lek. Cezary Piotr Maciejewski

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.

Rozprawa na stopień doktora nauk medycznych i nauk o zdrowiu
w dyscyplinie nauki medyczne

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

Obrona rozprawy doktorskiej przed Radą Dyscypliny Nauk Medycznych
Warszawskiego Uniwersytetu Medycznego

Warszawa 2024

Słowa kluczowe: elektroniczna dokumentacja medyczna, eksploracja tekstu, procesowanie języka naturalnego, migotanie przedsionków, leczenie przeciwkrzepliwe, skala CHA2DS2VASc, skala HAS-BLED.

Key words: electronic health records (EHR), text-mining, natural language processing (NLP), atrial fibrillation (AF), anticoagulation, CHA2DS2VASc scale, HAS-BLED scale.

Panu Prof. dr hab. n. med. Pawłowi Balsamowi dziękuję za inspirację,
umożliwienie rozwoju naukowego i klinicznego,
oraz cenne uwagi merytoryczne.

Panu Dr hab. n. med. Krzysztofowi Ozierańskiemu dziękuję za życzliwość,
poświęcony czas oraz motywację do pracy naukowej.

Panu Dr hab. n. med. Andrzejowi Cacko za dziękuję za serdeczność,
poświęcony czas oraz cenne uwagi merytoryczne.

Mojej Matce dziękuję za wiarę w moje możliwości
oraz wsparcie na całej mojej drodze edukacji.
Mojemu Ojcu dziękuję za bycie pierwszym wzorem pracowitości.
Mojemu Bratu dziękuję za poczucie humoru w trudnych chwilach.
Mojej Babci dziękuję za wsparcie i życzliwość.
Mojej Żonie dziękuję za zrozumienie oraz codzienne, nieocenione wsparcie.

WYKAZ PUBLIKACJI STANOWIĄCYCH PRACĘ DOKTORSKĄ

Lp.	Artykuł	IF	Punkty MNiSW
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	-	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.	4.900	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.*	4.800	200
	Łącznie	**9.700**	**480**

Spis treści

1. Wykaz stosowanych skrótów

EBM- ang. Evidence Based Medicine – Medycyna oparta na dowodach
EDM- Elektroniczna dokumentacja medyczna
LLMs- ang. Large language models- Duże modele językowe
NFZ- Narodowy Fundusz Zdrowia
NLP- Procesowanie języka naturalnego
RCT- ang. Randomized controlled trial- randomizowane kontrolowane badania kliniczne

2. 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.

3. Streszczenie oraz tytuł rozprawy w języku angielskim

Title: The utilization of modern techniques for textual data analysis in electronic medical records to create tools expediting the acquisition of scientifically valuable structured data and automated risk scales in cardiology.

Summary:

Randomized controlled trials are the main source of knowledge on the effectiveness and safety of medical interventions in evidence-based medicine (EBM) as practiced today. Nevertheless, observational studies remain an important source of knowledge in clinical medicine. Methods such as causal inference are likely to contribute to the growing role of such studies in practicing EBM, especially considering the vast amount of data nowadays being generated during patient care in the „big data" era.

The analysis of data from daily clinical practice collected in the form of electronic medical records (EMRs) can provide important information about populations of cardiology patients, especially groups underrepresented in clinical trials. According to the European Society of Cardiology, this data can be useful for conducting retrospective studies and registries, generating new research hypotheses, monitoring adverse events, and planning and reducing the costs of RCTs. The main limitation of using EMRs for research is that most of the collected information is unstructured data, such as descriptive forms (hospitalization notes, diagnoses lists, discharge recommendations, observations, descriptions of imaging studies), which require time-consuming manual analysis and input into a structured database format by medical staff. This limits the possibility to conduct analyses on large cohorts of patients or to quickly identify patient groups with specific, unique characteristics. Consequently, the opportunity to verify certain research hypotheses in large historical or currently treated patient populations is practically precluded.

Efforts to mitigate these limitations in using EMRs include leveraging structured data created during the provision of healthcare services, such as administrative data. For example, diagnostic and procedural codes (ICD-10, ICD-9) reported to healthcare payers such as the National Health Fund (NHF) in Poland are attractive for research due to the opportunity to quickly obtain data on large patient populations. Unfortunately, their use is fraught with risks of errors in inference. Coded diagnoses and medical procedures may not accurately reflect the actual clinical status of patients because their primary purpose is administrative billing rather than collecting high-quality medical data. These limitations in the context of conducting cardiology research have been identified in various healthcare systems worldwide but were scarcely studied in Poland.

Due to the aforementioned limitations of using administrative data, other methods for effectively obtaining data for scientific research based on EMRs are sought. One such method is the use of natural language processing (NLP) techniques on EMR data to automate the process of collecting clinical data for research purposes. The aim of this dissertation was to: (1) analyze the currently used structured data (diagnostic codes) in the context of cardiology research in Poland to identify limitations of these data, (2) develop alternative solutions using text processing techniques in EMRs to obtain data for cardiology research, and (3) present them in the context of manual data acquisition and clinical characteristics based on administrative data.

In the course of study (publication No. 1), comprehensive comparison of clinical data in medical documentation to data based on ICD-10 codes registered with the NHF in Poland was performed. The study was conducted on a large historical cohort of 3,338 patients diagnosed with atrial fibrillation. Significant discrepancies were identified between patient characteristics based on NHF data compared to medical documentation due to inaccuracies and missing data in

reported ICD-10 codes. Additionally, it was noted that the data held by the payer did not include relevant information for clinical cardiology research such as: detailed diagnoses, information about prescribed medications and dosages, and echocardiographic data.

In the subsequent stage (publication No. 2), an original solution named "AssistMED" was developed, which utilizes NLP techniques within specific types of descriptive EMR data in Polish to automatically obtain a broad clinical profile of large populations of cardiology patients: clinical diagnoses, prescribed medications and dosages, numerical echocardiographic parameters. The assumptions of the solution were described in detail from both clinical and technological perspectives. Quantitative and qualitative analyses of the tool's utilization on a population of 400 anonymized patient records were conducted to comprehensively characterize the limitations resulting from the application of NLP methods. The quantitative analysis showed that using AssistMED for data collection yielded highly convergent results with manual data entry. In the qualitative analysis, the main causes of errors identified included: lack of advanced context analysis (limitations of the NLP techniques), random algorithm errors, typos in the EMRs, and complex medication dosage descriptions.

In publication No. 3, the developed AssistMED tool was used on an anonymized cohort of 10,314 patients from a cardiology department (years 2016-2019). The automated method was compared with manual data verification by humans to characterize a retrospective cohort of patients diagnosed with atrial fibrillation. Very small discrepancies between automatically and manually obtained data were demonstrated, with significantly shorter data acquisition time using the automated method.

In conclusion, although administrative data such as ICD-10 codes are a time-efficient and valuable data source for scientific research, they have significant limitations. Therefore, there is a need to develop other techniques aimed at automating data acquisition from EMRs. It was shown that the use of developed NLP techniques on EMR data can allow for obtaining a broad population profile for cardiology research in a short time and with high compliance in comparison to manual data collection.

4. Wstęp uzasadniający połączenie wskazanych publikacji w jeden cykl, jak i komentujący osiągnięcie na tle dotychczasowego stanu wiedzy.

Choroby sercowo-naczyniowe są jedną z głównych przyczyn zgonów na świecie. Prowadzą do pogorszeniem jakości życia, przyczyniają się do zmniejszenia produktywności poprzez przedwczesne zaprzestanie aktywności zawodowej oraz wiążą się z wysokimi kosztami dla systemów opieki zdrowotnej. W dziedzinie kardiologii prowadzone są liczne badania, wdrażane są kolejne terapie o wykazanej skuteczności, które przyczyniają się do poprawy długości i jakości życia populacji.

Współcześnie praktykowana medycyna oparta na dowodach (EBM) [1] wymaga wykazania skuteczności i bezpieczeństwa terapii w określonej grupie chorych w randomizowanych badaniach klinicznych (RCT-randomized controlled trial) [2]. Badania RCT oraz metaanalizy tych badań dostarczają danych naukowych o najwyższej jakości, ponieważ cechują się wysokim stopniem kompletności danych oraz ograniczają wpływ czynników zakłócających [3]. Interwencje o wykazanej skuteczności w badaniach RCT są wdrażane do standardu postępowania w codziennej praktyce klinicznej z najwyższą rekomendacją w wytycznych postępowania [4]. Ograniczeniem badań randomizowanych jest ich kosztowność, długi czas przeprowadzania oraz selekcjonowanie populacji chorych nie w pełni odzwierciedlających złożoność codziennej praktyki klinicznej [5, 6].

Z tego względu, badania obserwacyjne nadal pozostają bardzo istotnym źródłem wiedzy w medycynie klinicznej. Dostępność ogromnych zasobów danych, rozwój technik data-science oraz postęp w metodologii wnioskowania przyczynowego pozwalają na limitowanie tradycyjnych ograniczeń badań obserwacyjnych [7]. Dokumentacja medyczna jest szeroko wykorzystywana w badaniach naukowych, ponieważ wnioski, które można formułować na podstawie jej analizy dotyczą rzeczywistego postępowania z pacjentem, a przez to odzwierciedlają skuteczność praktyczną interwencji leczniczych oraz codzienną praktykę kliniczną. Badania obserwacyjne mogą dostarczyć informacji o grupach pacjentów rzadziej reprezentowanych w badaniach klinicznych. Ponadto dane obserwacyjne mogą usprawnić rekrutację do badań RCT oraz przyczynić się do lepszego uogólnienia wyników [8].

Grupa ekspercka przy Europejskim Towarzystwie Kardiologicznym (ESC) [9] wskazuje na duży potencjał jaki ma analiza danych tworzonych w trakcie codziennej opieki nad pacjentem i zapisywanej w dokumentacji medycznej w:

- prowadzeniu badań i rejestrów retrospektywnych,
- generowaniu nowych hipotez badawczych,
- monitorowaniu zdarzeń niepożądanych
- planowaniu i zmniejszaniu kosztów prowadzenia badań RCT.

