Akceptuję

WOJSKOWY INSTYTUT MEDYCZNY-
PAŃSTWOWY INSTYTUT BADAWCZY
Klinika Kardiologii i Chorób Wewnętrznych
płk prof. dr hab. n. med. i n. o zdr. Paweł Krzesiński
04-141 Warszawa, ul. Szaserów 128
tel. kom. 665 707 580

Warszawa, 12.07.2024

Ocena rozprawy doktorskiej lek. Cezarego Piotra 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"

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

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

I Katedra i Klinika Kardiologii, Warszawski Uniwersytet Medyczny

Kierownik Kliniki: prof. dr hab. n. med. Marcin Grabowski

Wprowadzenie do recenzji

Analiza danych powstających w codziennej praktyce klinicznej, gromadzonych w formie elektronicznej dokumentacji medycznej (EDM) może przyczynić się do istotnego postępu medycyny. W erze dynamicznie rozwijających się informatycznych narzędzi analizy dużych zbiorów danych nasze możliwości wnioskowania o problemach zdrowotnych, diagnostyce, leczeniu i rokowaniu pacjentów reprezentujących w sposób jak najbardziej naturalny populację ogólną osiągają istotnie wyższy poziom. To ogromna szansa dla rozwoju medycyny, którą jednak należy wykorzystać ze świadomością potencjalnych „pułapek" i ograniczeń. Podstawowym warunkiem uzyskiwania wyników odzwierciedlających rzeczywistość jest bowiem poprawność, kompletność oraz wiarygodność kliniczna danych, które poddamy analizie lub użyjemy w budowaniu modeli i algorytmów. Niestety trudno tego oczekiwać od baz danych budowanych na bazie raportów dla celów administracyjnych, statystycznych i rozliczeniowych, takich jak choćby klasyfikacja ICD-10. Kolejnym wyzwaniem jest fakt, że większość gromadzonych informacji o istotnych znaczeniu klinicznym i naukowym to tzw. dane nieustrukturyzowane np. w formie opisowej (opisy hospitalizacji, listy rozpoznań, zalecenia lekarskie, obserwacje, opisy badań obrazowych), wymagające czasochłonnej manualnej analizy i wprowadzenia do ustrukturyzowanego formatu bazy danych przez personel medyczny. Potrzebujemy zatem narzędzi informatycznych, które zastąpią oko ludzkie w analizie dokumentacji i wyekstrahują z niej istotne dane w zakresie nie gorszym niż zrobiłby to lekarz.

Takie narzędzia dla dokumentacji prowadzonej w języku angielskim już powstają ale warunkiem ich rozpowszechnienia w Polsce jest dostępność wersji zdolnych do analizy naszego ojczystego języka. Nie

jest to jedynie kwestia prostego przetłumaczenia wersji oprogramowania, bowiem każdy język ma swoją unikalną strukturę i słownik pojęć medycznych, co może oznaczać że techniki efektywne w języku angielskim nie zawsze sprawdzą się w analizie polskich danych.
Mimo że do tej pory moje uwagi wstępne mogą wydawać się mało związane z kardiologią, to byłoby to wrażenie jedynie pozorne. To właśnie wnioskowanie z dobrze opracowanych danych rzeczywistych jest kluczowe aby do naszej codziennej praktyki klinicznej skutecznie adoptować obowiązujące wytyczne, oparte w dużej mierze na badaniach randomizowanych, a zatem nie w pełni reprezentatywnych dla populacji ogólnej. Przedstawiona do recenzji rozprawa doktorska lek. Cezarego Piotra 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"* dotyczy zatem tematu jak najbardziej aktualnego, odnosząc się do dynamicznie rozwijającej się obszaru medycyny. Należy docenić zaangażowanie lek. Cezarego Piotra Maciejewskiego w tworzenie narzędzi informatycznych, ponieważ to właśnie udział profesjonalistów medycznych, przyszłych użytkowników końcowych, gwarantuje ich poprawne wdrażanie w praktyce klinicznej.

Omówienie rozprawy

Opis ogólny

Rozprawa doktorska opiera się na omówieniu trzech ściśle powiązanych ze sobą publikacji oryginalnych, opublikowanych w renomowanych czasopismach o zasięgu międzynarodowym:

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. IF - ; MNiSW 140
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. IF 4.900; MNiSW 140
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. IF 4.800; MNiSW 200

We wszystkich pracach lek. Cezary Piotr Maciejewski jest pierwszym autorem ze zdecydowanie dominującyh wkładem, odpowiednio 86%, 81% i 82%.

