0,17*	-0,13*	0,29*	0,02	0,25*
cm- wzrost	0,01	-0,23*	0,01	-0,38*	0,11	-0,34*	-0,08	-0,22*	0,16	-0,17*
BMI	-0,08	0,42*	-0,10	0,47*	-0,04	0,37*	-0,10	0,45*	-0,06	0,41*
pierwsze kołatanie (wiek) #	-0,27*	0,55*	-0,21*	0,56*	-0,35*	0,55*	-0,24*	0,53*	-0,34*	0,59*
TSH	0,08	-0,12	0,01	-0,04	0,23*	-0,22*	0,06	-0,14*	0,16	-0,03
Na+	-0,06	0,15*	-0,11	0,23*	0,02	0,06	-0,06	0,14*	-0,06	0,13
K+	-0,09	0,17*	-0,02	0,21*	-0,18*	0,09	-0,09	0,17*	-0,10	0,15*
LK	-0,11*	0,23*	-0,12	0,25*	-0,04	0,22*	-0,14*	0,26*	-0,05	0,24*
LP	-0,06	0,35*	-0,02	0,37*	-0,08	0,36*	-0,11	0,37*	0,04	0,39*
PK	-0,13*	0,26*	-0,12	0,28*	-0,10	0,21*	-0,16*	0,29*	-0,09	0,23*
d.-b. śc. LK	-0,07	0,37*	-0,04	0,37*	-0,05	0,40*	-0,12*	0,42*	0,03	0,35*
PMK	-0,06	0,36*	-0,09	0,41*	0,04	0,28*	-0,09	0,42*	-0,02	0,34*
Ao	-0,14*	0,26*	-0,10	0,29*	-0,18*	0,28*	-0,20*	0,31*	-0,03	0,26*
CL w EP	<b>-0,73*</b>	0,26*	<b>-0,72*</b>	0,26*	<b>-0,74*</b>	0,25*	-0,69*	0,27*	<b>-0,83*</b>	0,33*
Czasokres objawów kołatania i częstoskurczu #	-0,03	0,49*	-0,10	0,54*	0,06	0,45*	-0,09	0,46*	0,09	0,47*

\* wynik istotny statystycznie ( $p < 0,05$ ) - istotne współczynniki korelacji Spearmana; „d.-b. śc.” –dolno-boczna ściana;

pogrubienie czcionki – współczynniki korelacji Spearmana przekraczający wartość 0,7;

# zmienne nieuwzględnione w modelach ze względu na silną zależność z inną zmienną niezależną, wiekiem

**Tab. X Korelacje w podgrupach cz.2**

	AVNRT; kobiety		AVNRT; mężczyźni		WPW; kobiety		WPW; mężczyźni	
	HR stand.	Wiek	HR stand.	Wiek	HR stand.	Wiek	HR stand.	Wiek
HR stand.	<b>1,00</b>	-0,30*	<b>1,00</b>	-0,50*	<b>1,00</b>	-0,26*	<b>1,00</b>	-0,25*
Wiek	-0,30*	<b>1,00</b>	-0,50*	<b>1,00</b>	-0,26*	<b>1,00</b>	-0,25*	<b>1,00</b>
kg-waga	-0,13	0,29*	0,12	0,01*	-0,04	0,40*	-0,01*	0,19*
cm- wzrost	-0,05	-0,46*	0,12	-0,41*	0,24	-0,22*	0,11	-0,31*
BMI	-0,09	0,47*	0,01	0,29*	-0,09	0,48*	-0,08	0,36*
pierwsze kołatanie (wiek)	-0,20*	0,56*	-0,33*	0,47*	-0,26*	0,53*	-0,38*	0,64*
TSH	-0,01	-0,10	0,25	-0,19	0,17*	0,20*	0,19	-0,18*
Na+	-0,11	0,24*	0,04	-0,01*	-0,13	0,19	0,00	0,09*
K+	-0,06	0,21*	-0,10	0,05*	0,12*	0,18	-0,28	0,11*
LK	-0,13*	0,22*	-0,01	0,22*	-0,06	0,35*	-0,07*	0,16*
LP	-0,03	0,33*	-0,16	0,39*	0,01	0,45*	0,03	0,38*
PK	-0,18*	0,26*	0,04	0,18*	-0,01	0,28*	-0,21*	0,20*
dolno-boczna ściana LK	-0,06	0,35*	-0,10	0,50*	0,03	0,44*	0,01*	0,30*
PMK	-0,10	0,42*	0,09	0,30*	-0,03	0,40*	-0,03	0,31*
Ao	-0,14*	0,24*	-0,26	0,37*	0,03*	0,40*	-0,08*	0,18*
CL w EP	<b>-0,71*</b>	0,24*	<b>-0,61*</b>	0,27*	<b>-0,75*</b>	0,45*	<b>-0,88*</b>	0,26*
Czasokres objawów kołatania i częstoskurczu	-0,14*	0,49*	0,00	0,40*	-0,04	0,61*	0,18	0,37*

\* wynik istotny statystycznie (p<0,05) - istotne współczynniki korelacji Spearmana;  
pogrubienie czcionki – współczynniki korelacji Spearmana przekraczający wartość 0,7

**Tab. XI Wartość należna HR częstoskurczu wobec wartości faktycznie zaobserwowanej (standardowe HR częstoskurczu)**

		podgrupy (płeć; arytmia)	N ważnych	Średnia	Mediana	min	max	Odch.std	p testu kolejności par Wilcoxon <sup>#</sup>
częstoskurcz standaryzowany należna HR max	całość		425	187,88	188,00	109,00	300,00	30,288	0,00042*
			517	183,14	183,35	157,35	210,00	10,044	
częstoskurcz standaryzowany należna HR max	K	AVNRT	186	191,26	190,50	126,00	300,00	30,347	0,00036*
			200	182,37	182,05	157,35	198,95	9,266	
częstoskurcz standaryzowany należna HR max	M	AVNRT	83	180,98	185,00	120,00	240,00	29,361	0,23603
			89	178,91	177,50	158,65	210,00	11,400	
częstoskurcz standaryzowany należna HR max	K	WPW	70	185,74	181,00	109,00	260,00	28,394	0,91843
			98	185,75	186,60	158,65	199,60	9,789	
częstoskurcz standaryzowany należna HR max	M	WPW	86	188,98	186,50	118,00	300,00	31,765	0,23550
			130	185,25	186,60	158,65	200,25	9,353	
częstoskurcz standaryzowany należna HR max	K		256	189,75	189,00	109,00	300,00	29,872	0,00177*
			298	183,48	184,65	157,35	199,60	9,558	
częstoskurcz standaryzowany należna HR max	M		169	185,05	185,00	118,00	300,00	30,779	0,09691
			219	182,67	183,35	158,65	210,00	10,675	
częstoskurcz standaryzowany należna HR max	AVNRT		269	188,09	189,00	120,00	300,00	30,367	0,00023*
			289	181,30	180,75	157,35	210,00	10,080	
częstoskurcz standaryzowany należna HR max	WPW		156	187,53	185,00	109,00	300,00	30,246	0,33004
			228	185,46	186,60	158,65	200,25	9,525	

# zarówno HR częstoskurczu standaryzowanego, jak i należna wartość częstoskurczu w badanych podgrupach miały wynik testu Lilliefors'a na poziomie  $p < 0,05$ ;

\* wynik istotny statystycznie ( $p < 0,05$ )

**Tab. XII Porównanie obu arytmii: zmienne ilościowe**

	średnia dla AVNRT	średnia dla WPW	p-testu #	AVNRT (N)	WPW (N)
wiek	44,30	37,75	0,000001*	288	228
waga (kg)	71,76	74,55	0,021209*	281	221
wzrost (cm)	168,77	172,02	0,000015*	280	221
BMI (kg/m <sup>2</sup> )	25,14	25,09	0,678462	280	221
Wywiad-częstoskurcz (HR)	183,24	191,42	0,016824*	244	108
TSH (mU/l)	1,66	1,59	0,724051	194	75
FT4 (pmol/l)	7,65	3,77	0,047642*	82	25
FT3 (pmol/l)	3,52	2,74	0,072045	43	16
Na+ (mmol/l)	141,73	141,29	0,107164	282	206
K+ (mmol/l)	4,26	4,25	0,387857	282	206
LK (mm)	45,22	46,14	0,025037*	276	189
LP (mm)	31,70	32,75	0,010843*	276	189
PK (mm)	24,18	24,30	0,944153	276	188
dolno-boczna ściana LK (mm)	9,34	9,66	0,027232*	276	189
PMK (mm)	9,42	9,95	0,002133*	276	189
Ao (mm)	28,34	29,18	0,036570*	276	189
EP: pre-ERP-Avna	242,59	244,67	0,895337	239	30
EP: po-ERP-Avna	275,85	261,13	0,002488*	253	168
EP: pre-ERP-Avnr	247,07	262,73	0,686311	247	11
EP: po-ERP-Avnr	242,18	308,70	0,000000*	237	123
EP: HR-Wenck-ante	180,16	182,90	0,798093	170	20
EP: po-HR-Wen-ante	167,41	196,04	0,000000*	261	137
EP: HR-Wenck-retro	187,50	178,29	0,755524	240	7
EP: po-HR-Wenck-retro	191,76	161,82	0,000000*	230	105

