Akceptuję

**Recenzja rozprawy na stopień doktora nauk medycznych i nauk o zdrowiu w dyscyplinie nauki medyczne lek. Cezarego Maciejewskiego**

„Wykorzystanie nowoczesnych technik analizy danych tekstowych w elektronicznej dokumentacji medycznej, w celu stworzenia narzędzi przyspieszających pozyskanie wartościowych naukowo danych ustrukturyzowanych oraz zautomatyzowanych skal ryzyka w kardiologii"

Promotor pracy: prof. dr hab. n med. Paweł Balsam

Promotor pomocniczy: dr hab. n. med. Krzysztof Ozierański

Współczesna medycyna coraz bardziej ewoluuje w kierunku tzw. medycyny opartej na faktach. To medycyna, w której praktycznie wszystkiego należy dowieźć w sposób naukowy, najlepiej w sposób nie pozostawiający żadnych wątpliwości, że dana interwencja, sposób leczenia czy też diagnostyki są skuteczne, bezpieczne, kosz-efektywne i mogą być zastosowane globalnie w codziennej praktyce klinicznej. Bez wątpienia dzisiaj „złotym standardem" w testowaniu interwencji medycznej są prospektywne badania randomizowane (RCT – randomised controlled trials). Wyniki takich dużych, najlepiej wieloośrodkowych badań są następnie podstawą do aktualizacji bądź nawet zmiany zaleceń medycznych przez międzynarodowe i narodowe towarzystwa naukowe. Aktualnie badania randomizowane mają najwyższą skalę zaleceń medycznych. Ale RCTs nie są oczywiście pozbawione wad. Ich główną wadą jest „selection bias", czyli niewłączanie wszystkich kolejnych chorych do przetestowania badanej interwencji medycznej. Kryteria włączenie i wyłączenia do RCTs czasami powodują, że bardzo znaczny odsetek chorych nie kwalifikuje się do nich, co prowadzi do szeregu konsekwencji, jak np. podważanie możliwości generalizowania wyników i ich

zastosowania w populacjach nieprzebadanych. Tą lukę generowaną przez RCTs w sposób znakomity uzupełniają badania obserwacyjne. Ich główną zaletą jest możliwość analizy danych wszystkich, kolejnych pacjentów (bez wyjątków), ze szczególnym uwzględnieniem rejestrowania działań niepożądanych i powikłań różnego rodzaju interwencji, co nazywane jest czasami „real-life evidence". Oczywistą wadą natomiast jest brak randomizacji – ale badania te nie służą, co do zasady, do oceny porównawczej interwencji medycznej. Tym samym badania obserwacyjne są niezbędnym elementem badania, analizowania i raportowania naszej codziennej aktywności medycznej. Pytanie zasadnicze brzmi – jak to zrobić? A w zasadzie jak to robić w sposób ciągły i dynamiczny, aby na bieżąco móc korzystać ze stale aktualizowanych danych?

W przedstawionej mi do recenzji rozprawie, Doktorant, lekarz Cezary Maciejewski, podjął się bardzo trudnego zadania, a mianowicie znalezienia sposobu na to, aby w sposób inny niż manualny (powszechnie dotąd stosowany), spróbować pozyskać cenne dane z dostępnej dokumentacji medycznej. Do tej pory, każdy z nas, próbujący jakiekolwiek działalności medycznej, tworzył tzw. „bazę danych" – czyli zbiór danych niezbędnych do szeregu dalszych analiz medycznych, naukowych, administracyjnych, rozliczeniowych itd. Postępowanie takie jest powszechne, bardzo czasochłonne i w zasadzie niemożliwe do aktualizacji (aktualizacją danej bazy jest ponowne stworzenie bazy danych praktycznie od nowa). Doktorant opracował natomiast autorskie rozwiązanie o nazwie „AssistMED" wykorzystujące techniki procesowania języka naturalnego (NLP) na danych w elektronicznej dokumentacji medycznej w celu automatyzacji procesu gromadzenia danych klinicznych na potrzeby badawcze. Tym samym możliwe stało się automatyczne pozyskiwanie danych dotyczących szerokiej charakterystyki klinicznej dużych populacji chorych kardiologicznych: rozpoznań klinicznych, stosowanych leków i ich dawkowania oraz liczbowych parametrów echokardiograficznych.