Łączna punktacja cyklu to **IF 9.800 pkt; MNiSW: 480 pkt.**

Należy podkreślić, że w pracach nr 2 i 3. wykorzystano autorskie rozwiązanie „AssistMED" oparte na technikach 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 oraz liczbowych parametrów echokardiograficznych.

Zasadnicza część rozprawa doktorskiej liczy 49 stron. Jest ona uzupełniona załącznikami – pisemnymi oświadczeniami współautorów prac.
W części rozpoczynającej rozprawę zamieszczono wykaz publikacji wchodzących w skład cyklu, spis stosowanych skrótów oraz streszczenia. Kolejną części stanowi **Wstęp**, następnie wymieniono precyzyjnie **Założenia i cel pracy**. Na kolejnych 6 stronach przedstawiono **opis poszczególnych publikacji** a następnie zaprezentowano je w **wersjach oryginalnych.** Zasadniczą części rozprawy kończy **Podsumowanie i wnioski. Piśmiennictwo** zostało dobrane trafnie, dominują prace z ostatniego dziesięciolecia, co podkreśla aktualność podjętego tematu. Pisemne oświadczenia współautorów prac wskazują precyzyjnie ich udział w publikacjach wchodzących w skład cyklu.
Rozprawa przygotowana jest bardzo przejrzyście, napisana poprawnym i komunikatywnym językiem. Również załączone publikacje są pod tym względem bardzo dobrze przygotowane. Zwraca uwagę duża staranność edytorska całego materiału.

Ocena merytoryczna

Wstęp i cele pracy

Wstęp jest napisany w sposób dojrzały i elegancki. Jasno sprecyzowano problem badawczy, argumenty za podjęciem badań i związanych z nim rozwojowych prac technologicznych, wskazano również potencjalne korzyści z jego rozwiązania. Trafnie wskazano szanse i wyzwania związane z rozwojem EDM oraz informatycznych narzędzi ich analizy. Lektura wstępu jednoznacznie wskazuje, że Doktorant miał wizję konsekwentnej etapowej realizacji badań.
Dobrze zdefiniowano hipotezy i cele badawcze. Uszeregowano je w logicznej kolejności:

1. Scharakteryzowanie dokładności i ograniczeń dostępnych w Polsce danych ustrukturyzowanych 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.

Założenia badawcze rozprawy świadczą o jej oryginalności.

Wyniki

W tej części przedstawiono wyniki poszczególnych wątków badawczych, odwołując się do publikacji wchodzących w skład cyklu. Zaprezentowano kluczowe rezultaty pracy, umieszczając obok komentarzy tekstowych najważniejsze ryciny. Mimo, że przed Doktorantem stanęło niełatwe zadanie zwięzłego przedstawienia technologii informatycznych, wykonał to z dużą dojrzałością i zdolnością prezentacji kwintesencji swoich badań.
Nie ma wątpliwości, że wykonanie badania była dużym wyzwaniem. Doktorant musiał pozyskać dane zarówno z ośrodka macierzystego, jak i z NFZ. Potrafił nawiązać współpracę wieloośrodkową a badania

i prace rozwojowe zaplanował bardzo szczegółowo i profesjonalnie. Należy podkreślić, że badania dotyczące oceny nowych technologii są mniej powszechne i charakteryzują się specyficznym podejście statystycznym do analizy wyników. W tym przypadku nie budzi ona zastrzeżeń i pozwala rzetelnie i szczegółowo ocenić wartość praktyczną rozwiązania „AssistMED".