**Tab. XII Porównanie obu arytmii: zmienne ilościowe c.d.**

	średnia dla AVNRT	średnia dla WPW	p-testu #	AVNRT (N)	WPW (N)
częstoskurcz standaryzowany	188,09	187,53	0,830719	269	156
należna HR max	181,30	185,46	0,000002*	289	228
czasokres częstoskurczu	17,09	12,22	0,000002*	288	221

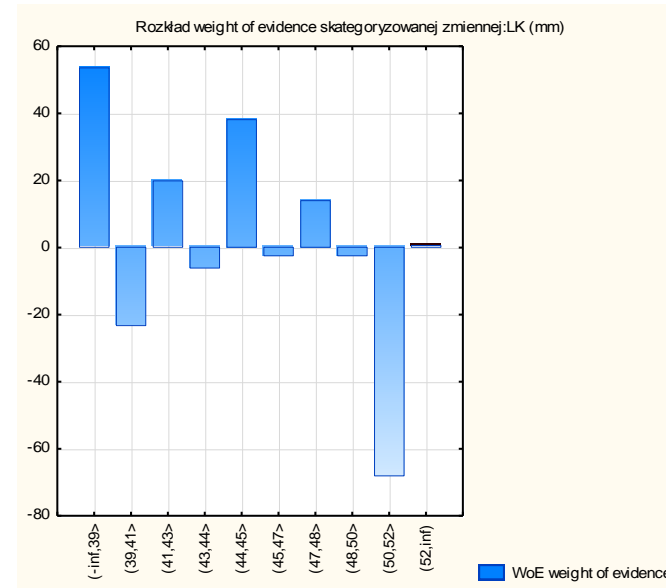
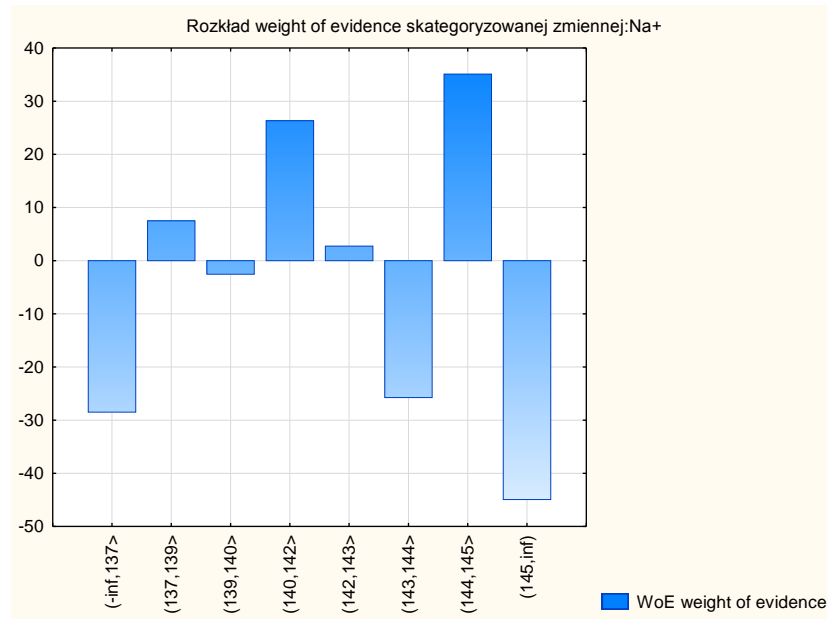
# Test porównujący dwie grupy: parametryczny t-Studenta lub Manna-Whitneya (w przypadku braku zgodności z rozkładem normalnym).

\* wynik istotny statystycznie ( $p < 0,05$ )

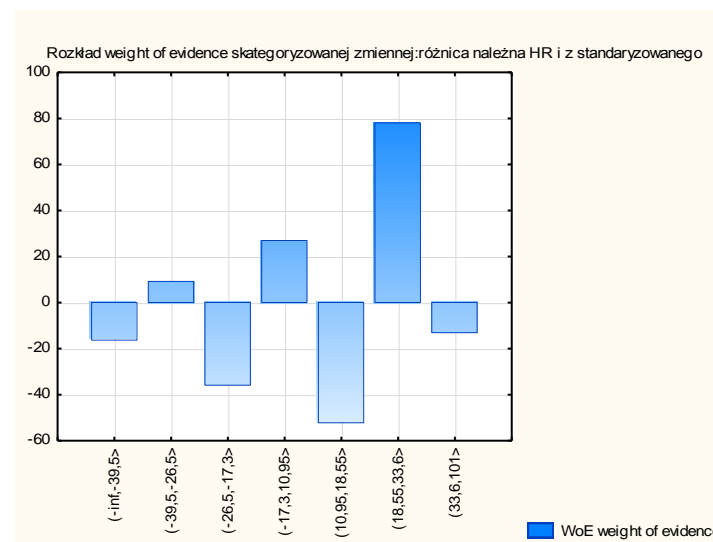
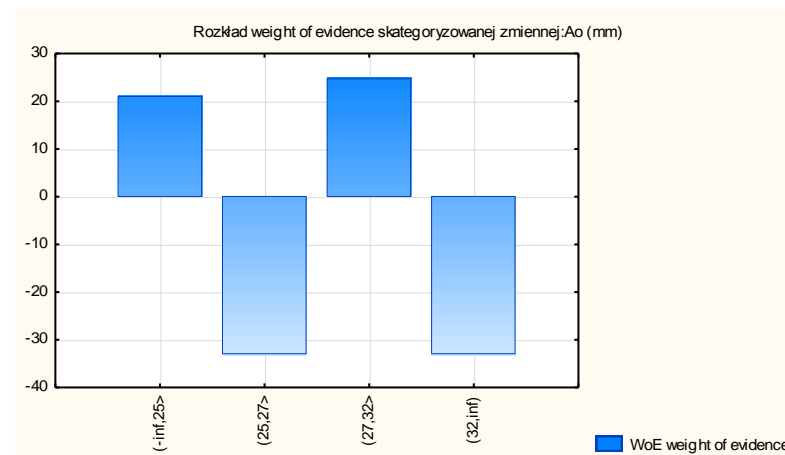
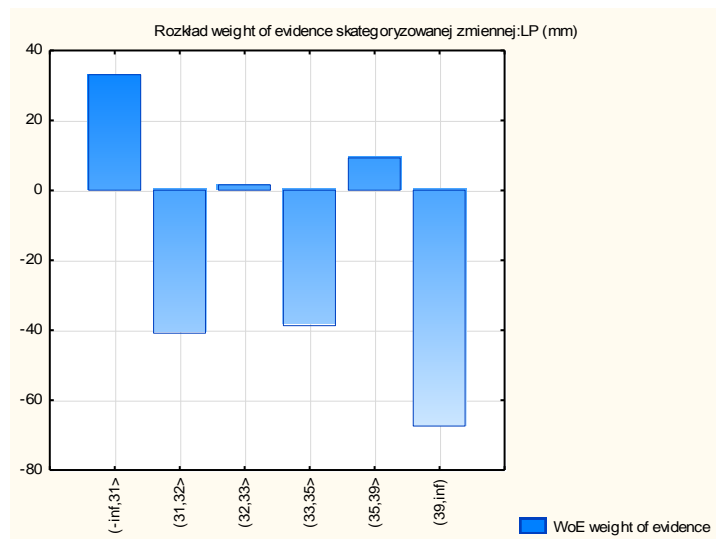
**Tab. XIII Wartości własne czynników wyodrębnionych podczas wstępnego przetwarzania analizowanej bazy**

Wyodrębnione czynniki	Wartość własna	% ogółu wariancji	skumulowany %
1	7,315014	18,75645	18,7564
2	6,261976	16,05635	34,8128
3	5,977702	15,32744	50,1402
4	5,448502	13,97052	64,1108
5	4,559659	11,69143	75,8022
6	2,764231	7,08777	82,8900
7	2,566820	6,58159	89,4715
8	2,364452	6,06270	95,5342
9	1,741645	4,46576	100,0000