Tradycyjnym ograniczeniem analiz opartych o dokumentację medyczną jest czasochłonny proces prowadzenia bazy danych, który wymaga zaangażowania personelu medycznego w ręczną analizę papierowej dokumentacji medycznej. Coraz więcej podmiotów leczniczych w Polsce, wdraża systemy elektronicznej dokumentacji medycznej. Od początku 2019 roku Ministerstwo Zdrowia wprowadziło obowiązek prowadzenia części dokumentacji medycznej w formie elektronicznej przez wszystkie podmioty publicznego sektora ochrony zdrowia w Polsce. Dostępność dokumentacji w formie elektronicznej, wpłynie na możliwość jej szerszego wykorzystania w badaniach naukowych. Istnieje jednak szereg ograniczeń, do których należy m.in. specyfika danych w elektronicznej dokumentacji medycznej. Dane te dzielą się na tzw. ustrukturyzowane (np. wiek, płeć, wyniki badań laboratoryjnych, jednostki chorobowe i procedury zakodowane za pomocą kodów ICD-10/ICD-9) oraz nieustrukturyzowane (np. dane opisowe). Większość (około 80% danych) w elektronicznej dokumentacji medycznej to informacje nieustrukturyzowane [10].

Wykorzystanie danych ustrukturyzowanych np. kodów ICD-10, może być obarczone ryzykiem błędów we wnioskowaniu, ponieważ główną przesłanką ich raportowania są względy administracyjne (rozliczenie finansowe udzielanych przez podmioty medyczne świadczeń zdrowotnych). W konsekwencji, dane te mogą nie odzwierciedlać precyzyjnie realnego stanu zdrowia badanej populacji, co dokumentowano w różnych systemach ochrony zdrowia [11, 12]. Rozpoznania kliniczne oraz procedury medyczne mogą być niepoprawnie kodowane, ponieważ są postrzegane przez pracowników ochrony zdrowia jako dodatkowe obciążenie administracyjne [13]. Rzadziej kodowane są rozpoznania chorobowe, które nie są niezbędne do celów rozliczeniowych, dlatego dane administracyjne mogą nie zawierać zakodowanych informacji o chorobach współistniejących [14]. Trafność kodów ICD-10 jest też bardzo zmienna, w zależności od lokalnych praktyk kodowania oraz liczby podmiotów administrującym tymi danymi.

Dane administracyjne są często stosowanym źródłem danych do wnioskowania do dużych badań obserwacyjnych w Polsce i na świecie, z uwagi na możliwość zebrania w krótkim czasie informacji o dużej populacji chorych. Jednak ograniczenia i możliwe błędy wynikające z ich zastosowania są niedostatecznie zbadane. W Polsce, Narodowy Fundusz Zdrowia jest głównym płatnikiem za usługi medyczne, stąd dane dotyczące świadczeń zdrowotnych oraz diagnoz klinicznych powinny być względnie kompleksowe. Dotychczas istnieją nieliczne opracowania dotyczące analizy zgodności tych danych, ze stanem faktycznym, w dziedzinie kardiologii w Polsce.

Dane nieustrukturyzowane np. dane w formie opisowej: lista rozpoznań klinicznych, epikryzy lekarskie, badanie podmiotowe i przedmiotowe zalecenia lekarskie czy opisy badań obrazowych zawierają w sobie większość informacji klinicznych. Stanowią efekt oceny klinicznej i syntezy informacji o pacjencie przeprowadzonej przez personel medyczny i są tworzone w celu dokumentacji procesu leczniczego. Takie dane nie mogą zostać poddane analizie statystycznej i wnioskowaniu, dopóki nie zostaną manualnie przeanalizowane i zakodowane w formie bazy danych. Jest to proces bardzo czasochłonny, co w praktyce uniemożliwia wykorzystanie danych z elektronicznej dokumentacji medycznej, do szybkiej weryfikacji hipotez badawczych.

Aby umożliwić szersze wykorzystanie danych nieustrukturyzowanych, z elektronicznej dokumentacji medycznej, niezbędne jest wypracowanie metod ich strukturyzacji. Do takich metod należą techniki procesowania języka naturalnego (NLP), które mogą umożliwić analizowanie zawartości i kodowanie tekstu w notatce, do uporządkowanych terminów medycznych oraz eksportu do bazy danych. Techniki procesowania tekstu notatek medycznych są intensywnie rozwijanym kierunkiem badań na świecie, głównie w odniesieniu do języka angielskiego z uwagi na względną łatwość jego analizy oraz publiczną dostępność dużych zasobów danych. Wśród użytecznych klinicznie przykładów zastosowań procesowania tekstu w medycynie można wymienić:

- śledzenie poważnych zdarzeń medycznych np. epizodów krwawień w trakcie hospitalizacji [15],
- ekstrakcja danych z badania echokardiograficznego [16],
- rozpoznawanie i kodowanie wartości parametrów życiowych takich jak ciśnienie, tętno, liczba oddechów [17].

Ponadto, wykorzystanie technik procesowania języka naturalnego, umożliwia analizę olbrzymich zbiorów danych w celu pozyskania informacji istotnych z punktu widzenia jakości terapii i bezpieczeństwa pacjenta. W jednym badaniu, zastosowanie tych technik, umożliwiło automatyczną analizę rekordów pacjentów w elektronicznej dokumentacji medycznej, w celu wykrycia osób z migotaniem przedsionków[18]. Analiza tekstu umożliwiła rozpoznanie ponad 30% więcej pacjentów z migotaniem przedsionków, niż wskazywałyby na to kody ICD-10, a analiza skal CHA_2DS_2-VASc oraz HAS-BLED w tej grupie pacjentów wykazała, że 13,6% osób z tej grupy miało wskazania do wdrożenia leczenia przeciwkrzepliwego w celu prewencji incydentów niedokrwiennych.

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.

W publikacji nr 1, zostały omówione aktualnie, szeroko stosowane techniki pozyskania charakterystyki klinicznej dużych populacji – kody ICD-10, raportowane do Narodowego Funduszu Zdrowia. Wykonano analizy porównawcze do danych zgromadzonych w trakcie hospitalizacji w tekstowej dokumentacji medycznej, w celu oceny zgodności danych w NFZ, na przykładzie retrospektywnej, zanonimizowanej kohorcie 3338 pacjentów z rozpoznaniem migotania przedsionków. Według mojej wiedzy, jest to pierwsza dostępna analiza, która dostarczyła tak kompleksowej informacji o dokładności danych rozliczeniowych w Polsce w dziedzinie kardiologii. Wartość pracy stanowi przedstawienie danych na przykładzie prawdziwej kohorty chorych, z jednoczasowym pokazaniem wielu chorób, co pozwoliło na praktyczną wizualizację, jak charakterystyka przykładowej grupy badanej zmieniłaby się, gdyby oprzeć ją wyłącznie o dane rozliczeniowe.

Techniki procesowania języka naturalnego, nie były dotąd używane w dziedzinie kardiologii w Polsce, w celu pozyskiwania kompleksowej charakterystyki populacji do badań medycznych z EDM. Doświadczenia zebrane w trakcie tworzenia publikacji 1 (między innymi struktury powiązań rozpoznań chorobowych, analiza poziomu ich

szczegółowości w kontekście potrzeb nauk klinicznych, zidentyfikowane ograniczenia danych ustrukturyzowanych) posłużyły do stworzenia autorskiego algorytmu wykorzystującego techniki procesowania naturalnego (projekt AssistMED) w celu pozyskiwania danych z wybranych typów danych tekstowych w EDM.

W publikacji 2, omówione zostały założenia merytoryczne, zarówno od strony klinicznej jak i technicznej, stojące u podstaw wdrożenia algorytmu. Ponadto, przeprowadzona została ewaluacja wyników algorytmu w porównaniu do analizy analogicznych części EDM przez człowieka. Dokonano podsumowania w formie analiz ilościowych i jakościowych, w celu precyzyjnego opisania ilości błędów popełnianych przez algorytm oraz przyczyn ich popełniania (kategoryzacja sytuacji w których algorytm popełniał błędy).

W publikacji nr 3 przedstawiono uzyskanie dużej kohorty chorych (3030 pacjentów) z wykorzystaniem wypracowanych w toku projektu AssistMED metod automatycznych oraz analizy danych przez człowieka. Przeanalizowano zbieżność danych uzyskanych automatycznie i manualnie, oraz określono czas uzyskania danych w obu metodach, w celu określenia wiarygodności i efektywności czasowej narzędzia opartego o NLP.

Przedstawiana rozprawa doktorska to cykl publikacji oryginalnych kompleksowo dokumentujący ograniczenia danych administracyjnych oraz przedstawiający autorskie rozwiązanie oparte na technikach NLP w celu pozyskania danych naukowych w dziedzinie kardiologii z elektronicznej dokumentacji medycznej.

5. Założenia i cel pracy

5.1. Hipotezy badawcze prac w cyklu:

- Dostępne w polskim systemie ochrony zdrowia dane rozliczeniowe są niewystarczające, w celu uzyskania precyzyjnego i kompleksowego opisu przykładowej kohorty chorych w kardiologii, w porównaniu do danych opisowych gromadzonych w dokumentacji medycznej.
- Wypracowanie narzędzia opartego o NLP pracującego w EDM w celu pozyskania danych do badań naukowych w dziedzinie kardiologii jest możliwe, ale wymaga zaangażowania kompetencji klinicznych i programistycznych w celu wypracowania odpowiednich struktur i zależności między uzyskiwanymi danymi.
- Automatyczna analiza określonych typów danych opisowych w EDM pozwala na pozyskanie danych wysoce zbieżnych z manualną analizą danych przez człowieka, ale w istotnie krótszym czasie.
- Wykorzystanie aktualnych technik NLP do pozyskania z danych z EDM ma ograniczenia skutkujące określonymi typami błędów w identyfikacji.

5.2. Cele ogólne prac w cyklu

Do celów ogólnych pracy należało

- 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.
- 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.

- 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.
- Scharakteryzowanie ograniczeń narzędzia opartego o wykorzystanie NLP w EDM.

5.3. Publikacja 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

W publikacji zbadano dokładność charakterystyki klinicznej pacjentów uzyskiwanej w oparciu o dane rozliczeniowe dostępne w Narodowym Funduszu Zdrowia, w kontekście określania ryzyka niedokrwiennego i krwotocznego kohorty pacjentów z migotaniem przedsionków. Z uwagi na obecność pojedynczego, publicznego płatnika ochrony zdrowia w Polsce - NFZ, pozyskane dane rozliczeniowe powinny w kompleksowy sposób odzwierciedlać stan zdrowia badanej populacji (w wielu krajach system ochrony zdrowia jest rozdzielony na wielu płatników, z których każdy prowadzi swoje bazy danych).