Cytując Autora: „Celem niniejszej rozprawy doktorskiej, było scharakteryzowanie występujących ograniczeń w zakresie wykorzystania elektronicznej dokumentacji medycznej do badań w dziedzinie kardiologii w Polsce oraz wypracowanie nowych rozwiązań opartych o techniki procesowania tekstu w elektronicznej dokumentacji medycznej." Szczegółowo cele pracy były następujące:

1. Scharakteryzowanie dokładności i ograniczeń dostępnych w Polsce danych ustrukturyzowanych (kody rozliczeniowe ICD-10) w kontekście prowadzenia badań w dziedzinie kardiologii - na przykładzie populacji pacjentów leczonych z powodu migotania przedsionków.
2. Opracowanie założeń merytorycznych i wdrożenie systemu wykorzystującego techniki NLP w celu zautomatyzowanego pozyskiwania danych ustrukturyzowanych z określonych typów danych tekstowych w EDM: rozpoznań klinicznych, substancji leczniczych i dawkowania oraz liczbowych parametrów echokardiograficznych.
3. Analiza dokładności i szybkości pozyskania danych z wykorzystaniem wypracowanego narzędzia opartego o NLP, w porównaniu do danych pozyskiwanych przez człowieka na przykładzie dużej populacji pacjentów leczonych z powodu migotania przedsionków.
4. Scharakteryzowanie ograniczeń narzędzia opartego o wykorzystanie NLP w EDM.

Przedstawiona mi do recenzji rozprawa doktorska lekarza Cezarego Maciejewskiego jest syntezą całego przedsięwzięcia związanego z omawianym zagadnieniem. Wieńczy je cykl trzech publikacji w recenzowanych czasopismach międzynarodowych z wysokim wskaźnikiem cytowań:

1. Maciejewski C., Ozierański K., Basza M., Lodziński P., Śliwczyński A., Kraj L., Krajsman M., Prado Paulino J., Tymińska A., Opolski G., Cacko A., Grabowski M., Balsam P.; Administrative Data in Cardiovascular Research—A Comparison of Polish National Health Fund and CRAFT Registry Data.; Int. J. Environ. Res. Public Health. 2022,19(19), 11964
2. Maciejewski C., Ozierański K., Barwiołek A., Basza M., Bożym A., Ciurla M., Krajsman M., Maciejewska M., Lodziński P., Opolski G., Grabowski M., Cacko A., Balsam P.; AssistMED project: transforming cardiology cohort characterisation from electronic health records through natural language processing – algorithm design, preliminary results, and field prospects.; Int J Med Inform. 2024 May:185:105380.
3. Maciejewski C., Ozierański K., Basza M., Barwiołek A., Ciurla M., Bożym A., Krajsman M., Lodziński P., Opolski G., Grabowski M., Cacko A., Balsam P.; Practical use case of natural language processing for observational clinical research data retrieval from electronic health records: AssistMED project.; Pol Arch Intern Med . 2024 Mar 19:16704.

Publikacja nr 1 dotyczyła kompleksowego porównanie zgodności danych klinicznych w dokumentacji medycznej do danych opartych o zarejestrowane w Narodowym Funduszu Zdrowia (NFZ) kody ICD-10. Badanie przeprowadzono na grupie 3338 pacjentów z rozpoznaniem migotania przedsionków i dostarczyło cennych informacji na temat istotnych rozbieżności między charakterystyką pacjentów opartą o dane administracyjne NFZ, w porównaniu do dokumentacji medycznej, wynikające z nietrafności i braków w raportowanych kodach ICD-10.