**Tab. XIV Wykresy wartości WoE dla wybranych zmiennych po dyskretyzacji**



**Tab. XIV Wykresy wartości WoE dla wybranych zmiennych po dyskretyzacji c.d.**





**Tab. XV Charakterystyka czynników składowych głównych dla wybranych trzech grup zmiennych.**

Grupa	Zmienne w grupie	p testu Bartletta	%łącznie wytłumaczonej wariacji #	Współczynniki wartości czynnikowych z rotacją Varimax znormalizowaną			Ładunki czynnikowe z rotacją Varimax znormalizowaną		
				Czynnik 1	Czynnik 2	Czynnik 3	Czynnik 1	Czynnik 2	Czynnik 3
pierwsza	Wiek	<0.000001**	72%	0,366315	0,054835	0,374518	0,568593	0,165047	0,600449
	BMI			0,740965	-0,022365	-0,118350	0,879150*	<del>0,030466</del>	<del>0,060976</del>
	Na+			0,159320	0,790896	-0,261384	0,201668	0,847950*	-0,150050
	K+			-0,379084	0,508612	0,387397	-0,302596	0,594295	0,478667
	HR stadnd.			0,072802	0,190420	-0,679233	<del>0,089860</del>	<del>0,098338</del>	-0,821047*
druga	TSH	0,000041**	87%	0,005287	-0,881362		<del>0,001839</del>	-0,973578*	
	FT4			0,568434	0,253317		0,871381*	0,286235	
	FT3			0,575658	-0,253625		0,876714*	-0,273701	
trzecia	LK	<0.000001**	74%	0,465732	-0,045319		0,724499*	0,328131	
	LP			0,206738	0,307173		0,594588	0,581314	
	PK			-0,371574	0,935330		0,171952	0,935102*	
	Ao			0,706992	-0,363883		0,853681*	0,105317	

\* Ładunki czynnikowe >0,7; przekreślono bezwzględne wartości wag mniejszych od 0,15.

# dla wybranej, opisanej tu, liczby czynników

\*\* wynik istotny statystycznie (p<0,05)

**Tab. XVI Charakterystyka czynników składowych głównych dla grupy 18 wybranych zmiennych.**

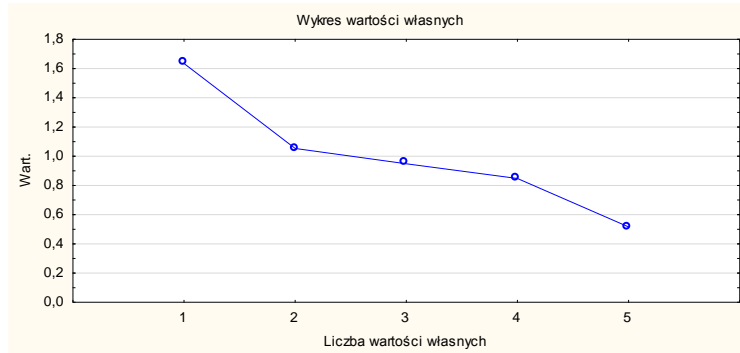
Zmienne w grupie #	Współczynniki wartości czynnikowych z rotacją Varimax znormalizowaną					Ładunki czynnikowe z rotacją Varimax znormalizowaną				
	Czynnik 1	Czynnik 2	Czynnik 3	Czynnik 4	Czynnik 5	Czynnik 1	Czynnik 2	Czynnik 3	Czynnik 4	Czynnik 5
Wiek	-0,009139	-0,016515	0,372436	-0,109168	0,003746	0,343921	0,208018	0,847945*	<del>-0,067465</del>	<del>-0,134841</del>
kg-waga	0,235636	-0,035862	-0,148772	0,073706	-0,054158	0,886477*	<del>0,091982</del>	<del>-0,082465</del>	0,088341	<del>-0,069992</del>
cm- wzrost	0,169801	0,043582	-0,327035	-0,250091	0,053738	0,435856	0,168729	-0,627205	-0,362950	0,139650
BMI	0,164097	-0,053933	0,008997	0,238286	-0,083264	0,731956*	<del>0,029086</del>	0,242915	0,318381	<del>-0,145570</del>
HR z wywiadu	0,044507	-0,271723	0,000684	-0,127419	0,005639	-0,171429	-0,840611*	-0,164445	-0,117248	0,052084
TSH	0,012291	0,045196	-0,065141	0,703421	0,022699	<del>0,090039</del>	<del>0,021386</del>	<del>0,006360</del>	0,855817*	0,029659
Na+	-0,032582	0,011493	0,024321	-0,225127	-0,514374	<del>0,027436</del>	0,150373	0,192836	-0,271277	-0,625830
K+	-0,003659	-0,030711	-0,069967	0,107146	-0,619512	<del>0,053384</del>	<del>-0,043527</del>	0,070950	0,133919	-0,721451*
LK	0,184002	0,006985	-0,223224	0,035315	-0,287349	0,689528	0,168827	-0,202963	0,021165	-0,317924
LP	0,192148	-0,033377	-0,003120	0,043895	0,120481	0,800218*	<del>0,112063</del>	<del>0,144576</del>	0,071082	0,097956
PK	0,158112	-0,041534	-0,021744	-0,198706	-0,097622	0,658247	<del>0,104670</del>	<del>0,099047</del>	-0,231464	<del>-0,145079</del>
d.-b. śc. LK	0,126337	-0,005493	0,113114	0,055470	0,099772	0,673228	0,202201	0,378692	<del>0,097742</del>	0,040107
PMK	0,131392	-0,040554	0,132005	-0,027155	0,100828	0,663974	<del>0,115102</del>	0,389727	<del>0,003763</del>	0,041007
Ao	0,131114	-0,022888	0,069691	-0,173647	0,284972	0,551620	<del>0,112459</del>	0,153860	-0,194465	0,288669
CL w EP	-0,050808	0,301108	-0,056354	-0,015922	0,025397	<del>0,099815</del>	0,905562*	<del>0,007029</del>	<del>-0,075618</del>	0,005653
HR stand.	0,054142	-0,285338	-0,009549	-0,124008	0,000187	-0,155169	-0,880159*	-0,184929	<del>-0,112156</del>	0,049441
HR w EP	0,044879	-0,293295	0,060180	0,092542	-0,011116	<del>-0,110662</del>	-0,898022*	<del>0,009008</del>	0,169665	0,009181
należna HR max	0,009139	0,016515	-0,372436	0,109168	-0,003746	-0,343921	-0,208018	-0,847945*	0,067465	0,134841

\*Ładunki czynnikowe >0,7; przekreślono bezwzględne wartości wag mniejszych od 0,15. ; „d.-b. śc” –dolno-boczna ściana

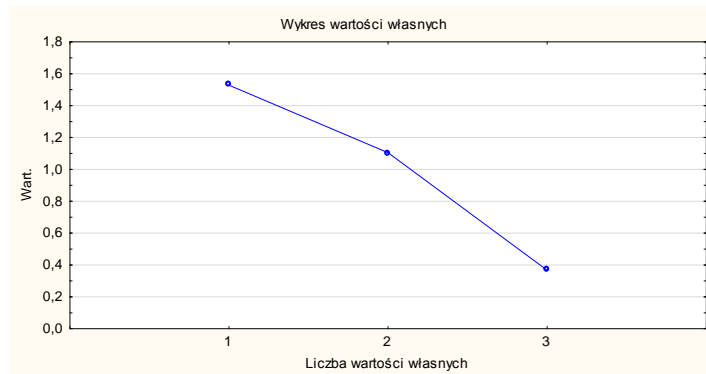
# %łącznie wytłumaczonej wariancji (przez wybrane 5 czynników) wynosił 71%; p testu Bartletta <0,000001 świadczy o istotnym skorelowaniu wybranych 18 zmiennych.