Do porównań wykorzystano historyczną kohortę pacjentów leczonych przeciwkrzepliwie z powodu migotania przedsionków scharakteryzowaną w oparciu o dane z dokumentacji medycznej dwóch ośrodków kardiologicznych. Równolegle pozyskane zostały dane z NFZ (m.in. kody rozpoznań chorobowych ICD-10), dotyczących tej samej grupy pacjentów. Dla kohorty 3338 pacjentów zidentyfikowanych zostało 565521 świadczeń zdrowotnych z przypisanymi im kodami ICD-10 (świadczenia dla poszczególnych jednostek chorobowych dla pojedynczego pacjenta powtarzają w się wielokrotnie- np. kolejne wizyty w poradni z powodu danej choroby).

Wyniki wskazały na istotne statystycznie różnice pomiędzy charakterystyką pacjentów, opartą o analizę dokumentacji medycznej w formie kart hospitalizacji 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 (NFZ = 83% vs. EDM = 100%), jednocześnie zawyżając odsetek pacjentów z innymi chorobami układu krążenia. Wskazywało to na stosunkowo niską wartość predykcyjną dodatnią danych rozliczeniowych, przy zadowalającej wartości predykcyjnej ujemnej. Istotnie wyższe wyniki w skali CHA2DS2VASc (mediana, [Q1-Q3]) (NFZ: 1, [0-2]; vs. EDM: 1, [0-1]; $p < 0,001$) i HAS-BLED (mediana, [Q1-Q3]) (NFZ: 4, [2-6] vs. EDM: 3, [2-5]; $p < 0,001$) zostały obliczone odpowiednio według danych z NFZ, w porównaniu z danymi EDM, co wskazywało na przeszacowywanie ryzyka niedokrwiennego i krwotocznego pacjentów według danych administracyjnych. Ponadto zauważono, że nie istnieją kody ICD-10, które pozwalałyby na odzwierciedlenie niektórych sytuacji klinicznych, niezbędnych do kalkulacji skali ryzyka krwawienia HAS-BLED (np. niekontrolowane nadciśnienie tętnicze, labilny czas protrombinowy). Ponadto dane z NFZ nie zawierały informacji o stosowanych lekach, które są niezbędne do kalkulacji skali HAS-BLED.

W dyskusji przybliżono wyniki identyfikacji poszczególnych chorób z innych systemów ochrony zdrowia, które wskazywały na analogiczne ograniczenia danych rozliczeniowych na świecie, w stosunku do tych zidentyfikowanych w pracy badawczej. Zaobserwowano, że w publikacjach badających jednoczasowo trafność kodów ICD-10 oraz NLP, wykorzystanie tych ostatnich umożliwiało pozyskanie bardziej wiarygodnych danych.

Warto podkreślić, że według mojej wiedzy, publikacja jest jedynym opracowaniem badającym zbieżność danych dotyczących jednostek chorobowych w dokumentacji medycznej i danych rozliczeniowych zaraportowanych płatnikowi w Polsce, w kontekście wykorzystania do badań naukowych w dziedzinie kardiologii. Publikacja przedstawia jednoczasowo wiele jednostek chorobowych. Ma to na celu ogólne przedstawienie, jak różniłaby się charakterystyka pacjentów oparta w pełni o dane administracyjne, w stosunku do opartej na dokumentacji medycznej.

Podsumowując, istnieją 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. Ponadto kody ICD-10, nie były w stanie uchwycić niektórych sytuacji klinicznych. Dane ustrukturyzowane nie zawierały również informacji o stosowanych lekach. Biorąc pod uwagę powyższe wyniki, powinny być poszukiwane nowe sposoby efektywnej klasyfikacji pacjentów np. z użyciem procesowania języka naturalnego w EDM.

5.4. Publikacja 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.

Wykorzystując obserwacje z publikacji nr 1, w kolejnym etapie został stworzony algorytm „AssistMED" wykorzystujący techniki procesowania języka naturalnego, w celu pozyskiwania danych z EDM dotyczących:

- rozpoznań klinicznych,
- substancji leczniczych i dawkowania
- liczbowych parametrów echokardiograficznych.

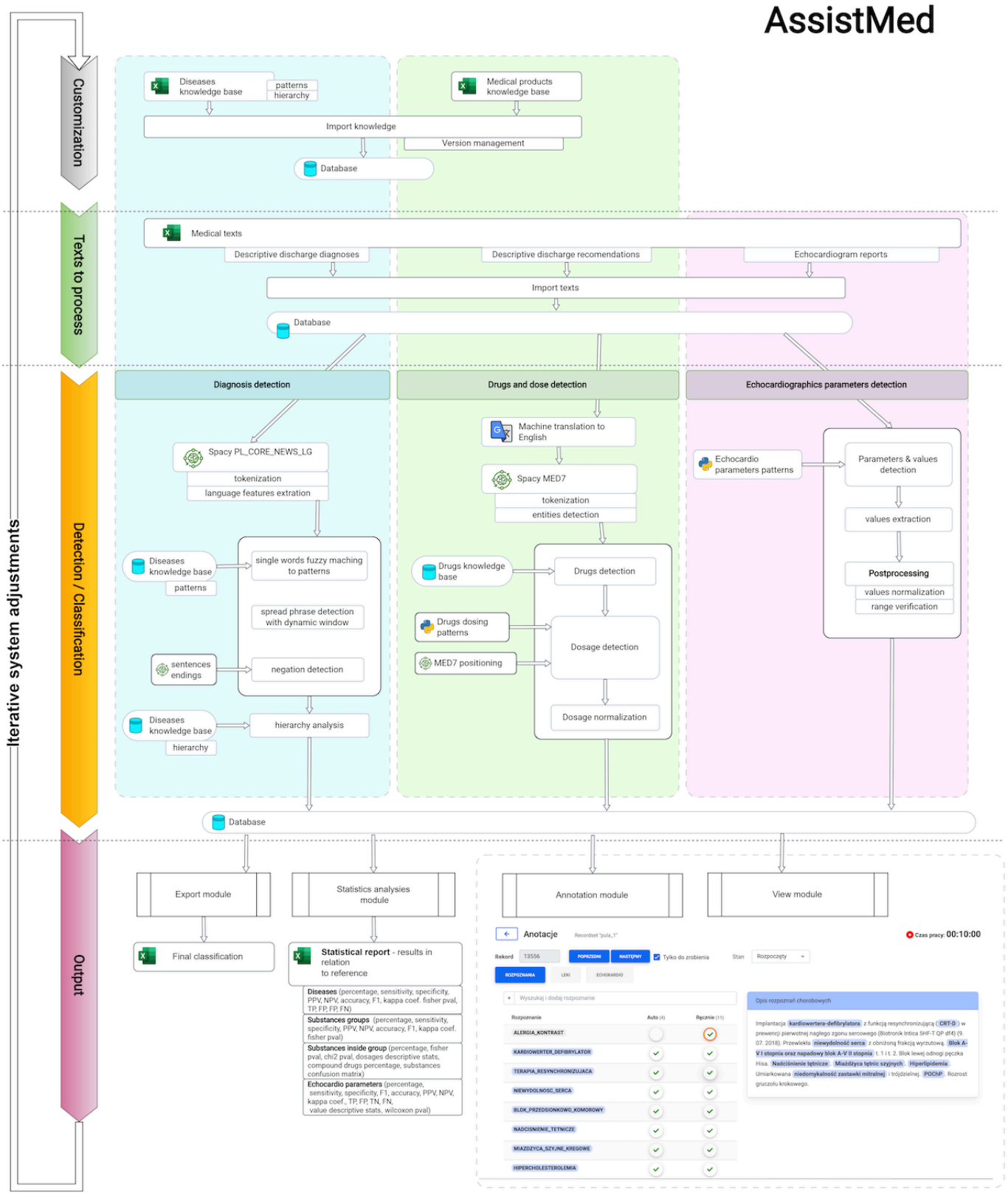
W publikacji 2 szczegółowo przedstawione zostały założenia merytoryczne, proces projektowania oraz aspekty techniczne wdrożenia projektu AssistMED. Rycina 1 przedstawia komponenty algorytmu, a rycina 2 przykład danych tekstowych które są procesowane przez algorytm.

W toku ewaluacji wypracowanego narzędzia, AssistMED przeanalizował obecność 56 stanów klinicznych, leków z 16 grup leków z dawkowaniem i 15 numerycznych parametrów echokardiograficznych na próbie losowych rekordów 400 pacjentów hospitalizowanych na oddziale kardiologii. Co istotne, 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, która potencjalnie może zostać osiągnięta jedynie poprzez zastosowanie najnowocześniejszych rozwiązań technicznych- z zastosowaniem dużych modeli językowych (LLMs - ang. Large language models).