Publikacja nr 2 dotyczyła opracowania autorskiego rozwiązania „AssistMED" wykorzystującego techniki NLP w obrębie określonych typów danych opisowych elektronicznej dokumentacji medycznej (EDM) w języku polskim w celu automatycznego pozyskiwania szerokiej charakterystyki klinicznej dużych populacji chorych kardiologicznych. Badania przeprowadzono na populacja 400 rekordów chorych i stwierdzono, że wykorzystanie AssistMED do gromadzenia danych, pozwalało na osiągnięcie wyników wysoce zbieżnych z manualnym wprowadzaniem danych. Główne przyczyny błędów dotyczyły między innymi: braku zaawansowanej analizy kontekstu, losowego błędu algorytmu, błędów literowych w EDM oraz złożonego opis dawkowania leku.

W Publikacji nr 3 użyto narzędzia AssistMED na kohorcie kolejnych 10314 pacjentów hospitalizowanych w oddziale kardiologicznym w latach 2016-2019, a metodę automatyczną porównano z manualną. Z dostępnych 10314 kart wypisowych zidentyfikowano odpowiednio 3030 i 3029 pacjentów za pomocą analizy przez człowieka i opartego na NLP, co odzwierciedlało >99% dokładność NLP w wykrywaniu pacjentów z migotaniem przedsionków. Charakterystyka pacjenta za pomocą NLP była szybsza od analizy wykonanej przez anotatora (3 godziny i 15 minut versus 71 godziny i 12 minut). Stwierdzono prawie idealną zgodność między NLP, a charakterystyką określoną przez człowieka w zakresie rozpoznań klinicznych, substancji lekowych i liczbowych parametrów echokardiograficznych. Identyfikacja dziennej dawki przyjmowanej substancji leczniczej była najmniej dokładną cechą NLP.

Na podstawie przeprowadzonych badań Doktorant wysunął następujące wnioski:

1. Dokładność danych administracyjnych pozyskiwanych z NFZ jest ograniczona w kontekście wnioskowania o charakterystyce klinicznej pacjentów. Dane administracyjne wyrażone w postaci kodów ICD-10, nie odzwierciedlają niektórych ważnych z punktu

widzenia badaczy aspektów klinicznych. Nie zawierają również informacji o stosowanych lekach czy parametrach echokardiograficznych, co ma istotne znaczenie w badaniach w kardiologii.

2. Techniki NLP, mogą pozwolić na dokładne i szybkie scharakteryzowanie pacjentów w populacji kardiologicznej, w porównaniu do analizy danych przez człowieka.
3. Techniki NLP charakteryzują się określonymi ograniczeniami w kontekście pozyskiwania trafnych, ustrukturyzowanych danych klinicznych z elektronicznej dokumentacji medycznej.
4. Rozwój algorytmów procesowania tekstu (w szczególności tzw. dużych modeli językowych dla języka polskiego) może umożliwić szerokie zastosowanie NLP, w celu prowadzenia badań w kardiologii. W celu pozyskania wiarygodnych danych, konieczne będzie zaangażowanie osób z wiedzą kliniczną. Niezbędna będzie też walidacja wypracowanych rozwiązań, w celu upewnienia się co do jakości danych uzyskiwanych automatycznie oraz dokumentacji ograniczeń wdrażanych narzędzi

Praca doktorska, którą mam przyjemność i zaszczyt recenzować jest, moim zdaniem, wzorem prac doktorskich. Gratuluję Doktorantowi, Promotorowi oraz Promotorowi Pomocniczemu, innym członkom zespołu i współautorom publikacji pomysłu i jego realizacji. Doktorant oraz współautorzy projektu zauważyli problem, zdiagnozowali go, wykazali zasadność rozwiązania alternatywnego, zaproponowali nowe, autorskie narzędzie medyczne, zaangażowali w jego realizację specjalistów z zakresu informatyki, zaproponowali zakres działania narzędzia AssistMED, zwalidowali jego skuteczność i rozpoznali słabe strony, a finalnie całość projektu opublikowali w recenzowanych czasopismach medycznych. To co Doktorant proponuje, czyli praktycznie automatyczne przetransferowanie danych z elektronicznej dokumentacji medycznej o bardzo wysokim odsetku zgodności ze stanem