## Tab. XVII Wykresy osypiska dla omawianych grup

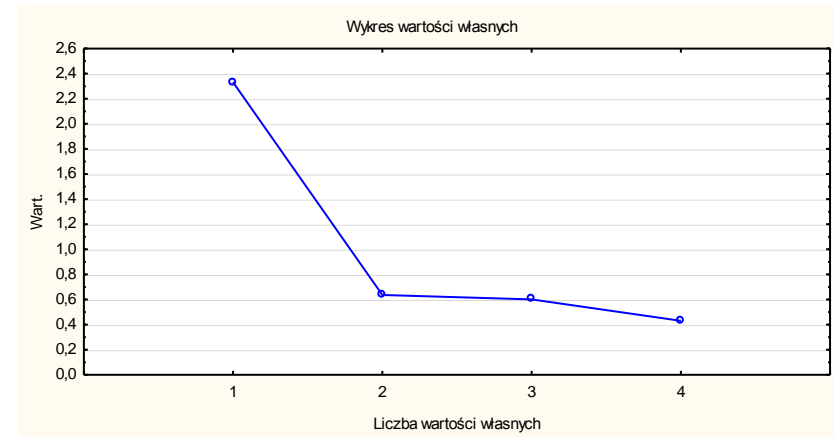
Grupa pierwsza: Wiek, BMI, Na+, K+ oraz częstotliwość HR częstoskurczu standaryzowanego



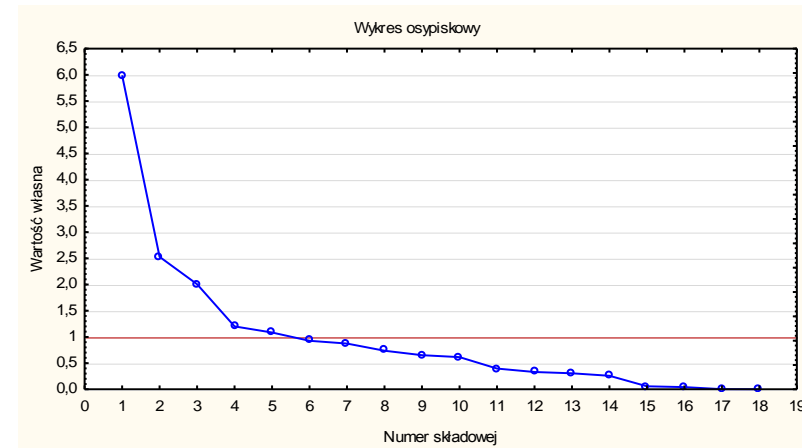
Grupa druga: TSH, FT3 i FT4



Trzecia grupa: LK, PK, LP oraz Ao



Grupa 18 zmiennych wybranych po wstępnej analizie struktury korelacyjnej



Tab. XVIII Modele regresji wielorakiej

Model	F testu analizy wariacji	SE (miara dopasowania modelu)	wartość p modelu (test F)	N (liczebność)	Normalność reszt (p testu Lillieforsa)	R (współczynnik korelacji wielorakiej)	R2 skorygowane	zmienne niezależne	współczynniki modelu $\beta$	SE $_{\beta}$	wartość p dla zmiennej (test t-Studenta)
A1	49,184	28,701	<0,0000001*	425	0,389	0,32	10,2%	wiek	-0,323	0,046	<0,0000001*
A2	21,055	26,318	0,000016*	83	0,250	0,45	19,7%	wiek	-0,454	0,099	0,000016*
B	25,358	28,685	<0,0000001*	425	0,249	0,33	10,3%	wiek	-0,319	0,046	<0,0000001*
								pleć	0,056	0,046	0,225028
C	25,790	28,659	<0,0000001*	425	0,420	0,33	10,5%	wiek	-0,336	0,047	<0,0000001*
								arytmia	-0,070	0,047	0,134322
E1	25,808	28,323	<0,0000001*	414	0,410	0,33	10,7%	wiek	-0,334	0,047	<0,0000001*
								Na+	-0,004	0,047	0,934784
E2	10,531	26,510	0,000088*	82	439,000	0,46	19,0%	wiek	-0,447	0,100	0,000026*
								Na+	0,077	0,100	0,444891
F1	25,923	28,316	<0,0000001*	414	0,311	0,33	10,8%	wiek	-0,330	0,047	<0,0000001*
								K+	-0,022	0,047	0,645888
F2	10,395	26,546	0,000098*	82	0,210	0,46	18,8%	wiek	-0,448	0,100	0,000026*
								K+	-0,061	0,100	0,542896
G1	26,292	28,065	<0,0000001*	403	0,525	0,34	11,2%	wiek	-0,305	0,049	<0,0000001*
								Ao	-0,087	0,049	0,075758
G2	9,221	26,050	0,000263*	78	0,131	0,44	17,6%	wiek	-0,378	0,111	0,001094*
								Ao	-0,134	0,111	0,233409
H1	24,655	28,167	<0,0000001*	403	0,618	0,33	10,5%	wiek	-0,340	0,051	<0,0000001*
								LP	0,025	0,051	0,615402

Tab. XVIII Modele regresji wielorakiej c.d.

Model	F testu analizy wariacji	SE (miara dopasowania modelu)	wartość p modelu (test F)	N (liczebność)	Normalność reszt (p testu Lillieforsa)	R (współczynnik korelacji wielorakiej)	R2 skorygowane	zmienne niezależne	współczynniki modelu $\beta$	SE $\beta$	wartość p dla zmiennej (test t-Studenta)
H2	8,485	26,257	0,000476*	78	0,170	0,43	16,3%	wiek	-0,405	0,113	0,000603*
								LP	-0,055	0,113	0,625775
I1	25,503	28,114	<0,0000001*	403	0,208	0,34	10,9%	wiek	-0,315	0,049	<0,0000001*
								LK	-0,064	0,049	0,184849
I2	8,733	26,187	0,000389*	78	0,075	0,43	16,7%	wiek	-0,444	0,106	0,000078*
								LK	0,085	0,106	0,424330
J1	24,986	28,146	<0,0000001*	403	0,513	0,33	10,7%	wiek	-0,318	0,049	<0,0000001*
								PK	-0,045	0,049	0,359264
J2	9,371	26,008	0,000233*	78	0,345	0,45	17,9%	wiek	-0,459	0,106	0,000047*
								PK	0,138	0,106	0,197815
K1	25,070	28,141	<0,0000001*	403	0,553	0,33	10,7%	wiek	-0,348	0,050	<0,0000001*
								tyl LK	0,050	0,050	0,319628
K2	10,182	25,786	0,000122*	78	0,188	0,46	19,3%	wiek	-0,516	0,115	0,000024*
								tyl LK	0,199	0,115	0,086604
L1	25,628	28,106	<0,0000001*	403	0,364	0,34	10,9%	wiek	-0,356	0,050	<0,0000001*
								PMK	0,071	0,050	0,159521
L2	10,738	25,637	0,000079*	78	0,306	0,47	20,2%	wiek	-0,488	0,107	0,000018*
								PMK	0,211	0,107	0,051248
Ł1	28,686	28,350	<0,0000001*	413	0,392	0,35	11,8%	wiek	-0,364	0,050	<0,0000001*
								BMI	0,043	0,050	0,386817

Tab. XVIII Modele regresji wielorakiej c.d.

Model	F testu analizy wariacji	SE (miara dopasowania modelu)	wartość p modelu (test F)	N (liczebność)	Normalność reszt (p testu Lillieforsa)	R (współczynnik korelacji wielorakiej)	R2 skorygowane	zmienne niezależne	współczynniki modelu $\beta$	SE $_{\beta}$	wartość p dla zmiennej (test t-Studenta)
Ł2	13,976	25,681	0,000007*	80	0,342	0,52	24,7%	wiek	-0,542	0,103	0,000001*
								BMI	0,188	0,103	0,070519
M1	10,475	29,783	0,000043*	246	0,112	0,28	7,2%	wiek	-0,281	0,062	<0,0000001*
								TSH	-0,018	0,062	0,764448
M2	5,900	28,043	0,004861*	56	0,021*	0,43	15,1%	wiek	-0,400	0,126	0,002508*
								TSH	0,097	0,126	0,442474
N1	5,103	29,330	0,007834*	99	0,428	0,31	7,7%	wiek	-0,292	0,097	0,003321*
								FT4	0,104	0,097	0,284328
N2	1,060	19,466	0,376662	15	0,996	0,39	0,9%	wiek	-0,393	0,273	0,175472
								FT4	0,146	0,273	0,603515
O1	1,663	29,022	0,199899	53	0,582	0,25	2,5%	wiek	-0,237	0,137	0,089988
								FT3	0,064	0,137	0,644645
O2	1,164	19,498	0,462149	5	0,282	0,73	7,6%	wiek	-0,225	0,519	0,706821
								FT3	0,619	0,519	0,355163

Modele z oznaczeniem „1” to modele dla całej populacji, natomiast „2” oznacza populację mężczyzn z AVNRT. Porównań odpowiednich modeli pełnych z modelami po usunięciu dodatkowej zmiennej (czyli zawierających jedną zmienną niezależną – wiek) dokonano testem F.