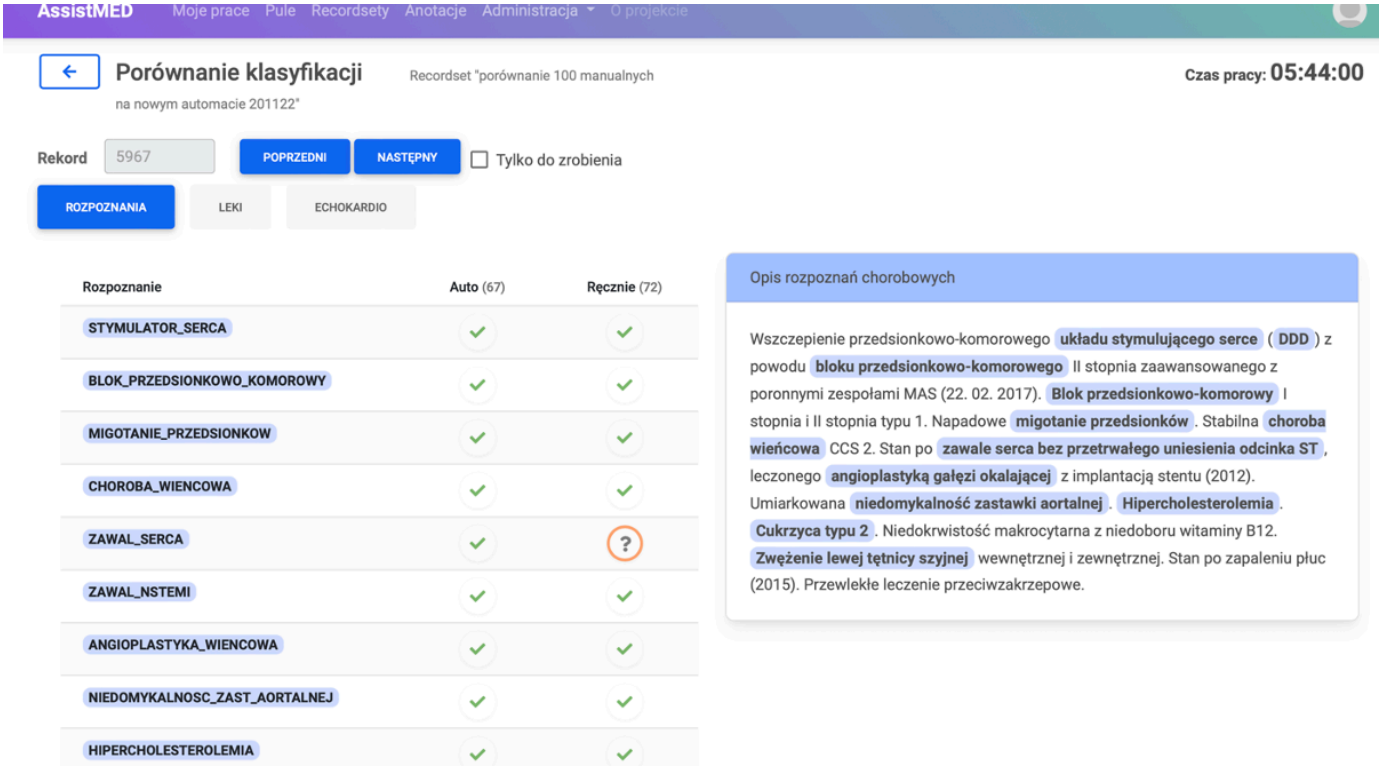
Wartością publikacji było: (1) wykazanie, że pozyskanie wiarygodnej, szerokiej charakterystyki klinicznej z danych opisowych w elektronicznej dokumentacji medycznej, prowadzonej w języku polskim w sposób automatyczny jest możliwe; (2) opisane zostały wyzwania merytoryczne i techniczne przy wdrażaniu algorytmu NLP, pracującego na danych medycznych; (3) scharakteryzowane zostały ograniczenia i wskazane potencjalne sposoby ich zaadresowania w przyszłości, z nakreśleniem konkretnych rozwiązań technicznych, które mogą zostać zastosowane.

Należy podkreślić, że publikację wyróżnia liczba parametrów charakteryzowanych w sposób jednoczesny przez AssistMED. Dotychczasowe publikacje w dziedzinie wykorzystania NLP, na danych tekstowych z EDM na świecie, demonstrują uzyskiwanie danych dla pojedynczej lub kilku diagnoz lub leków lub parametrów echokardiograficznych. Nie prezentują one takich danych jednocześnie w celu prezentacji kompleksowej charakterystyki klinicznej grupy pacjentów. Brak jest również danych, dokumentujące tego rodzaju system w Polsce.

Rycina 1. Komponenty algorytmu AssistMED.



Rycina 2. Przykład danych opisowych procesowanych przez AssistMED.



5.5. Publikacja 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.*

W publikacji nr 3 przedstawiony jest przykład praktycznego użycia narzędzi wypracowanych w ramach projektu AssistMED, w celu pozyskania danych do dużego badania obserwacyjnego pacjentów z migotaniem przedsionków, hospitalizowanych w jednym oddziale kardiologii. Ukazane zostaje jego zastosowanie do pozyskiwania rozległej i 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.

Na wspomnianej kohorcie chorych wykazano wysoką zbieżność wskazań metody automatycznej opartej o NLP z weryfikacją przez człowieka, przy jednocześnie znacznie krótszym czasie pozyskania zbioru danych. Z dostępnych 10314 kart wypisowych zidentyfikowano odpowiednio 3030 i 3029 pacjentów za pomocą analizy przez człowieka i opartego na NLP, co odzwierciedlało 99,93% dokładność NLP w wykrywaniu pacjentów z migotaniem przedsionków. 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). Obliczone punktacje w skalach CHA2DS2VASc i HAS-BLED oparte na obu metodach nie różniły się istotnie (człowiek vs NLP; mediana, IQR, wartość p): 3 (2-5) vs 3 (2-5) p=0,74 i 1 (1-2) vs 1(1-2) p=0,63. 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. Identyfikacja dziennej dawki przyjmowanej substancji leczniczej była najmniej dokładną cechą NLP.

W oparciu o wyniki własne i dane literaturowe, 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. Skonkludowano, że w określonych warunkach, przy użyciu narzędzi NLP w EDM, można wysnuć wnioski na temat charakterystyki kohorty chorych, zbieżne z uzyskiwanymi poprzez analizę przez człowieka, przy znacznym skróceniu czasu jej uzyskania. Wykrywanie dziennej dawki dla niektórych grup lekowych stanowiło największe wyzwanie dla algorytmu.

Warto podkreślić, że przedstawione opracowanie jest jedynym dostępnym, przedstawiającym sposób uzyskania jednoczasowo kompleksowej charakterystyki dużej kohorty chorych, na potrzeby badań w kardiologii, przy użyciu technik NLP w EDM w języku polskim oraz zwalidowanej przez człowieka w celu opisania ograniczeń metody.

6. Kopie opublikowanych prac

6.1. Publikacja 1

International Journal of
Environmental Research and Public Health

MDPI

Article

Administrative Data in Cardiovascular Research—A Comparison of Polish National Health Fund and CRAFT Registry Data

Cezary Maciejewski [1,2], Krzysztof Ozierański [1,*], Mikołaj Basza [3], Piotr Lodziński [1], Andrzej Śliwczyński [4], Leszek Kraj [5], Maciej Janusz Krajsman [6], Jefte Prado Paulino [1], Agata Tymińska [1], Grzegorz Opolski [1], Andrzej Cacko [1,6], Marcin Grabowski [1] and Paweł Balsam [1]

1 1st Chair and Department of Cardiology, Medical University of Warsaw, 02-091 Warszawa, Poland
2 Doctoral School, Medical University of Warsaw, 02-091 Warszawa, Poland
3 Medical University of Silesia in Katowice, 40-055 Katowice, Poland
4 Satellite Campus in Warsaw, University of Humanities and Economics in Lodz, 90-212 Lodz, Poland
5 Department of Molecular Biology, Institute of Genetics and Animal Biotechnology, Polish Academy Science, Postępu 36A, 05-552 Magdalenka, Poland
6 Department of Medical Informatics and Telemedicine, Medical University of Warsaw, 02-091 Warszawa, Poland
* Correspondence: krzysztof.ozieranski@wum.edu.pl

Abstract: (1) Background: Administrative data allows for time- and cost-efficient acquisition of large volumes of individual patient data invaluable for evaluation of the prevalence of diseases and clinical outcomes. The aim of the study was to evaluate the accuracy of data collected from the Polish National Health Fund (NHF), from a researcher's perspective, in regard to a cohort of atrial fibrillation patients. (2) Methods: NHF data regarding atrial fibrillation and common cardiovascular comorbidities was compared with the data collected manually from the individual patients' health records (IHR) collected in the retrospective CRAFT registry (NCT02987062). (3) Results: Data from the NHF underestimated the proportion of patients with AF (NHF = 83% vs. IHR = 100%) while overestimating the proportion of patients with other cardiovascular comorbidities in the cohort. Significantly higher CHA2DS2VASc (Median, [Q1–Q3]) (NHF: 1, [0–2]; vs. IHR: 1, [0–1]; $p < 0.001$) and HAS-BLED (Median, [Q1–Q3]) (NHF: 4, [2–6] vs. IHR: 3, [2–5]; $p < 0.001$) scores were calculated according to NHF in comparison to IHR data, respectively. (4) Conclusions: Clinical researchers should be aware that significant differences between IHR and billing data in cardiovascular research can be observed which should be acknowledged while drawing conclusions from administrative data-based cohorts. Natural Language Processing of IHR could further increase administrative data quality in the future.

Keywords: billing data; administrative data; NHF; NLP; cardiology; AF; atrial fibrillation

Citation: Maciejewski, C.; Ozierański, K.; Basza, M.; Lodziński, P.; Śliwczyński, A.; Kraj, L.; Krajsman, M.J.; Prado Paulino, J.; Tymińska, A.; Opolski, G.; et al. Administrative Data in Cardiovascular Research—A Comparison of Polish National Health Fund and CRAFT Registry Data. *Int. J. Environ. Res. Public Health* **2022**, *19*, 11964. https://doi.org/10.3390/ijerph191911964

Academic Editor: Paul B. Tchounwou

Received: 26 July 2022
Accepted: 13 September 2022
Published: 22 September 2022

1. Introduction

Atrial fibrillation (AF) is a common cardiac arrhythmia affecting 2–4% of adults in the European population [1]. Its increasing prevalence is related to the ageing of modern societies and the presence of other comorbidities (i.e., hypertension, coronary artery disease, heart failure) [1]. Large cohort studies may allow for monitoring of the quality of treatment and patient outcomes in those with AF and possibly lead to new discoveries.

Manual chart review or clinician-driven prospective data collection are regarded as the most accurate methods for clinical research database formation [2]. However, these methods are extremely laborious and time-consuming. They are not feasible for studies that require large cohorts of patients and, therefore, alternatives are being sought.

Administrative (billing) data, most frequently relying on International Classification of Diseases (e.g., ICD-9 and ICD-10) diagnostic codes, are becoming frequently used in

observational clinical research. The analysis of ICD codes allows for time- and cost-efficient acquisition of large volumes of individual patient data. The utilization of this data source may in turn allow for the evaluation of real-life clinical outcomes of patients or generation and initial verification of new hypotheses that otherwise could not be tested due to high costs of prospective registry and randomized studies [3]. Administrative data are used extensively for cardiovascular observational clinical studies especially in Northern America and Scandinavian countries due to the availability of large databases [3–5]. However, a crucial issue is the reliability of the gathered information. Identification of common cardiovascular diseases in administrative datasets has often shown poor sensitivity and was characterized by a high degree of variability in the past [6,7].

