**Streszczenie w języku polskim**

Badania randomizowane są głównym źródłem wiedzy o skuteczności i bezpieczeństwie interwencji medycznych we współczesnej medycynie opartej na dowodach (EBM - ang. Evidenced Based Medicine). Badania obserwacyjne pozostają ważnym źródłem wiedzy w medycynie klinicznej. Metody badawcze, takie jak wnioskowanie przyczynowe, prawdopodobnie przyczynią się do rosnącej roli tego typu badań w medycyny opartej na dowodach z uwagi na niższe koszty. Szczególnie biorąc pod uwagę relatywnie łatwą dostępność dużych ilości danych tworzonych w trakcie opieki nad pacjentem w erze „big data”.

Analiza danych powstających w codziennej praktyce klinicznej, gromadzonych w formie elektronicznej dokumentacji medycznej (EDM), dostarcza ważnych danych na temat populacji pacjentów kardiologicznych, w szczególności grup słabo reprezentowanych w badaniach klinicznych. Według stanowiska Europejskiego Towarzystwa Kardiologicznego, dane te mogą być użyteczne do prowadzenia badań i rejestrów retrospektywnych, generowaniu nowych hipotez badawczych, monitorowaniu zdarzeń niepożądanych oraz planowaniu i zmniejszaniu kosztów prowadzenia badań randomizowanych. Głównym ograniczeniem wykorzystania elektronicznej dokumentacji medycznej do badań, jest fakt, że większość gromadzonych informacji to tzw. dane nieustrukturyzowane np. w formie opisowej (opisy hospitalizacji, listy rozpoznań, zalecenia lekarskie, obserwacje, opisy badań obrazowych). Dane nieustrukturyzowane wymagają czasochłonnej manualnej analizy i wprowadzenia do ustrukturyzowanego formatu bazy danych przez personel medyczny. Ogranicza to możliwość prowadzenia analiz na dużych kohortach chorych lub wyszukania grup pacjentów o ściśle określonej charakterystyce, w krótkim czasie. Stanowi to przeszkodę do efektywnej weryfikacji hipotez badawczych w historycznych oraz aktualnie leczonych populacjach chorych.

Wspomniane utrudnienia w wykorzystaniu elektronicznej dokumentacji medycznej próbuje się mitygować poprzez wykorzystanie danych ustrukturyzowanych tworzonych podczas udzielania świadczeń zdrowotnych np. danych administracyjnych. Kody rozpoznań chorobowych i procedur leczniczych (ICD-10, ICD-9) raportowane do płatników usług zdrowotnych takich jak polski Narodowy Fundusz Zdrowia (NFZ) są atrakcyjnym obiektem badań z uwagi na możliwość szybkiego uzyskania danych dotyczących ogromnych populacji chorych. Jednak ich wykorzystanie jest obarczone ryzykiem błędów we wnioskowaniu. Kodowane rozpoznania oraz procedury medyczne, mogą nie odzwierciedlać trafnie faktycznego stanu klinicznego pacjentów, ponieważ głównym ich celem jest prowadzenie rozliczeń administracyjnych.

Powyższe ograniczenia, w kontekście prowadzenia badań w kardiologii, były identyfikowane w rożnych systemach ochrony zdrowia na świecie, w Polsce natomiast dane na ten temat są skąpe.

Trwają poszukiwania innych sposobów efektywnego pozyskiwania danych do badań naukowych opartych o elektroniczną dokumentację medyczną. Jednym z nich, jest wykorzystanie technik procesowania języka naturalnego (NLP) na danych w elektronicznej dokumentacji medycznej w celu automatyzacji procesu gromadzenia danych klinicznych na potrzeby badawcze. Dlatego też celem omawianej rozprawy doktorskiej było:

(1) analiza obecnie wykorzystywanych danych ustrukturyzowanych (kody rozpoznań chorobowych), w badaniach w dziedzinie kardiologii w Polsce, w celu identyfikacji ograniczeń dostępnych danych ustrukturyzowanych,

(2) wypracowanie rozwiązań alternatywnych wykorzystujących techniki procesowania tekstu w elektronicznej dokumentacji medycznej w celu pozyskiwania danych do badań w dziedzinie kardiologii,

(3) porównanie zaproponowanego narzędzia „AssistMED", opartego o NLP, z manualnym pozyskiwaniem danych.

W ramach wykonanych badań (publikacja nr 1), przeprowadzono kompleksowe porównanie zgodności danych klinicznych w dokumentacji medycznej do danych opartych o zarejestrowane w Narodowym Funduszu Zdrowia kody ICD-10. Badanie przeprowadzono na historycznej kohorcie 3338 pacjentów z rozpoznaniem migotania przedsionków. Zidentyfikowano istotne rozbieżności między charakterystyką pacjentów opartą o dane administracyjne NFZ, w porównaniu do dokumentacji medycznej, wynikające z nietrafności i braków w raportowanych kodach ICD-10. Ponadto zauważono, że dane posiadane przez płatnika, nie zawierały istotnych z punktu widzenia wnioskowania w kardiologii, informacji klinicznych takich jak: dane o niektórych szczegółowych rozpoznaniach chorobowych, dane o stosowanych lekach i ich dawkowaniu, danych z badania echokardiograficznego.

W kolejnym etapie (publikacja nr 2), opracowano autorskie rozwiązanie „AssistMED" wykorzystujące techniki NLP w obrębie określonych typów danych opisowych EDM w języku polskim w celu automatycznego pozyskiwania szerokiej charakterystyki klinicznej dużych populacji chorych kardiologicznych: rozpoznań klinicznych, stosowanych leków i ich dawkowania, liczbowych parametrów echokardiograficznych. Wykonano analizy ilościowe i jakościowe wykorzystania narzędzia na populacji 400 zanonimizowanych rekordów pacjentów w stosunku do pozyskiwania danych przez człowieka, w celu kompleksowego scharakteryzowania ograniczeń wynikających z zastosowania metod NLP. Analiza ilościowa wykazała, że wykorzystanie AssistMED do gromadzenia danych, pozwalało na osiągnięcie wyników wysoce zbieżnych z manualnym wprowadzaniem danych. W analizie jakościowej jako główne przyczyny błędów zidentyfikowano między innymi: brak zaawansowanej analizy kontekstu (ograniczenia technik NLP), losowe błędy algorytmu, błędy literowe w EDM, złożony opis dawkowania leku.

W publikacji nr 3, narzędzia AssistMED użyto na zanonimizowanej kohorcie 10314 pacjentów hospitalizowanych w oddziale kardiologicznym (lata 2016-2019). Metodę automatyczną porównano z manualną weryfikacją danych, w celu scharakteryzowania retrospektywnej kohorty pacjentów z rozpoznaniem migotania przedsionków. Wykazano niewielkie rozbieżności między danymi uzyskiwanymi w sposób automatyczny i manualny, przy jednocześnie wielokrotnie krótszym czasie pozyskania danych w sposób automatyczny.

Skonkludowano, że chociaż dane administracyjne np. kody ICD-10 są relatywnie szybkim do uzyskania źródłem danych do badań naukowych, to mają istotne ograniczenia. Wykazano, że wykorzystanie wypracowanych w toku badań technik opartych o NLP na danych z EDM, może pozwolić na uzyskiwanie szerokiej charakterystyki populacji na potrzeby badań w kardiologii, w krótkim czasie i o wysokiej zbieżności z danymi pozyskiwanymi poprzez analizę danych przez człowieka.