\* wynik istotny statystycznie (p<0,05)

**Tab. XIX Modele logistyczne**

Model logistyczny	p testu ilorazu wiarygodności	N	jakość modelu: p testu Hosmera - Lemeshowa	R2 (Nagelkerke)	R2 (Coxa-Snella)	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)	porównanie z modelem zredukowanym <sup>a</sup>
<b>A</b>	0,12785	407	0,74030	1,580%	1,006%	częstoskurcz standaryzowany wiek	1,0025 1,0171	0,57150 0,04375*	
<b>dla krótszego czasokresu HR (&lt;20lat)</b>	0,00888*	275	0,20307	5,280%	3,377%	częstoskurcz standaryzowany wiek	1,0046 1,0320	0,38008 0,00225*	
<b>dla dłuższego czasokresu HR (&gt;=20lat)</b>	0,60263	132	0,15626	1,215%	0,764%	częstoskurcz standaryzowany wiek	0,9940 0,9786	0,4903 0,3353	
<b>A1</b>	0,12232	407	0,33798	2,219%	1,412%	częstoskurcz standaryzowany wiek krótki czasokres częstoskurczu (<20lat)	1,0029 1,0225 1,4592	0,5073 0,0166* 0,1994	0,1955

**Tab. XIX Modele logistyczne c.d.**

Model logistyczny	p testu ilorazu wiarygodności	N	jakość modelu: p testu Hosmera - Lemeshowa	R2 (Nagelkerke)	R2 (Coxa-Snella)	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)	porównanie z modelem zredukowanym <sup>a</sup>
<b>B</b>	0,44239	407	0,55811	0,629%	0,400%	częstoskurcz standaryzowany	0,9998	0,96273	
						rodzaj arytmii	1,3801	0,19957	
<b>C</b>	0,96370	407	0,82118	0,029%	0,018%	częstoskurcz standaryzowany	0,9997	0,93898	
						płeć	1,0694	0,79094	
<b>D</b>	0,28672	407	0,54048	0,962%	0,612%	częstoskurcz standaryzowany	1,0006	0,89055	
						choroba typ I	1,4962	0,11825	
<b>E</b>	0,39383	407	0,97472	0,718%	0,457%	częstoskurcz standaryzowany	1,0007	0,86938	
						choroba typ II	1,4675	0,16691	



**Tab. XIX Modele logistyczne c.d.**

Model logistyczny	p testu ilorazu wiarygodności	N	jakość modelu: p testu Hosmera - Lemeshowa	R2 (Nagelkerke)	R2 (Coxa-Snella)	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)	porównanie z modelem zredukowanym <sup>a</sup>
<b>F</b>	0,00019*	373	0,82546	7,166%	4,500%	częstoskurcz standaryzowany stany przedomdleniowe	0,9978	0,62986	0,62953 <sup>b</sup>
							6,9754	0,00137*	
<b>dla M</b>	0,08775	143	0,16848	5,392%	3,346%	częstoskurcz standaryzowany stany przedomdleniowe	1,0019	0,7998	
							4,1052	0,0659	
<b>dla K</b>	0,00117*	230	0,74201	9,019%	5,704%	częstoskurcz standaryzowany stany przedomdleniowe	0,9952	0,4088	
							12,6904	0,0136*	
<b>dla choroby typu I - brak</b>	0,10300	153	0,96432	5,217%	2,928%	częstoskurcz standaryzowany stany przedomdleniowe	0,9939	0,4265	
							5,5337	0,1033	
<b>dla choroby typu I - TAK</b>	0,00094*	220	0,56107	9,285%	6,141%	częstoskurcz standaryzowany stany przedomdleniowe	1,0023	0,6907	
							7,9734	0,0053*	

**Tab. XIX Modele logistyczne c.d.**

Model logistyczny	p testu ilorazu wiarygodności	N	jakość modelu: p testu Hosmera - Lemeshowa	R2 (Nagelkerke)	R2 (Coxa-Snella)	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)	porównanie z modelem zredukowanym <sup>a</sup>
<b>F1</b>	0,00006*	373	0,71460	9,193%	5,773%	częstoskurcz standaryzowany stany przedomdleniowe Choroby typ I	0,9992 7,1520 1,8762	0,8671 0,0012* 0,0287*	0,0253*
<b>F2</b>	0,00065*	373	0,33146	7,173%	4,504%	częstoskurcz standaryzowany stany przedomdleniowe płeć	0,9978 6,9636 1,0360	0,6220 0,0014* 0,8983	0,8983
<b>G</b>	0,30108	389	0,86197	0,964%	0,615%	częstoskurcz standaryzowany naprzemienność zespołu QRS	0,9993 0,6123	0,88227 0,17086	
<b>H</b>	0,96847	407	0,32233	0,025%	0,016%	częstoskurcz standaryzowany krótki czasokres częstoskurczu	0,9997 1,0673	0,94090 0,80599	
<b>I</b>	0,32015	152	0,61099	2,598%	1,488%	częstoskurcz standaryzowany gr.CLw EP poniżej 280,5	1,0057 1,6973	0,48284 0,34790	

**Tab. XIX Modele logistyczne c.d.**

Model logistyczny	p testu ilorazu wiarygodności	N	jakość modelu: p testu Hosmera - Lemeshowa	R2 (Nagelkerke)	R2 (Coxa-Snella)	zmiennie niezależne	OR	wartość p dla zmiennej (test Walda)	porównanie z modelem zredukowanym <sup>a</sup>
<b>J</b>	0,14956	407	0,32116	1,460%	0,929%	wiek gr.HR>=170	1,0158 1,0228	0,06075 0,93831	
<b>P</b>	0,05047	443	0,40619	2,130%	1,339%	LP (mm) PK (mm)	1,0359 1,0436	0,18318 0,22033	
<b>P1</b>	0,00647*	443	0,26830	4,350%	2,735%	LP (mm) PK (mm) LP x PK	1,5527 1,7674 0,9839	0,00970* 0,00990* 0,01576*	0,012*
<b>R_gr1</b>	0,19577	391	0,75304	1,888%	1,193%	czynnik1_gr1 czynnik2_gr1 czynnik3_gr1	1,2230 1,1796 1,0990	0,11485 0,19516 0,47185	
<b>R_gr2</b>	0,14276	30	0,47273	20,493%	12,171%	czynnik1_gr2 czynnik2_gr2	0,3382 1,9819	0,17796 0,25506	
<b>R_gr3</b>	0,18366	385	0,97704	1,364%	0,877%	czynnik1_gr3 czynnik2_gr3	1,2150 1,1303	0,10667 0,32639	
<b>R_gr_18</b>	0,13381	111	0,74890	12,721%	7,318%	czynnik1_gr_duza czynnik2_gr_duza czynnik3_gr_duza czynnik4_gr_duza czynnik5_gr_duza	0,8489 0,5814 1,8851 1,2130 1,2697	0,57200 0,07081 0,04429* 0,40170 0,39654	

**Tab. XIX Modele logistyczne c.d.**

Model logistyczny	p testu ilorazu wiarygodności	N	jakość modelu: p testu Hosmera - Lemeshowa	R2 (Nagelkerke)	R2 (Coxa-Snella)	zmienne niezależne	OR	wartość p dla zmiennej (test Walda)	porównanie z modelem zredukowanym <sup>a</sup>
<b>W</b>	0,00595*	149	0,64601	21,597%	12,461%	Wiek_kat Na+_kat LP (mm)_kat PK (mm)_kat PMK (mm)_kat częstoskurcz standaryzowany_kat	0,9816 0,9817 1,0147 0,9838 0,9892 0,9940	0,03147* 0,10682 0,09873 0,01098* 0,17545 0,60245	
<b>S</b>	0,00001*	466	0,75369	6,296%	4,019%	scoring	0,9783	0,00002*	
<b>T</b>	0,67440	260	0,97919	0,473%	0,303%	Wiek TSH	1,0090 1,0099	0,37703 0,93061	0,493614
<b>dla K</b>	0,75485	166	0,03275*	0,517%	0,338%	Wiek TSH	1,0046 0,9064	0,73987 0,52301	0,930354
<b>T1</b>	0,00157*	259	0,02398*	9,007%	5,738%	Wiek TSH TSH x wiek	0,9467 0,1242 1,0392	0,00706* 0,00125* 0,00090*	0,000126*
<b>dla K</b>	0,00549*	165	0,48237	11,349%	7,374%	Wiek TSH TSH x wiek	0,9393 0,0777 1,0452	0,00894* 0,00281* 0,00249*	0,000434*

\* wynik istotny statystycznie ( $p < 0,05$ ); <sup>a</sup> Porównanie modeli: p- testu ilorazu wiarygodności porównującego kolejno powstające (sąsiednie) modele; W modelu zredukowanym pominięto ostatnią zmienną niezależną; <sup>b</sup> Dla modelu F: w modelu zredukowanym pominięto częstoskurcz; „\_kat” oznacza zmienną kategoryjalną, po dyskretyzacji.