Following the international trends, administrative data is also increasingly being used for clinical research in Poland. In Poland, the NHF provides almost universal healthcare coverage in both inpatient and outpatient settings to its citizens. Since it is the single public health fund of the country, its data makes for a very promising opportunities in clinical research.

The current study aimed to evaluate the accuracy of administrative NHF data from a clinical researcher perspective. NHF data is compared against the data collected manually from the individual patients' medical documentation in the retrospective CRAFT registry (NCT02987062) [8]. We evaluated the main disease (AF) and several common comorbidities. To the authors' best knowledge this is the first study of its kind, being based on Polish administrative data and one of the few studies simultaneously evaluating several cardiovascular comorbidities, thus broadening perspectives on the topic.

2. Materials and Methods

Due to the retrospective character of the study, the approval of a local ethics committee and patient-provided written informed consent were waived.

2.1. Individual Health Record (IHR)—Data Obtained through Manual Chart Review

The current study is based on the cohort of patients collected in the MultiCenter expeRience in AFib patients treated with oral anticoagulation registry (CRAFT NCT02987062). This was a retrospective observational cohort study that included consecutive patients aged $\geq$18 years, with a diagnosis of AF treated with anticoagulants and hospitalized between 2011–2016 at one academic and one district hospital in Poland. Details about the study design and main results have been reported elsewhere [9]. Case ascertainment of diseases within medical charts was based on: list of discharge diagnoses, hospitalization summary, discharge recommendations and laboratory tests results. Participants with valvular AF were excluded from the analysis due to difficulty in selection of the optimal ICD-10 codes constellation for this clinical diagnosis.

2.2. National Health Fund (NHF)—Administrative Data

Unidentified billing data on medical services were acquired from the Polish National Health Fund. NHF provides health care for Polish citizens, with an enrollment rate of approximately 94% of the Polish population. NHF gathers data about medical services that it finances, e.g., exact dates of provision, voivodeship (similar to province), setting (emergency department, inpatient and outpatient), primary diagnosis (ICD-10 code- each medical service has 1 primary diagnosis assigned), procedures (ICD-9 code). The primary diagnosis dictates the need for treatment and/or diagnostic tests and is mainly responsible for the use of resources.

We established the list of ICD-10 codes that served as proxies for actual diagnoses evaluated in the CRAFT study. The set of ICD-10 codes was identified through agreement of two physicians after the analysis of the ICD-10 textbook and is presented in the supplementary material (Supplementary Table S1). These codes were utilized in order to obtain clinical characteristics of patients according to NHF data at the time of the CRAFT study data collection. We analyzed the entire medical history (all types of medical services) registered

in the NHF database before and until 30 days after discharge from the hospital at the time of inclusion in the CRAFT registry. We allowed for this 30-day period after hospitalization in order to register additional ICD-10 codes that were likely related to the hospitalization. We decided that such an approach might allow for detection of additional diagnoses acquired from referrals to outpatient care recommended by the treating physician at discharge from the hospital (and thus increase the sensitivity of disease detection). The total number of medical services with assigned ICD-10 codes for this cohort of 3338 patients was 565,521.

2.3. CHA2DS2VASc and HASBLED Scores

With regard to CHA2DS2VASc score [10], proxies for all components of the scale could be identified in NHF data.

HASBLED scale was calculated only for data available for evaluation using ICD-10 codes therefore waiving: uncontrolled hypertension, labile prothrombin time, concomitant use of non-steroid anti-inflammatory drugs and antiplatelets. This resulted in a maximum possible score of 6 out of 9 total points in the scale (in this regard, the same calculation method was utilized for IHR and NHF data). Additionally, one component of the NHF-based HASBLED score, "history of severe bleeding", was analyzed only in the emergency department or inpatient registered billing data in order to identify clinically significant bleeding. Renal disease and liver disease was considered positive according to NHF if any of the selected ICD-10 codes for the respective diseases was present.

2.4. Statistical Analysis

In all analyses, IHR data was treated as a reference for NHF data. The results were presented as medians and quartiles for continuous variables and as frequencies and percentages for categorical and ordinal variables. The frequencies of the categorical and ordinal variables were compared with Fisher's exact test and continuous variables by Mann–Whitney U test respectively. p value below 0.05 was considered significant for all tests. All tests were two-tailed.

Sensitivity, specificity, PPV (Positive Predictive Value) and NPV (Negative Predictive Value) were calculated for NHF identified diseases.

Inter-rater reliability between IHR and NHF data with regard to reported diagnoses was analyzed through calculation of Cohen's Kappa coefficient. The results of this statistic should be analyzed as follows: ≤ 0 as indicating no agreement between analyzed data sources; 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

Statistical analyses and all calculations were performed using R software, version 3.6.2 (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1. Study Population

The final dataset for analysis consisted of records of 3338 patients with both manually collected data from IHR and NHF data collected through a set of ICD-10 codes detection (Supplementary Table S1). A patient flow diagram is presented in Figure 1. The entire available database consisted of 3427 patients; 89 patients were excluded from the current analysis due to valvular AF diagnosis. For the remaining 3338 records, successful matching with administrative data was achieved.

3.2. IHR vs. NHF Data

Table 1 presents the comparison of the entire cohort between IHR and NHF with respective statistics. IHR data is treated as a reference. In all diagnoses there were significant differences present between IHR and NHF. In general, NHF data had a propensity to identify more patients with the respective diagnosis than IHR. NHF underestimated the proportion of AF (all the patients in the CRAFT registry had confirmed diagnosis of AF) and CKD in the cohort while overestimating the proportion of patients with other

conditions. The highest sensitivity and PPV was present for CHA2DS2VASc for guideline-recommended anticoagulation use (class I recommendation for anticoagulation use in AF-2 points for men and 3 points for woman in CHA2DS2VASc score), hypertension and atherosclerosis. The highest specificity was present for liver disease, smoking, severe bleeding and alcohol consumption. The highest NPV was present for alcohol, CKD for HASBLED and HASBLED $\geq$ 3. In the analysis of inter-rater reliability (Cohens kappa) for most diagnoses, there was a fair and slight agreement between IHR and NHF data. The highest agreement was noted for diabetes and prediabetic conditions (moderate agreement) and the lowest for smoking history (none to slight agreement).

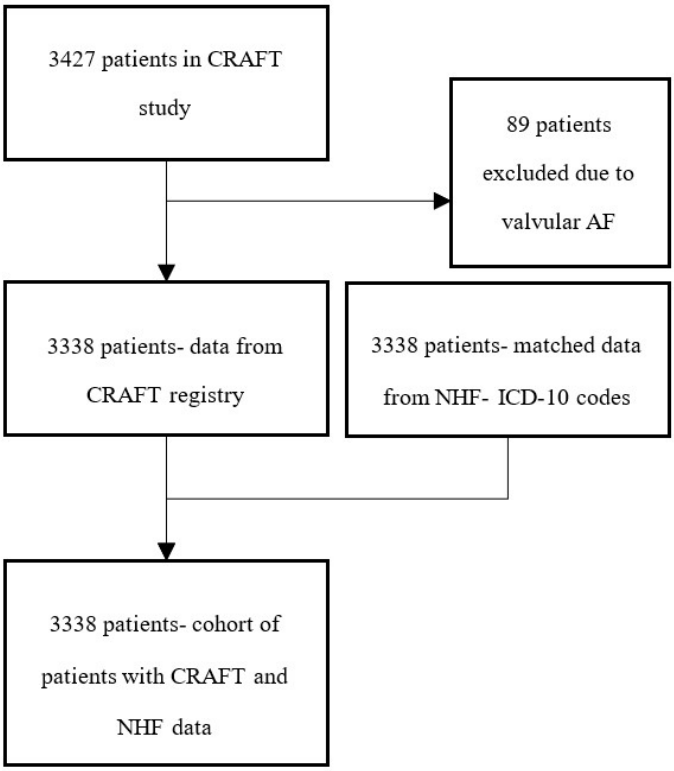


Figure 1. Patient-flow diagram in the study.

Table 1. Comparison between IHR and NHF data. IHR data is treated as reference.

Condition	IHR N, % (CI)	NHF N, % (CI)	*p*-Value	Sensitivity	Specificity	PPV	NPV	Accuracy	Cohen's Kappa *
AF	3338 100% (99.9–100.0)	2766 83% (81.5–84.1)	**<0.001**	0.83	-	-	-	-	-
Severe bleeding	255/3336 7.6% (6.8–8.6)	409 12.3% (11.2–13.4)	**<0.001**	0.24	0.89	0.15	0.93	0.84	0.01

Table 1. *Cont.*

Condition	IHR N, % (CI)	NHF N, % (CI)	*p*-Value	Sensitivity	Specificity	PPV	NPV	Accuracy	Cohen's Kappa *
Alcohol consumption	33/3330 1% (0.7–1.4)	377 11.3% (10.3–12.4)	**<0.001**	0.48	0.89	0.04	0.99	0.88	0.06
CKD for HASBLED	94/3325 2.8% (2.3–3.4)	557 16.7% (15.5–18)	**<0.001**	0.56	0.84	0.1	0.99	0.84	0.12
CKD	706/3325 21.2% (19.9–22.7)	557 16.7% (15.5–18)	**<0.001**	0.34	0.88	0.43	0.83	0.76	0.23
Liver disease	80/3148 2.5% (2–3.2)	346 10.4% (9.4–11.4)	**<0.001**	0.15	0.90	0.04	0.98	0.88	0.02
HF	1207/3333 36.2% (34.6–37.9)	1823 54.6% (52.9–56.3)	**<0.001**	0.82	0.61	0.55	0.86	0.69	0.39
Hypertension	2389/3334 71.7% (70.1–73.2)	2768 82.9% (81.2–84.2)	**<0.001**	0.89	0.32	0.77	0.53	0.73	0.23
Diabetes and prediabetic conditions	874/3325 26.3% (25–27.8)	1108 33.2% (32–34.8)	**<0.001**	0.79	0.83	0.63	0.92	0.82	0.58
Stroke/TIA/ other thromboembolic events	430/3330 12.9% (11.8–14.1)	850 25.5% (24–27)	**<0.001**	0.69	0.81	0.35	0.95	0.79	0.35
Atherosclerosis	1430 42.8% (41.2–44.5)	2390 71.6% (70–73.1)	**<0.001**	0.88	0.40	0.53	0.83	0.61	0.26
CAD	1386 41.5% (40–43.2)	2298 68.9% (67.3–70.4)	**<0.001**	0.86	0.43	0.52	0.81	0.61	0.26
COPD	293/3333 8.8% (7.8–9.8)	735 22% (20.6–23.5)	**<0.001**	0.71	0.83	0.28	0.97	0.82	0.32
Smoking history	175/3328 5.3% (4.6–6.1)	326 9.8% (8.8–10.8)	**<0.001**	0.10	0.90	0.06	0.95	0.86	0.004
HASBLED $\geq$ 3	86/3124 2.8% (2.2–3.4)	487 14.6% (13.4–15.8)	**<0.001**	0.38	0.86	0.07	0.98	0.85	0.08
CHA2DS2VASc for recommended anticoagulation	2390/3316 72.1% (70.5–73.6)	2816 84.4% (83.1–85.6)	**<0.001**	0.96	0.44	0.82	0.79	0.81	0.46

CI—confidence interval; AF—Atrial Fibrillation; CKD for HASBLED—dialysis, transplant, Cr > 2.26 mg/dL or >200 µmol/L; CKD—any evidence of chronic kidney disease; HF—heart failure; TIA—transient ischemic attack; CAD—coronary artery disease; COPD—chronic obstructive pulmonary disease. * Cohen's kappa statistic interpretation: $\leq$0 as indicating no agreement between analyzed data sources; 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Numbers after slash "/" refer to available number of cases if there is missing data. Statistically significant differences are marked as **bolded**.