W publikacji nr 1 ocenie poddano dokładność charakterystyki klinicznej pacjentów uzyskiwanej w oparciu o dane rozliczeniowe dostępne w NFZ (kody ICD-10), w kontekście określania ryzyka niedokrwiennego i krwotocznego pacjentów z migotaniem przedsionków, odnosząc je do wyników analizy kart wypisowych dokonanej przez personel. Dla kohorty 3338 pacjentów zidentyfikowanych zostało 565521 świadczeń zdrowotnych z przypisanymi im kodami ICD-10. Wyniki wskazały na istotne statystycznie różnice pomiędzy charakterystyką pacjentów, opartą o analizę dokumentacji medycznej wykonaną przez człowieka, w porównaniu do charakterystyki opartej na diagnozach medycznych zaraportowanych do NFZ w formie kodów ICD-10: dane z NFZ zaniżały odsetek pacjentów z migotaniem przedsionków w kohorcie, jednocześnie zawyżając odsetek pacjentów z innymi chorobami układu krążenia.. Dane NFZ dały istotnie wyższe wyniki w skali CHA2DS2VASc i HAS-BLED, co wskazywało na przeszacowywanie ryzyka niedokrwiennego i krwotocznego pacjentów według danych administracyjnych. W dyskusji przybliżono wyniki identyfikacji poszczególnych chorób z innych systemów ochrony zdrowia, które wskazywały na analogiczne ograniczenia danych rozliczeniowych na świecie. Wykazano zatem znaczące różnice między dokumentacją medyczną a danymi rozliczeniowymi w zakresie stwierdzanych jednostek chorobowych w dziedzinie kardiologii, które mogą mieć istotny wpływ na wnioskowanie w badaniach naukowych.
Wyniki tej pracy uzasadniają podjęcie badań nad wykorzystaniem nowej technologii „AssistMED", które przedstawiono w dwóch kolejnych publikacjach.
W publikacji nr 2 szczegółowo przedstawiono założenia merytoryczne, proces projektowania oraz aspekty techniczne wdrożenia projektu AssistMED. Opiera się on na technice procesowania języka naturalnego, w celu pozyskiwania danych z EDM dotyczących w zakresie rozpoznań klinicznych, substancji leczniczych i dawkowania, liczbowych parametrów echokardiograficznych. Pod uwagę wzięto znaczącą liczbę zmiennych. Nie odnotowano statystycznie istotnych różnic w charakterystyce pacjentów pomiędzy algorytmem a podwójną analizą przez człowieka. Analiza jakościowa wykazała, że niezgodności algorytmu z ręczną anotacją, wynikały przede wszystkim z przypadkowych błędów algorytmu, błędnej anotacji przez człowieka i braku zaawansowanej świadomości kontekstu naszego narzędzia NLP. Wykazano, że pozyskanie wiarygodnej, szerokiej charakterystyki klinicznej z danych opisowych w EDM prowadzonej w języku polskim w sposób automatyczny jest możliwe; Opisano wyzwania merytoryczne i techniczne przy wdrażaniu algorytmu NLP; scharakteryzowano ograniczenia i wskazane potencjalne sposoby ich zaadresowania w przyszłości, z nakreśleniem konkretnych rozwiązań technicznych. W szczególności warte docenienia jest poddanie skrupulatnej analizie przypadków rozbieżności kwalifikacji danych przez algorytm AssistMED i człowieka, co w wybranych przypadkach ujawniło niedoskonałości pracy ludzkiej, zarówno w zakresie wpisywania danych, jak i ich wtórnej analizy.
W publikacji nr 3 przedstawiono praktyczne zastosowanie AssistMED, do pozyskiwania szczegółowej charakterystyki klinicznej 3030 pacjentów, w porównaniu z rejestrowaniem danych przez człowieka. Przedmiotem analizy byli kolejni pacjenci wypisywani z oddziału kardiologii w latach 2016-2019. Wykazano wysoką zbieżność wskazań metody automatycznej opartej o NLP z weryfikacją przez człowieka, przy jednocześnie znacznie krótszym czasie pozyskania zbioru danych: kompleksowa charakterystyka pacjenta za pomocą NLP była szybsza niż analiza przez anotatora (3 godziny i 15 minut

w porównaniu z 71 godzinami i 12 minutami). W przypadku większości danych stwierdzono prawie idealną zgodność między NLP, a charakterystyką określoną przez człowieka w zakresie identyfikowanych rozpoznań klinicznych, substancji lekowych i liczbowych parametrów echokardiograficznych. Obliczone punktacje w skalach CHA2DS2VASc i HAS-BLED oparte na obu metodach nie różniły się istotnie, a zatem były zbieżne. Najmniej dokładną cechą NLP była identyfikacja dziennej dawki przyjmowanej substancji lecznicze, co w pracy precyzyjnie wyjaśniono. W publikacji podkreślono przyszłe perspektywy wykorzystania technik NLP w przetwarzaniu EDM do badań naukowych. Nakreślone zostały również ograniczenia różnych technik NLP, w kontekście analizy danych klinicznych, w celu przybliżenia tematyki klinicystom nie zaangażowanym w tematykę NLP. Dyskusja w tej pracy wskazuje na szeroką wiedzę Doktoranta w tematyce rozprawy.

Należy podkreślić, że wszystkie prace zostały perfekcyjne przygotowane od strony metodycznej. Wyniki zaprezentowano przejrzyście i wyczerpująco. Dyskusje są poprowadzono w sposób wyważony i zawierają elementy edukacyjne, co ma wartość dodaną, szczególnie istotną wobec ograniczonego rozpowszechnienia wiedzy na ten temat w środowisku medycznym. Uwagę zwracają również bardzo dobrze zidentyfikowane i omówione ograniczenia badań i prac.