**Tab. XX Miary oceniające przydatność modeli logistycznych do klasyfikacji (v-krotna walidacja)**

Model logistyczny	Czułość	Swoistość	Precyzja	F-miara			G-miara	Krzywa ROC		
				F1	F2	F3		Pkt. odcięcia	AUC	p-testu
<b>A</b>	21,43%	79,68%	7,23%	10,81%	15,38%	17,91%	41,32%	0,27	0,5327	0,35605
<b>dla krótszego czasokresu częstoskurczu (&lt;20lat)</b>	1,75%	100,00%	100,00%	3,45%	2,18%	1,95%	13,25%	0,46	0,6280	0,00293*
<b>dla dłuższego czasokresu częstoskurczu (&gt;= 20lat)</b>	0,00%	99,06%	0,00%	-	-	-	0,00%	0,30	0,5771	0,22404
<b>A1</b>	6,02%	96,30%	29,41%	10,00%	7,16%	6,54%	24,09%	0,30	0,5754	0,03393*
<b>B</b>	3,61%	99,07%	50,00%	6,74%	4,44%	3,98%	18,92%	0,24	0,5389	0,27354
<b>C</b>	7,23%	93,83%	23,08%	11,01%	8,38%	7,76%	26,04%	0,21	0,5142	0,68953
<b>D</b>	0,00%	99,69%	0,00%	-	-	-	0,00%	0,26	0,5374	0,29277
<b>E</b>	13,70%	91,67%	28,57%	18,52%	15,29%	14,45%	35,44%	0,31	0,6354	0,00033*
<b>F</b>	6,85%	95,67%	27,78%	10,99%	8,06%	7,41%	25,60%	0,25	0,6042	0,00574*
<b>dla M</b>	0,00%	100,00%	-	-	-	-	0,00%	0,25	0,6138	0,06593
<b>dla K</b>	34,78%	76,63%	27,12%	30,48%	32,92%	33,83%	51,63%	0,25	0,6301	0,00638*
<b>dla choroby typu I brak jawnych objawów</b>	0,00%	100,00%	-	-	-	-	0,00%	0,30	0,6270	0,05703
<b>dla choroby typu I TAK, występują jawne objawy</b>	7,84%	96,45%	40,00%	13,11%	9,35%	8,53%	27,50%	0,30	0,6294	0,00513*
<b>F1</b>	0,00%	100,00%	-	-	-	-	0,00%	0,29	0,6465	0,00010*
<b>F2</b>	0,00%	100,00%	-	-	-	-	0,00%	0,26	0,6073	0,00444*
<b>G</b>	55,00%	51,78%	22,80%	32,23%	42,88%	48,19%	53,37%	0,22	0,5488	0,17865

**Tab. XX Miary oceniające przydatność modeli logistycznych do klasyfikacji (v-krotna walidacja) c.d.**

Model logistyczny	Czułość	Swoistość	Precyzja	F-miara			G-miara	Krzywa ROC		
				F1	F2	F3		Pkt. odcięcia	AUC	p-testu
<b>H</b>	6,02%	98,15%	45,45%	10,64%	7,29%	6,60%	24,32%	0,21	0,5057	0,87329
<b>I</b>	4,35%	98,45%	33,33%	7,69%	5,26%	4,76%	20,69%	0,27	0,5846	0,19688
<b>J</b>	0,00%	99,69%	0,00%	-	-	-	0,00%	0,31	0,5694	0,05113
<b>P</b>	0,00%	99,72%	0,00%	-	-	-	0,00%	0,36	0,5915	0,00810*
<b>P1</b>	4,60%	100,00%	100,00%	8,79%	5,68%	5,08%	21,44%	0,39	0,61	0,00117*
<b>R_gr1</b>	3,85%	99,04%	50,00%	7,14%	4,72%	4,24%	19,52%	0,30	0,57	0,04111*
<b>R_gr2</b>	40,00%	88,00%	40,00%	40,00%	40,00%	40,00%	59,33%	0,80	0,77	0,06229
<b>R_gr3</b>	1,23%	97,70%	12,50%	2,25%	1,51%	1,36%	10,98%	0,77	0,57	0,04729*
<b>R_gr_18</b>	29,41%	96,81%	62,50%	40,00%	32,89%	31,06%	53,36%	0,36	0,68	0,01718*
<b>W</b>	39,13%	87,30%	36,00%	37,50%	38,46%	38,79%	58,45%	0,26	0,66	0,01353*
<b>S</b>	37,50%	81,89%	34,95%	36,18%	36,96%	37,23%	55,42%	0,17	0,62	0,00017*
<b>T</b>	11,11%	98,54%	66,67%	19,05%	13,33%	12,12%	33,09%	0,25	0,53	0,48687
<b>dla K</b>	0,00%	100,00%	-	-	-	-	0,00%	0,78	0,61	0,04184*
<b>T1</b>	13,21%	97,57%	58,33%	21,54%	15,63%	14,31%	35,90%	0,36	0,64	0,00119*
<b>dla K</b>	2,78%	100,00%	100,00%	5,41%	3,45%	3,08%	16,67%	0,76	0,67	0,00151*

\*wynik istotny statystycznie (p<0,05)

**Tab. XXI Wyniki analizy jednoczynnikowej – test ilorazu wiarygodności LR.**

	Nazwa zmiennej	p (LR)
1	LP (mm)	0,0362*
2	PK (mm)	0,0402*
3	PK (mm)_kat	0,0607
4	LP (mm)_kat	0,1092
5	K+ (mmol/l)	0,1023
6	BMI (kg/m <sup>2</sup> )	0,1107
7	Choroby współistniejące typu II	0,1332
8	grupa CL_kat	0,1383
9	kat BMI_kat	0,1669
10	PMK (mm)_kat	0,1672
11	Ao (mm)_kat	0,1681
12	różnica należna HR i z standaryzowanego_kat	0,1758
13	Choroby współistniejące I	0,1929
14	Wiek_kat	0,1968
15	Wiek	0,1992
16	Ao (mm)	0,2166
17	waga (kg)	0,2366
18	PMK (mm)	0,2439
19	LK (mm)	0,2613
20	częstoskurcz standaryzowany_kat	0,5075
21	CL z EP	0,5145
22	wzrost (cm)	0,5272
23	Na+_kat	0,6111
24	LK (mm)_kat	0,6383
25	Pora dnia	0,7394
26	Pora dnia_kat	0,8389
27	Na+ (mmol/l)	0,8862
28	HR częstoskurczu z badania EP	0,9232
29	TSH (mU/l)	0,9375
30	częstoskurcz standaryzowany	0,953

\*wynik istotny statystycznie (p<0,05); Nazwa zmiennej zakończona „\_kat” oznacza zmienną kategoryjną, po dyskretyzacji.

**Tab. XXII Interakcje wykryte testem LR**

	Zmienna 1	Zmienna 2	p (LR)
1	Wiek	TSH (mU/l)	0,0002*
2	LK (mm)_kat	LP (mm)	0,0009*
3	grupa CL z EP z EP	LK (mm)_kat	0,0013*
4	PK (mm)_kat	Wiek_kat	0,0015*
5	Ao (mm)_kat	PK (mm)	0,0016*
6	Ao (mm)_kat	LK (mm)_kat	0,0035*
7	LP (mm)	TSH (mU/l)	0,0039*
8	PK (mm)_kat	Ao (mm)_kat	0,0046*
9	Ao (mm)	TSH (mU/l)	0,0051*
10	Wiek_kat	TSH (mU/l)	0,0053*
11	PK (mm)_kat	Na+ _kat	0,0066*
12	LP (mm)_kat	Na+ _kat	0,0076*
13	Wiek_kat	LK (mm)_kat	0,0094*
14	LP (mm)_kat	TSH (mU/l)	0,0096*
15	LP (mm)	PK (mm)	0,0120*
16	Ao (mm)_kat	kg-waga	0,0140*
17	Wiek_kat	Wiek	0,0148*
18	Pora dnia_kat	Na+ (mmol/l)	0,0199*
19	PMK (mm)	cm- wzrost	0,0200*
20	Ao (mm)_kat	cm- wzrost	0,0223*
21	Pora dnia	BMI (kg/m <sup>2</sup> )	0,0233*
22	grupa CL z EP _kat	LP (mm)	0,0258*
23	grupa CL z EP	LP (mm)	0,0258*
24	PMK (mm)_kat	cm- wzrost	0,0259*
25	PMK (mm)_kat	Wiek_kat	0,0273*
26	Pora dnia	PK (mm)	0,0287*
27	CL z EP	cm- wzrost	0,0306*
28	grupa CL z EP	LK (mm)	0,0328*
29	grupa CL z EP _kat	LK (mm)	0,0328*
30	Pora dnia	LK (mm)	0,0361*
31	Pora dnia_kat	TSH (mU/l)	0,0387*
32	Pora dnia	LP (mm)	0,0388*
33	CL z EP	TSH (mU/l)	0,0423*
34	PMK (mm)_kat	Ao (mm)	0,0438*
35	kat2_kat	Na+ (mmol/l)	0,0649
36	Pora dnia	Na+ (mmol/l)	0,0730
37	Wiek	Na+ (mmol/l)	0,1611
38	LK (mm)	Na+ (mmol/l)	0,4924
39	K+ (mmol/l)	Na+ (mmol/l)	0,6893
40	Ao (mm)	Na+ (mmol/l)	0,8722
41	LP (mm)	Na+ (mmol/l)	0,9313

\*wynik istotny statystycznie (p<0,05)

Nazwa zmiennej zakończona „\_kat” oznacza zmienną kategorialną, po dyskretyzacji.