CHA2DS2VASc and HASBLED scores data are presented in several ways which might be confusing to the reader at first but provide in-depth insight to the data gathered in the study. In general, CHA2DS2VASc and HASBLED scores were significantly higher according to NHF data (Table 2). Table 1 presents a comparison of the percentage of patients fulfilling the criteria for class I recommendation for anticoagulation use in AF (2 points for men and 3 points for woman in CHA2DS2VASc score) [1]. NHF data identified more patients fulfilling the criterion. Similarly, NHF identified more patients with HASBLED of $\geq$3 which

is referred to as a population with a high risk of bleeding in the current guidelines [1]. Figures 2 and 3 show the comparison of distribution of CHA2DS2VASc and HASBLED scores between IHR and NHF. For both scales, a clear tendency towards higher scoring in the NHF databank is visible. Additionally, confusion matrices for CHA2DS2VASc and HASBLED scores are available in the supplementary materials providing in depth insight to the data (Supplementary Tables S2–S5).

Table 2. Comparison of HASBLED and CHA2DS2VASc scores in IHR and NHF data.

	IHR Median [Q1–Q3]	**NHF Median [Q1–Q3]**	***p* Value**
HASBLED	1 [0–1] *3124*	1 [0–2]	**<0.001**
CHA2DS2VASc	3 [2–5] *3316*	4 [2–6]	**<0.001**

Q1, Q3—1st and 3rd quartile; Numbers after slash "/" refer to available number of cases if there is missing data. Statistically significant differences are marked as **bolded**. Numbers in *italics* refer to available number of cases with complete data for respective scale calculation.

NHF data had high sensitivity, moderate PPV, low specificity and low NPV with regard to identification of patients with class I anticoagulation recommendation. However, this result should be analyzed with caution as the criterion for CRAFT registry inclusion was AF and current anticoagulation intake; therefore, the major proportion of patients fulfill the indication for chronic anticoagulation due to AF with only a minority having transient indication. This biases the results towards high PPV and low NPV [11].

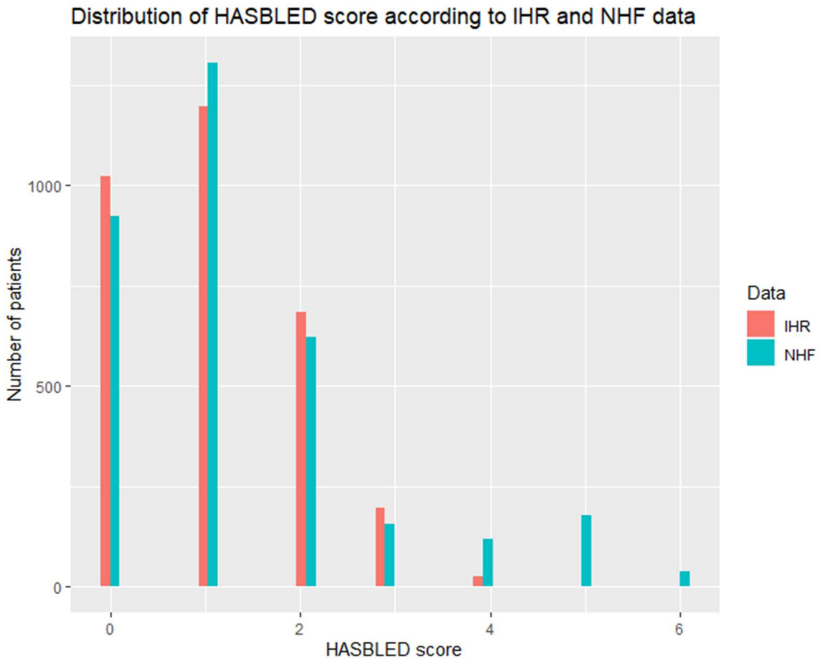


Figure 2. Distribution of HASBLED score within the cohort according to IHR and NHF data.

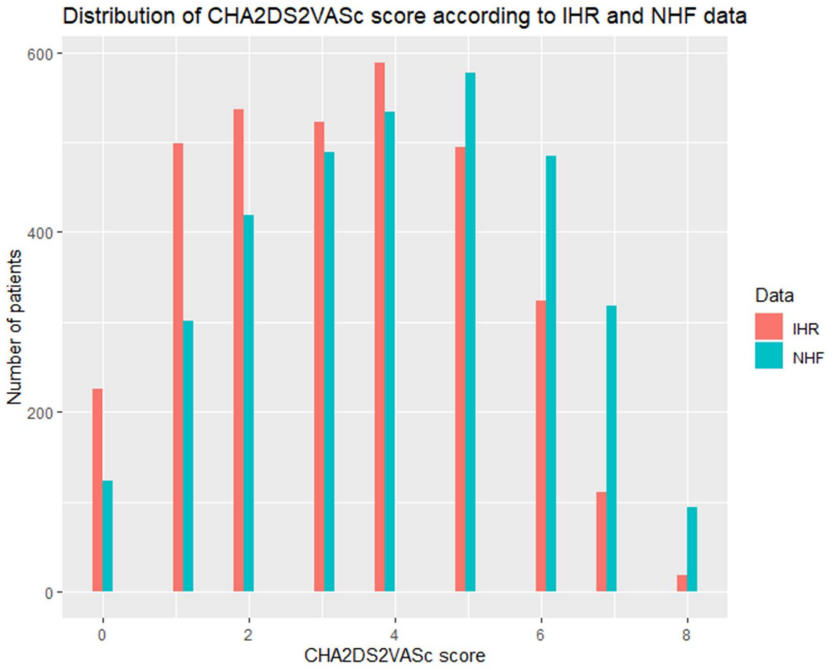


Figure 3. Distribution of CHA2DS2VASc score within the cohort according to IHR and NHF data.

With regard to identification of the population at high risk of bleeding (HASBLED $\geq$ 3 points), NHF had low sensitivity, low PPV, moderate specificity and high NPV. In this regard, the results are more informative, as the design of CRAFT registry as such did not greatly influence the distribution of HASBLED components.

Table 3 presents the AF patients cohort according to NHF data in comparison with the IHR based cohort. This analysis is presented in order to inform the readers performing clinical studies how utilization of solely administrative data could influence the shape of the final AF patients' cohort. The NHF based cohort is smaller due to 572 unidentified AF cases. Statistically significant differences in regards to all analyzed comorbidities are present and the NHF based cohort appears to be more burdened in general.

Table 3. Summary statistics of the cohort of AF patients in IHR and NHF datasets.

Condition	IHR N = 3338 N, % (CI)	NHF N = 2766 N, % (CI)	*p*-Value
Severe bleeding	255/3336 7.6% (6.8–8.6)	378 13.7% (12.4–15)	**<0.001**
Alcohol consumption	33/3330 1% (0.7–1.4)	360 13% (11.8–14.3)	**<0.001**
CKD for HASBLED	94/3325 2.8% (2.3–3.4)	524 18.9% (17.5–20.5)	**<0.001**

Table 3. *Cont.*

Condition	IHR N = 3338 N, % (CI)	NHF N = 2766 N, % (CI)	*p*-Value
CKD	706/3325 21.2% (19.9–22.7)	524 18.9% (17.5–20.5)	**0.03**
Liver disease	80/3148 2.5% (2–3.2)	343 12.4% (11.2–13.7)	**<0.001**
HF	1207/3333 36.2% (34.6–37.9)	1576 57% (55.1–59)	**<0.001**
Hypertension	2389/3334 71.7% (70.1–73.2)	2408 87% (85.8–88)	**<0.001**
Diabetes	874/3325 26.3% (25–27.8)	951 34.4% (32.6–36.2)	**<0.001**
Stroke/TIA/other thromboembolic events	430/3330 12.9% (11.8–14.1)	758 27.4% (25.8–29.1)	**<0.001**
Atherosclerosis	1430 42.8% (41.2–44.5)	2066 74.7% (73–76.3)	**<0.001**
COPD	293/3333 8.8% (7.8–9.8)	671 24.3% (22.7–25.9)	**<0.001**
CAD	1386 41.5% (40–43.2)	1998 72.2% (70.5–73.9)	**<0.001**
Smoking history	175/3328 5.3% (4.6–6.1)	323 11.7% (10.5–13)	**<0.001**
HASBLED $\geq$ 3	86/3124 2.8% (2.2–3.4)	470 17% (15.6–18.4)	**<0.001**
HASBLED, median [Q1–Q3]	1 [0–1]	1 [0–2]	**<0.001**
CHA2DS2VASc for recommended anticoagulation	2390/3316 72.1% (70.5–73.6)	2364 85.5% (84.1–86.7)	**<0.001**
CHA2DS2VASc, median [Q1–Q3]	3.000 [2–5]	4.0 [3–6]	**<0.001**

CI—confidence interval; Q1,Q3—1st and 3rd quartile; CKD for HASBLED—dialysis, transplant, Cr > 2.26 mg/dL or >200 µmol/L; CKD—any evidence of chronic kidney disease; HF—heart failure; TIA—transient ischemic attack; COPD—chronic obstructive pulmonary disease; CAD—coronary artery disease; Numbers after slash "/" refer to available number of cases if there is missing data. Statistically significant differences are marked as **bolded**.