Wnioski

Wnioski zostały sformułowane bardzo precyzyjnie. W pełni odpowiadają postawiony celom oraz mają pełne oparcie w otrzymanych wynikach:

- 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.
- Techniki NLP, mogą pozwolić na dokładne i szybkie scharakteryzowanie pacjentów w populacji kardiologicznej, w porównaniu do analizy danych przez człowieka.
- Techniki NLP charakteryzują się określonymi ograniczeniami w kontekście pozyskiwania trafnych, ustrukturyzowanych danych klinicznych z elektronicznej dokumentacji medycznej.
- 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.

Omawiane prace stanowią spójny cykl i źródło cennych, unikalnych informacji o problematyce pozyskiwania danych wysokiej jakości z przystającej dynamicznie dokumentacji elektronicznej. Lektura rozprawy umożliwia szerokie spojrzenia na współczesne metody szybkiego pozyskiwania danych do badań w dziedzinie kardiologii, z podkreśleniem roli technik procesowania tekstu za pomocą autorskiego rozwiązania informatycznego.

Jest to pierwszą w Polsce próbą aplikacji tego rodzaju technik do danych nieustrukturyzowanych zawartych w elektronicznej dokumentacji medycznej, prowadzonej w języku polskim, w celu stworzenia praktycznego narzędzia użytecznego w prowadzeniu badań oraz zautomatyzowanych skal ryzyka w kardiologii.

Dostrzegam również ważny dodatkowy walor tej pracy – uświadamia ona jak istotne jest rzetelne prowadzenie dokumentacji medycznej, która jest i zawsze będzie podstawą postępu medycyny. Dbałość o jakość danych oznacza dbałość o wyniki badań naukowych, a co za tym idzie odpowiedzialność za metody leczenia jakie wdrażamy na ich podstawie.

Pytania i uwagi

Po zapoznaniu się z rozprawą doktorską lek. Cezarego Piotra Maciejewskiego w pierwszej kolejności kieruję do Doktoranta, Promotorów i Zespołu współpracowników gratulacje. Temat jest bardzo ciekawy i ma zastosowanie praktyczne, w związku z czym pozwolę sobie zadać Doktorantowi następujące pytania:

1. *Czy prezentowanr narzędzia analityczne mogłby znaleźć zastosowanie powszechne i jakie kroki należałoby w tym kierunku podjąć?*
2. *Czy takie oprogramowanie wymaga certyfikacji jako wyrób medyczny?*
3. *Czy tego typu oprogramowanie mogłoby integrować dane w czasie rzeczywistym i wspomagać lekarzy w ich bieżącej pracy, np. na przykładzie chorych z migotaniem predsionków – czy możliwe byłaby autmatyczne, w czasie rzeczywistym, generowanie oceny pacjenta w skalach ryzyka/predykcyjnych na podstawie wprowadzanych do EDM danych lub też nawet generowania podpowiedzi jakie dane lekarz powinien uzupełnieć aby ocena na takich skalach była pełnowartościowa?*

W zasadniczej części rozprawy wskazane byłoby by znalazła się informacja, że rozwiązanie powstało w projekcie zrealizowanym w ramach „Inkubatora Innowacyjności 4.0"

Wnioski końcowe

Rozprawę doktorską lek. Cezarego Piotra Maciejewskiego oceniam bardzo pozytywnie. Doktorant wykazał przygotowanie do planowania badań naukowych, analizy ich wyników, wnioskowania na ich podstawie oraz umiejętność realizacji innowacyjnych prac technologicznych. Otrzymane wyniki mają istotne znaczenie dla rozpowszechniania i rozwijania nowych narzędzi analizy dużych zbiorów danych i optymalizacji poprawności wnioskowania na temat stanu zdrowia polskich pacjentów.

Podsumowując stwierdzam, że przedstawiona do oceny rozprawa doktorska lek. Cezarego Piotra Maciejewskiego spełnia warunki określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668). Niniejszym mam zaszczyt przedstawić Wysokiej Radzie Naukowej Dyscypliny Nauk Medycznych WUM **wniosek o dopuszczenie lek. . Cezarego Piotra Maciejewskiego do dalszych etapów przewodu doktorskiego.**

Równocześnie uwzględniając wysoką wartość kliniczną cyklu prac i rozprawy, jej dojrzałość naukową, a w szczególności interdyscyplinarność i innowacyjność (szczegółowe argumenty przedstawiłem powyżej), **wnioskuję o nagrodzenie rozprawy wyróżnieniem.**

prof. dr hab. n. med. i n. o zdr. Paweł Krzesiński

prof. dr hab. n. med. i n. o zdr.
Paweł Krzesiński
specjalista chorób wewnętrznych
i kardiologii
1316559