**Tab. XXIII Reguły związane z drzewami wygenerowanymi w analizie losowy las (zmienne po dyskretyzacji).**

Zmienna 1	Zmienna 2	Treść reguły	Licz ność	Bad rate	Przyros t (złe)	Przyrost (dobre)
różnica należna HR i z standaryzowanego_kat	TSH_kat	('różnica należna HR i z standaryzowanego_kat'=(18,55,33,6>' OR 'różnica należna HR i z standaryzowanego_kat'=(-17,3,10,95>' ) AND ('TSH_kat'=(-inf,0,499>' OR 'TSH_kat'=(1,2,1,37>' OR 'TSH_kat'=(1,95,inf)' )	45	0	0	1,262
Ao (mm)_kat	różnica należna HR i z standaryzowanego_kat	('Ao (mm)_kat'=(27,29>' OR 'Ao (mm)_kat'=(-inf,25>' ) AND ('różnica należna HR i z standaryzowanego_kat'=(-inf,-39,5>' OR 'różnica należna HR i z standaryzowanego_kat'=(-26,5,-17,3>' OR 'różnica należna HR i z standaryzowanego_kat'=(18,55,33,6>' )	42	0	0	1,321
częstoskurcz standaryzowany_kat	TSH_kat	('częstoskurcz standaryzowany_kat'=(200,214>' OR 'częstoskurcz standaryzowany_kat'=(150,180>' OR 'częstoskurcz standaryzowany_kat'=(214,inf)' ) AND ('TSH_kat'=(1,95,inf)' )	35	0	0	1,289
TSH_kat	PK (mm)_kat	('TSH_kat'=(1,95,inf)' OR 'TSH_kat'=(-inf,0,499>' ) AND ('PK (mm)_kat'=(19,22>' OR 'PK (mm)_kat'=(24,25>' OR 'PK (mm)_kat'=(27,29>' OR 'PK (mm)_kat'=(29,inf)' )	43	0	0	1,277
PK (mm)_kat	TSH_kat	('PK (mm)_kat'=(19,22>' OR 'PK (mm)_kat'=(27,29>' ) AND ('TSH_kat'=(-inf,0,499>' OR 'TSH_kat'=(1,95,inf)' )	36	0	0	1,248
PK (mm)_kat	TSH_kat	('PK (mm)_kat'=(19,22>' OR 'PK (mm)_kat'=(24,25>' ) AND ('TSH_kat'=(-inf,0,499>' OR 'TSH_kat'=(1,2,1,37>' OR 'TSH_kat'=(1,37,1,95>' OR 'TSH_kat'=(1,95,inf)' )	40	0	0	1,274

**Tab. XXIVII Reguły związane z drzewami wygenerowanymi w analizie losowy las (zmienne po dyskretyzacji) c.d.**

Zmienna 1	Zmienna 2	Treść reguły	Liczność	Bad rate	Przyrost (złe)	Przyrost (dobre)
cm- wzrost_kat	różnica należna HR i z standaryzowanego_kat	('cm- wzrost_kat' <> '(163,170)' AND 'cm- wzrost_kat' <> '(178,inf)' ) AND ('różnica należna HR i z standaryzowanego_kat' <> '(-26,5,-17,3)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(-39,5,-26,5)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(-inf,-39,5)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(18,55,33,6)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(33,6,101)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(-17,3,10,95)' )	21	0,762	3,048	0,317
TSH_kat	różnica należna HR i z standaryzowanego_kat	('TSH_kat' <> '(1,95,inf)' AND 'TSH_kat' <> '(1,2,1,37)' AND 'TSH_kat' <> '(0,499,1,2)' ) AND ('różnica należna HR i z standaryzowanego_kat' <> '(18,55,33,6)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(33,6,101)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(-17,3,10,95)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(-26,5,-17,3)' AND 'różnica należna HR i z standaryzowanego_kat' <> '(-39,5,-26,5)' )	21	0,762	3,373	0,308

W tabeli przedstawiono wybrano reguły wykluczające omdlenia (bad rate = 0) oraz pokazano dwie reguły o najwyższym bad rate.

Nazwa zmiennej zakończona „\_kat” oznacza zmienną kategorialną, po dyskretyzacji.

**Tab. XXV Reguły związane z drzewami wygenerowanymi w analizie losowy las (zmienne przed dyskretyzacją).**

Zmienna 1	Zmienna 2	Treść reguły	Liczność	Bad rate	Przyrost (złe)	Przyrost (dobre)
K+	CL z EP	('K+'<=4,365) AND ('CL'>353)	29	0	0	1,236
K+	K+	('K+'>4,325) AND ('K+'>4,715)	26	0	0	1,222
LP	TSH	('LP'<=40,5) AND ('TSH'>1,7)	66	0,015	0,091	1,182
K+		('K+'<=3,99)	65	0,031	0,155	1,209
K+	HR częstoskurczu z badania EP	('K+'<=4,37) AND (HR częstoskurczu z badania EP'>177,5)	122	0,033	0,164	1,209
kg-waga		('kg-waga'>101)	20	0,7	3,794	0,368

W tabeli przedstawiono wybrano reguły wykluczające omdlenia (bad rate = 0) oraz pokazano jedną regułę o najwyższym bad rate.;

CL – dotyczy wartości CL częstoskurczu podczas badania EP

**Tab. XXVI Tablica scoringowa – raport cech**

zmienna	Cecha:	Złe	Dobre	Suma	Procent złych	Procent dobrych	IV	WoE	Punktacja
<b>HR częstoskurczu standaryzowanego</b>	(-inf,150>	12	37	49	14,46%	11,42%	0,01	-23,59	73
	(150,180>	23	105	128	27,71%	32,41%	0,01	15,66	80
	(180,200>	26	86	112	31,33%	26,54%	0,01	-16,57	74
	(200,214>	5	35	40	6,02%	10,80%	0,03	58,40	87
	(214,inf)	17	61	78	20,48%	18,83%	0,00	-8,42	75
	Razem	83	324	407	100,00%	100,00%	0,05		
<b>PMK (mm)</b>	(-inf,8>	15	96	111	17,24%	26,89%	0,04	44,45	91
	(8,9,5>	26	87	113	29,89%	24,37%	0,01	-20,40	71
	(9,5,11>	30	129	159	34,48%	36,13%	0,00	4,68	78
	(11,12>	11	24	35	12,64%	6,72%	0,04	-63,17	57
	(12,inf)	5	21	26	5,75%	5,88%	0,00	2,33	78
	Razem	87	357	444	100,00%	100,00%	0,09		
<b>PK (mm)</b>	(-inf,19>	9	36	45	10,34%	10,11%	0,00	-2,27	76
	(19,22>	12	99	111	13,79%	27,81%	0,10	70,12	110
	(22,24>	19	58	77	21,84%	16,29%	0,02	-29,30	63
	(24,25>	8	41	49	9,20%	11,52%	0,01	22,51	88
	(25,27>	20	65	85	22,99%	18,26%	0,01	-23,04	66
	(27,29>	7	32	39	8,05%	8,99%	0,00	11,08	82
	(29,inf)	12	25	37	13,79%	7,02%	0,05	-67,51	45
	Razem	87	356	443	100,00%	100,00%	0,18		