4. Discussion

In general, NHF data tended to have relatively low PPV values, indicating that often there are patients classified as having a certain disease according to billing data who do not have it according to individual health records. At the same time, NHF data in most cases showed reasonable NPV. Therefore, if no information about a certain disease is present in the administrative data, then it is very likely the individual does not have it. The results of the performed study suggest that diagnoses collected in administrative data may carry

a varying degree of both under-coding (not registering ICD-10 code when the disease is present, which decreases sensitivity and NPV) and over-coding (registering ICD-10 code when the disease is absent, which decreases specificity and PPV). From the authors' own everyday clinical experience one of the situations when undercoding may arise is when an important diagnosis is not expressed as a corresponding ICD-10 code (even though it is clearly stated in the discharge summary). On the other hand, overcoding may, for instance, emerge when the patient with suspicion of a certain disease is referred to the specialist for evaluation with ICD-10 code of the suspected disease already assigned (not the code expressing suspicion of the disease as should be done)—in such a scenario, even if the disease is excluded after diagnostic process, the ICD-10 diagnostic code will be at least once registered in the patient's billing data. Both the above-mentioned situations may pose a challenge in designing an observational clinical study based on billing data. The results provided suggest that every clinical condition is defined in the billing data by its own individual qualities and this has to be acknowledged by researchers performing clinical studies utilizing administrative data.

Multiple other studies have evaluated billing data for different cardiovascular diseases and reported conflicting results. In the subsequent paragraphs we will summarize key results of these studies to provide a broader perspective on the topic and present our results in light of other publications. Most of them evaluated the administrative data for one or only a few diseases at a time. For clarity, key statistics in the discussion will be always presented in the following order: sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV). Importantly, not all studies provided all of the above-mentioned metrics.

Yao et al. [12] performed a systematic review evaluating the accuracy of AF detection in administrative data that included 24 studies utilizing data from different countries. The pooled estimates were: Se: 80% (95% CI 72–86%); Sp: 98% (96–99%); PPV 88% (82–94%); NPV 97% (94–99%). Authors concluded that billing data may fail to identify a significant proportion of patients with AF and this may affect estimates of quality of care and prognosis in this patient group. In another study on AF, authors evaluated the impact of different strategies for automatic detection of AF in both administrative and electronic health records in USA [13]. Administrative data based on AF diagnosis had a sensitivity of 88%. The utilization of the model employing Natural Language Processing (NLP) of IHR (textual data in electronic health records) detected an additional 22% of patients with AF. The highest predictional value of the presence of AF were achieved for models using a combination of ICD-10 and NLP of individual patients' electronic health records (EHR). Through a series of simulations with different cohort determination methods (administrative data only, NLP of EHR only, combination of administrative and NLP data with trained machine learning models) it was found that the final number of AF patients that would be included in the cohort could vary by an absolute range of up to 30%, depending on which method had been used for cohort detection. The sensitivity of AF detection based on administrative data in our study corresponds to that of the abovementioned studies.

In a meta-analysis of 11 studies evaluating the accuracy of heart failure diagnosis in administrative databases the calculated statistics were as follows: pooled sensitivity 75.3% (95% CI: 74.7–75.9); pooled specificity 96.8% (95% CI: 96.8–96.9); PPV $\geq$87% in the majority of studies [14]. Our cohort displayed similar sensitivity and somewhat lower specificity and PPV.

So L. et al. [15] evaluated the accuracy of identification of acute myocardial infarction and comorbidities within Canadian administrative dataset with the following results: HF (Se: 80.0%; Sp: 97.8%; PPV: 93.6%; NPV: 92.5%), CKD (Se: 72.2%; Sp: 98.3%; PPV: 81.3%; NPV: 97.2%); diabetes (Se: 66.7%; Sp: 98.9%; PPV: 83.3%; NPV: 97.2%). Xu et al. [16] performed an analysis of different algorithms including relying solely on ICD-10 codes from Canadian dataset in detection of HF (Se: 60.0; Sp: 99.1; PPV: 93.2; NPV 92.1). The researchers also tested algorithms relying on keyword searches, text mining, and a machine learning model trained with combination of text mining and ICD-10 codes, achieving the

highest accuracy with the latter. In another study performed on a Canadian cohort, the following statistics were found: CKD (PPV: 94.3%); COPD (Se: 60.9%; Sp: 94.5%; PPV 84.3%); diabetes (Se: 85.9%; Sp: 97.2%; PPV: 96.2%); liver disease (Se: 13.3%; Sp: 99.7%; PPV: 25.0%); CKD (Se: 50.2%; Sp: 96.6%; PPV: 71.6%). The authors concluded that the prevalence of heart failure and common comorbidities were underestimated in administrative data. Our study differed in that we utilized not only ICD-10 codes related to hospitalization but also to outpatient visits which could influence the final results. The results display similar limitations (e.g., very comparable, poor results in terms of liver disease detection in comparison to our cohort).

The detection of bleeding episodes was evaluated in the study by Joos et al. [17] in USA administrative data. In this study, authors examined charts of patients treated with anticoagulants who were admitted to the hospital. Presence of bleeding related ICD-10 code in any diagnosis position was deemed as positive for bleeding. The results were as follows: Se: 91.4%; Sp: 90.2%; PPV: 52.5%; NPV: 98.9%. The authors concluded that due to a high number of false positive rates, ICD-10 codes should not be used for identifying bleeding complications without confirmatory chart review.

Chang et al. [18] performed an analysis of the association of ICD-9 billing codes with actual diagnoses in the Paul Coverdell National Acute Stroke Program (PCNASP) database and demonstrated high agreement between this registry and administrative data; the Cohen's kappa coefficient was above 0.9 (almost perfect agreement). The calculated Cohen's kappa in our study was significantly lower—0.35 (fair agreement). It should be noted, however, that data was analyzed only for hospitals participating in the PCNASP program, therefore likely putting a higher emphasis on the entire care process (including coding accuracy) than non-participating institutions. This may limit the possibility of generalization of these results to the entire healthcare system.

Our study is unique in that we evaluated the main disease (atrial fibrillation) and several common cardiovascular comorbidities simultaneously which was rarely done in prior publications. This offers the reader a comprehensive view on the topic. The results of our study show that characteristics of patients based solely on administrative data may differ from that collected from IHR through manual chart review. Our results and prior evidence cited in the discussion give an important insight into the use of administrative data in cardiovascular research. The final cohort of AF patients based on NHF would be significantly smaller and in general more burdened than that obtained through IHR analysis. These observations are consistent with prior evidence available from other healthcare systems. Clinical researchers should therefore be aware of potential limitations of studies that utilize billing data as the only source of information for diseases and outcome determinations.

The authors believe that administrative data, despite limitations, is an invaluable tool in the arsenal of methodologies that a clinical researcher can utilize in cardiovascular studies. Continuous progress should be made to augment the accuracy of administrative data in order to further expand its use in cardiovascular research. Text mining and natural language processing (NLP) leverages the unstructured narrative from routine care and is another option for identifying patient cohorts. Future efforts should probably focus on increasing the usage of NLP of textual IHR data and artificial intelligence algorithms on top of the analyzed textual data and administrative data which—as pointed out in the discussion—have so far provided promising results. Such efforts may increase the data quality in cardiovascular studies [16]. Such developments may be carried out shortly on a large scale in Poland since the NHF is transitioning to a universal, central electronic documentation platform that is responsible for gathering various types of textual data (discharge summaries, discharge recommendations) and laboratory tests results related to a patient's given healthcare encounter. This abundance of data, if used efficiently through a combination of ICD codes analysis augmented with text processing and laboratory examinations, may provide researchers with the tools needed for efficient conduct of large, real-world data analysis based observational studies grounded in superior data quality.

Notably, availability of the mentioned data types could have substantially minimized most of the limitations we faced during conduct of the present study.

Limitations

In this study, the manual chart review of patients' health records was a reference. This is a limitation since the CRAFT registry was collected retrospectively from the patients' health records and therefore carries inherent limitations of such study design, e.g., missing data. Additionally, underappreciation of certain diseases in IHR data is possible, since the information regarding certain diseases included in the discharge summary is often based on medical history taken from the patient which may be a subject for recall bias. These two drawbacks of our IHR-reference could falsely decrease specificity and PPV of the billing data. Although we consider our manually reviewed IHR data to be of high quality, the true gold-standard would require prospective data collection with source data regarding disease diagnosis verification which could prevent errors resulting from data loss and recall bias.

Secondly, we utilized only the main-diagnosis ICD-10 code as we did not have access to secondary diagnoses ICD-10 codes gathered by NHF which could also affect our results by decreasing sensitivity and NPV of administrative databank. This drawback may be mitigated by the fact that with a databank as big as the NHF, even if one disease is not coded by one provider, another one will likely introduce the code for it at some point of patient's disease course and it will eventually become evident. However, it cannot be precluded that utilization of secondary diagnosis codes would bring some degree of improvement in the detection of diagnoses; therefore, future studies should aim for their inclusion.

The third limitation is that the exact HASBLED score could not be calculated due to an inability to design adequate proxies for all of its components in administrative data; some of the factors of the scale were simply omitted (the reason for 6 points maximum) as described in the methods section. The others, namely renal disease, liver disease and alcohol use, are very precisely defined in the HASBLED scale (including laboratory thresholds). This level of detail could not be well reflected with a set of ICD-10 codes and likely leads to overestimation of prevalence. This partially explains why significantly more patients are scored for renal disease, liver disease and alcohol use in administrative claims; thus serves as a possible justification for the overall higher HASBLED scores according to NHF than IHR data.

Lastly, our study cannot be broadly generalized; the results are applicable only to NHF data as administrative data accuracy can vary widely, depending not only upon the country and region but also the time period. In Poland, continuous efforts are made by the NHF to increase coding accuracy. As this study concerns data up until the year 2016, many things could have changed and likely for better since then.