rzeczywistym, do dostępnej obróbce statystycznej bazy danych to marzenie każdego naukowca. Jeśli do tego dołożymy bardzo szybki czas generowania danych oraz możliwość ich ciągłej aktualizacji, to otrzymujemy bazę idealną. Cała rozprawa doktorska jest opisem i podsumowaniem projektu, o którym mowa powyżej. Tekst rozprawy jest dobrze i płynnie napisany, logiczny i spójny, prowadzi czytającego chronologicznie od prawidłowo zaprojektowanych założeń i celów badania, aż po wyniki końcowe. Fakt zaprojektowania tak trudnego zadania, jego realizacja oraz publikacje dowodzą, że Doktorant wykazuje bardzo dobre przygotowanie merytoryczne i znajomość struktury badań naukowych.

Gratulując raz jeszcze osiągniecia, pozwolę sobie na kilka subiektywnych komentarzy dotyczących całego przedsięwzięcia:

1. Zanim tacy wyspecjalizowani asystenci medyczni typu AssistMED staną się narzędziem, które będzie rutynowo wykorzystywane do ekstrakcji danych medycznych do statystycznych baz medycznych, musimy mieć pewność, że rzeczywiście mamy do czynienia z wiarygodnymi danymi i dobrym kontekstem medycznym ich analizy. Należy pamiętać, że przy analizie dużych grup chorych, błędy powstają również w grupie kontrolnej (po stronie anotatorów). Doktorant, który jest jednocześnie trzecim anotatorem, był odpowiedzialny za rozwiązywanie niezgodności między dwoma anotatorami, gromadzącymi dane manualnie, a urządzeniem medycznym. W przypadku powszechnego użycia urządzenia nie będzie już takiej możliwości. Dlatego, moim zdaniem, narzędzie należałoby przetestować i zwalidować wieloośrodkowo.
2. Recenzent rozumie, że aktualne (i tak już bardzo duże) możliwości urządzenia można by w toku jego udoskonalania w przyszłości rozszerzyć o ekstrakcję również innych, bardzo ważnych danych medycznych dotyczących między innymi inwazyjnych badań hemodynamicznych, elektrofizjologicznych, diagnostyki obrazowej itd.

3. Drobne uwagi dotyczące samego tekstu rozprawy doktorskiej, drobnych błędów językowych i stylistycznych. Byłoby również świetnie, gdyby strony rozprawy zostały ponumerowane, szczególnie kiedy Autor na początku rozprawy zamieszcza spis treści.

Komentarze i uwagi recenzenta wynikają z ciekawości i chęci pogłębienia znajomości zagadnienia i w żaden sposób nie umniejszą jej wartości, którą uznaję, jak już wspominałem wcześniej, za znakomitą. Podsumowując, stwierdzam, że praca doktorska lek. Cezarego Maciejewskiego pt. „Wykorzystanie nowoczesnych technik analizy danych tekstowych w elektronicznej dokumentacji medycznej, w celu stworzenia narzędzi przyspieszających pozyskanie wartościowych naukowo danych ustrukturyzowanych oraz zautomatyzowanych skal ryzyka w kardiologii" jest bardzo wartościowym i oryginalnym opracowaniem oraz dowodzi bardzo wysokiego kunsztu naukowego Doktoranta.

Przedstawiona mi do recenzji rozprawa doktorska spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668). W związku z tym wnioskuję do Wysokiej Rady o dopuszczenie lek. Cezarego Maciejewskiego do dalszych etapów przewodu doktorskiego. Jednocześnie, z uwagi na wysoką wartość merytoryczną i kliniczną rozprawy, wprowadzenie nowego, autorskiego narzędzia medycznego oraz opublikowanie wyników rozprawy w czasopismach z listy filadelfijskiej wnioskuję również o jej wyróżnienie.

Zabrze, dnia 31.08.2024 r.

dr hab. med. Michał Mazurek
Oddział Kliniczny Kardiologii
Katedry Kardiologii, Wrodzonych
Wad Serca i Elektroterapii
Śląskie Centrum Chorób Serca w Zabrzu

dr hab. n. med.
Michał Mazurek
kardiolog 2290957

Michał Mazurek