**Tab. XXV Tablica scoringowa – raport cech c.d.**

<b>zmienna</b>	<b>Cecha:</b>	<b>Złe</b>	<b>Dobre</b>	<b>Suma</b>	<b>Procent złych</b>	<b>Procent dobrych</b>	<b>IV</b>	<b>WoE</b>	<b>Punktacja</b>
<b>LP (mm)</b>	(-inf,31>	31	177	208	35,63%	49,58%	0,05	33,03	63
	(31,32>	11	30	41	12,64%	8,40%	0,02	-40,85	94
	(32,33>	6	25	31	6,90%	7,00%	0,00	1,53	76
	(33,35>	14	39	53	16,09%	10,92%	0,02	-38,73	93
	(35,39>	14	63	77	16,09%	17,65%	0,00	9,22	73
	(39,inf)	11	23	34	12,64%	6,44%	0,04	-67,42	105
	Razem	87	357	444	100,00%	100,00%	0,13		
<b>Na+ (mmol/l)</b>	(-inf,137>	12	36	48	13,19%	9,92%	0,01	-28,49	62
	(137,139>	10	43	53	10,99%	11,85%	0,00	7,51	81
	(139,140>	9	35	44	9,89%	9,64%	0,00	-2,54	76
	(140,143>	31	150	181	34,07%	41,32%	0,01	19,31	87
	(143,144>	12	37	49	13,19%	10,19%	0,01	-25,75	63
	(144,145>	6	34	40	6,59%	9,37%	0,01	35,11	96
	(145,inf)	11	28	39	12,09%	7,71%	0,02	-44,92	53
Razem	91	363	454	100,00%	100,00%	0,06			
<b>Wiek</b>	(-inf,41>	41	193	234	42,71%	52,16%	0,02	20,00	88
	(41,46>	14	34	48	14,58%	9,19%	0,02	-46,19	52
	(46,51>	10	39	49	10,42%	10,54%	0,00	1,18	78
	(51,62>	24	66	90	25,00%	17,84%	0,02	-33,76	59
	(62,inf)	7	38	45	7,29%	10,27%	0,01	34,25	95
	Razem	96	370	466	100,00%	100,00%	0,08		

**Tab. XXVII Wybrane reguły uzyskane za pomocą algorytmu BRACID**

Pokrycie reguły	Pewność reguły	Wiek	BMI	Kołatania-age-of-onset	TSH	Pora_dnia {dzien,Noc,dzien,noc,dz ien_noc}	LK(mm)	ty_scLK(mm)	Ao(mm)	Stany przedomdleniowe {T,N}	naprzemienność zespołu QRS {nie,tak}	HR częstoskurczu (standaryzowana)	Decyzja (omdlenie T/N)
24	0,75	38,1:68,0	23,73:36,14	11,0:52,0	14,8:14,8	*	45,1:53,0	9,0:14,0	27,0:38,0	*	*	142,0:204,5	T
24	0,75	37,2:77,0	23,33:38,57	14,0:73,0	0,9:4,89	*	37,0:53,0	8,0:12,0	26,0:36,0	T	nie	139,0:210,2	T
24	0,75	39,0:68,0	23,17:38,57	36,1:52,0	1,49:2,56	*	43,0:53,0	8,0:13,0	26,0:36,0	*	*	142,0:221,0	T
23	0,74	42,1:60,9	23,17:36,14	34,1:54,0	0,48:1,91	*	45,0:58,0	8,0:13,0	26,0:34,0	*	*	150,0:230,0	T
23	0,74	35,1:49,9	18,71:29,69	11,0:43,0	0,56:1,69	*	41,0:51,0	9,0:12,0	25,0:31,0	*	*	142,0:250,0	T
21	0,71	38,2:68,0	22,91:36,14	38,1:54,9	1,27:1,55	*	41,1:52,0	10,0:13,0	30,0:36,0	*	*	120,0:221,0	T
20	0,85	38,1:60,0	23,17:33,79	34,1:54,0	0,93:1,83	dzień	45,0:53,0	8,0:13,0	26,0:36,0	*	*	142,0:221,0	T

Dla zmiennych nieuwzględnionych w regule zaznaczono \*. W przypadku danych numerycznych – zanotowano przedziały.

---

## Załącznik nr 2

---

### Teoria zbiorów przybliżonych

Teoria zbiorów przybliżonych (RS ang. rough sets) sformułowana została przez polskiego uczonego – Zdzisława Pawłaka [77]. W teorii tej do zdefiniowania wiedzy mamy pewien zbiór obiektów, które chcemy klasyfikować oraz zbiór cech, które te obiekty opisują. Ponadto oznaczmy:

**U** - uniwersum, czyli zbiór wszystkich przykładów,

**A** - zbiór atrybutów (zmiennych wyznaczonych w dowolnej skali pomiarowej),

**V** - zbiór wartości.

Zakładamy, że uniwersum jest zbiorem niepustym i skończonym. W uniwersum możemy wskazać podzbiór lub rodzinę podzbiorów. Każdy podzbiór  $X \subseteq U$  nazywamy pojęciem. Jak łatwo zauważyć, każde pojęcie  $X$  ustala na zbiorze uniwersum  $U$  podział na obiekty należące ( $u \in X$ ) i nie należące ( $u \in U \setminus X$ ) do pojęcia  $X$ . Uniwersum  $U$  można także podzielić na większą liczbę pojęć (rozłącznych podzbiorów), nazywanych klasyfikacją.

Zbiór danych przedstawiony jest w postaci tablicy, której kolumny są etykietowane przez atrybuty (parametry, własności, cechy), wiersze zaś odpowiadają obiektom. Na przecięciu wierszy i kolumn znajdują się wartości odpowiednich atrybutów dla poszczególnych obiektów.

System informacyjny (SI, ang. information system), nazywany także tablicą informacyjną, zapisywany jest często jako para uporządkowana:

$$A=(U,A),$$

gdzie:

każdy atrybut  $a \in A$  jest funkcją  $a: U \rightarrow V_a$ ,

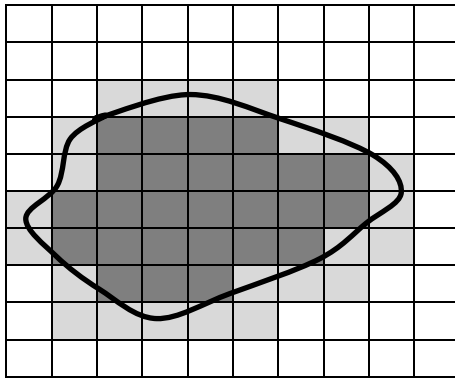
$V_a$  jest zbiorem wartości atrybutu  $a$  zwanym zakresem atrybutu  $a$ .

Tabela 39. Przykładowy system informacyjny (SI)

U	Atrybuty warunkowe			Atrybut decyzyjny
	Cecha 1 ( $a_1$ )	Cecha 2 ( $a_2$ )	Cecha 3 ( $a_3$ )	Omdlenie ( $d$ )
Pacjent (obiekt)				
1	Dużo	Tak	2	NIE
2	Mało	Nie	4	NIE
3	Średnio	Tak	1	TAK
4	Dużo	Tak	2	NIE
5	Mało	Tak	1	NIE

Tabela 39 przedstawia przykładowy system informacyjny zawierający wyniki badań dla grupy pacjentów. System ten składa się z 5 obiektów oraz 3 atrybutów warunkowych i jednego atrybutu decyzyjnego. W powyższym przykładzie tylko jeden obiekt należy do klasy omdlenia - TAK.

Prezentowana teoria zastępuje dowolny zbiór przybliżony parą zbiorów dokładnych zwanych dolnym i górnym przybliżeniem tego zbioru. Dolne przybliżenie jest zbiorem wszystkich obiektów, które na pewno należą do zbioru, zaś górne przybliżenie zawiera wszystkie obiekty, które prawdopodobnie należą do zbioru. Graficzną interpretację wprowadzonych pojęć przedstawiono na Rycinie 37 (wg [76] – w modyfikacji własnej):



Rycina 37. Dolne i górne przybliżenie zbioru - interpretacja graficzna.

Rycina 37 przedstawia zbiór wszystkich analizowanych obiektów, natomiast analizowany zbiór oznaczony został czarną krzywą (np. zbiór osób z omdleniami). Ciemnoszare kwadraty oznaczają dolne przybliżenie, natomiast jasnoszare razem z ciemnoszarymi stanowią górne przybliżenie. Jasnoszare pola, to brzeg zbioru, który jest różnicą pomiędzy górnym i dolnym przybliżeniem ocenianego zbioru.

Logicznie regułę decyzyjną możemy zapisać jako:

$$\mathbf{R} := (\mathbf{a}_1 \in \mathbf{V}_1) \wedge \dots \wedge (\mathbf{a}_k \in \mathbf{V}_k) \rightarrow (\mathbf{decyzja} = \mathbf{d})$$

gdzie:

koniunkcja testów na atrybutach jest warunkiem reguły

$R = \{r_1, r_2, \dots, r_n\}$  to zbiór reguł

Przykład wyliczania miar jakości prostej reguły dla Tabeli 39:

**„Jeżeli** ( $a_1 = \text{„Dużo”}$ ) i jednocześnie ( $a_2 = \text{„TAK”}$ ) **to** decyzja= „NIE”

Wsparcie reguły: 5; Liczba pozytywnych przykładów reguły: 5

Dokładność reguły wynosi zatem:  $5/5 = 1$  (100%)

Liczba wszystkich obiektów z decyzją NIE: 370

Pewność reguły:  $5/370 = 0,0135$  (czyli jedynie ok 1,35%)