5. Conclusions

In the present study we evaluated for the first time the accuracy of administrative NHF data for detection of common cardiovascular comorbidities. Although billing databanks remain an invaluable data source, clinical researchers should be aware of their potential limitations as described in the study. Future efforts should probably focus on implementation of Natural Language Processing of individual health records which could further increase data accuracy.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijerph191911964/s1, Table S1: ICD-10 codes of clinical diagnoses analyzed in the study; Table S2: Confusion matrix for HASBLED score; Table S3: Confusion matrix for HASBLED score according to $\geq$3 points cutoff—high risk of bleeding according to 2020 ESC AF guidelines; Table S4: Confusion matrix for CHA2DS2VASc score; Table S5: Confusion matrix for CHA2DS2VASc score according to $\geq$2 points for men and $\geq$3 points for woman cutoff—class I recommendation for chronic anticoagulation in atrial fibrillation according to 2020 ESC AF guidelines.

Author Contributions: Conceptualization, C.M.; methodology, C.M., K.O., M.B., A.Ś. and M.G.; formal analysis, C.M.; investigation, C.M.; data curation, C.M., A.Ś., L.K., P.L. and A.T.; writing—original draft preparation, C.M., K.O. and M.B.; writing—review and editing, P.L., L.K., M.J.K., J.P.P., A.T., G.O., A.C., M.G. and P.B.; visualization, C.M. and K.O.; supervision, K.O., A.C., M.G. and P.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Due to the retrospective character of the study, the approval of a local ethics committee was waived.

Informed Consent Statement: Due to the retrospective character of the study based on anonymized data, patient provided written informed consent was waived.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hindricks, G.; Potpara, T.; Dagres, N.; Arbelo, E.; Bax, J.J.; Blomström-Lundqvist, C.; Boriani, G.; Castella, M.; Dan, G.A.; Dilaveris, P.E.; et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). *Eur. Heart J.* **2021**, *42*, 373–498. [CrossRef] [PubMed]
2. Lee, D.S.; Donovan, L.; Austin, P.C.; Gong, Y.; Liu, P.P.; Rouleau, J.L.; Tu, J.V. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Med. Care* **2005**, *43*, 182–188. [CrossRef] [PubMed]
3. Lip, G.Y.H.; Keshishian, A.; Li, X.; Hamilton, M.; Masseria, C.; Gupta, K.; Luo, X.; Mardekian, J.; Friend, K.; Nadkarni, A.; et al. Effectiveness and Safety of Oral Anticoagulants Among Nonvalvular Atrial Fibrillation Patients. *Stroke* **2018**, *49*, 2933–2944. [CrossRef] [PubMed]
4. Ray, W.A.; Chung, C.P.; Murray, K.T.; Smalley, W.E.; Daugherty, J.R.; Dupont, W.D.; Stein, C.M. Association of Oral Anticoagulants and Proton Pump Inhibitor Cotherapy With Hospitalization for Upper Gastrointestinal Tract Bleeding. *JAMA* **2018**, *320*, 2221–2230. [CrossRef]
5. Schmidt, M.; Schmidt, S.A.J.; Sandegaard, J.L.; Ehrenstein, V.; Pedersen, L.; Sørensen, H.T. The Danish National Patient Registry: A review of content, data quality, and research potential. *Clin. Epidemiol.* **2015**, *7*, 449–490. [CrossRef] [PubMed]
6. Quach, S.; Blais, C.; Quan, H. Administrative data have high variation in validity for recording heart failure. *Can. J. Cardiol.* **2010**, *26*, 306–312. [CrossRef]
7. Kaspar, M.; Fette, G.; Güder, G.; Seidlmayer, L.; Ertl, M.; Dietrich, G.; Greger, H.; Puppe, F.; Störk, S. Underestimated prevalence of heart failure in hospital inpatients: A comparison of ICD codes and discharge letter information. *Clin. Res. Cardiol.* **2018**, *107*, 778–787. [CrossRef] [PubMed]
8. Balsam, P.; Tymińska, A.; Ozierański, K.; Zaleska, M.; Żukowska, K.; Szepietowska, K.; Maciejewski, K.; Peller, M.; Grabowski, M.; Lodziński, P.; et al. Randomized controlled clinical trials versus real-life atrial fibrillation patients treated with oral anticoagulants. Do we treat the same patients? *Cardiol. J.* **2020**, *27*, 590–599. [CrossRef] [PubMed]
9. Balsam, P.; Ozieranski, K.; Tyminska, A.; Zukowska, K.; Zaleska, M.; Szepietowska, K.; Maciejewski, K.; Peller, M.; Grabowski, M.; Lodzinski, P.; et al. Comparison of clinical characteristics of real-life atrial fibrillation patients treated with vitamin K antagonists, dabigatran, and rivaroxaban: Results from the CRAFT study. *Kardiol. Pol.* **2018**, *76*, 889–898. [CrossRef] [PubMed]
10. Lip, G.Y.; Nieuwlaat, R.; Pisters, R.; Lane, D.A.; Crijns, H.J. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The euro heart survey on atrial fibrillation. *Chest* **2010**, *137*, 263–272. [CrossRef] [PubMed]
11. Parikh, R.; Mathai, A.; Parikh, S.; Chandra Sekhar, G.; Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **2008**, *56*, 45–50. [CrossRef] [PubMed]
12. Yao, R.J.R.; Andrade, J.G.; Deyell, M.W.; Jackson, H.; McAlister, F.A.; Hawkins, N.M. Sensitivity, specificity, positive and negative predictive values of identifying atrial fibrillation using administrative data: A systematic review and meta-analysis. *Clin. Epidemiol.* **2019**, *11*, 753–767. [CrossRef] [PubMed]
13. Shah, R.U.; Mukherjee, R.; Zhang, Y.; Jones, A.E.; Springer, J.; Hackett, I.; Steinberg, B.A.; Lloyd-Jones, D.M.; Chapman, W.W. Impact of Different Electronic Cohort Definitions to Identify Patients With Atrial Fibrillation From the Electronic Medical Record. *J. Am. Heart Assoc.* **2020**, *9*, e014527. [CrossRef]
14. McCormick, N.; Lacaille, D.; Bhole, V.; Avina-Zubieta, J.A. Validity of heart failure diagnoses in administrative databases: A systematic review and meta-analysis. *PLoS ONE* **2014**, *9*, e104519. [CrossRef] [PubMed]
15. So, L.; Evans, D.; Quan, H. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. *BMC Health Serv. Res.* **2006**, *6*, 161. [CrossRef] [PubMed]
16. Xu, Y.; Lee, S.; Martin, E.; D'Souza, A.G.; Doktorchik, C.T.A.; Jiang, J.; Lee, S.; Eastwood, C.A.; Fine, N.; Hemmelgarn, B.; et al. Enhancing ICD-Code-Based Case Definition for Heart Failure Using Electronic Medical Record Data. *J. Card. Fail.* **2020**, *26*, 610–617. [CrossRef] [PubMed]

17. Joos, C.; Lawrence, K.; Jones, A.E.; Johnson, S.A.; Witt, D.M. Accuracy of ICD-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events. *Thromb. Res.* **2019**, *181*, 71–76. [CrossRef] [PubMed]
18. Chang, T.E.; Lichtman, J.H.; Goldstein, L.B.; George, M.G. Accuracy of ICD-9-CM Codes by Hospital Characteristics and Stroke Severity: Paul Coverdell National Acute Stroke Program. *J. Am. Heart Assoc.* **2016**, *5*, e003056. [CrossRef] [PubMed]

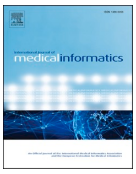
6.2. Publikacja 2

International Journal of Medical Informatics 185 (2024) 105380

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



AssistMED project: Transforming cardiology cohort characterisation from electronic health records through natural language processing – Algorithm design, preliminary results, and field prospects

Cezary Maciejewski [a,b,e], Krzysztof Ozierański [a,*], Adam Barwiołek [c], Mikołaj Basza [d], Aleksandra Bożym [a], Michalina Ciurla [a], Maciej Janusz Krajsman [e], Magdalena Maciejewska [b], Piotr Lodziński [a], Grzegorz Opolski [a], Marcin Grabowski [a], Andrzej Cacko [a,e], Paweł Balsam [a]

[a] 1st Chair and Department of Cardiology, Medical University of Warsaw, 02-091 Warszawa, Poland
[b] Doctoral School, Medical University of Warsaw, 02-091 Warszawa, Poland
[c] Codifive sp. z o.o., Lindleya 16, 02-013 Warszawa, Poland
[d] Medical University of Silesia in Katowice, 40-055 Katowice, Poland
[e] Department of Medical Informatics and Telemedicine, Medical University of Warsaw, 02-091 Warszawa, Poland

ARTICLE INFO

Keywords:
Natural language processing
NLP
Text-mining
Epidemiology
Cardiology

ABSTRACT

Introduction: Electronic health records (EHR) are of great value for clinical research. However, EHR consists primarily of unstructured text which must be analysed by a human and coded into a database before data analysis- a time-consuming and costly process limiting research efficiency. Natural language processing (NLP) can facilitate data retrieval from unstructured text. During AssistMED project, we developed a practical, NLP tool that automatically provides comprehensive clinical characteristics of patients from EHR, that is tailored to clinical researchers needs.

Material and methods: AssistMED retrieves patient characteristics regarding clinical conditions, medications with dosage, and echocardiographic parameters with clinically oriented data structure and provides researcher-friendly database output. We validate the algorithm performance against manual data retrieval and provide critical quantitative and qualitative analysis.

Results: AssistMED analysed the presence of 56 clinical conditions, medications from 16 drug groups with dosage and 15 numeric echocardiographic parameters in a sample of 400 patients hospitalized in the cardiology unit. No statistically significant differences between algorithm and human retrieval were noted. Qualitative analysis revealed that disagreements with manual annotation were primarily accounted to random algorithm errors, erroneous human annotation and lack of advanced context awareness of our tool.

Conclusions: Current NLP approaches are feasible to acquire accurate and detailed patient characteristics tailored to clinical researchers' needs from EHR. We present an in-depth description of an algorithm development and validation process, discuss obstacles and pinpoint potential solutions, including opportunities arising with recent advancements in the field of NLP, such as large language models.

1. Introduction

Electronic health records (EHRs) offer a distinct perspective on the day-to-day clinical care of patients by providing comprehensive healthcare data. EHR analysis may facilitate hypothesis generation, enable efficient multicenter clinical research trials, assess study feasibility, and support the execution of entirely EHR-based observational studies [1–6]. However, the full utilisation of EHRs to realise the above benefits is still limited due to a significant portion of unstructured data in the EHR, which is unavailable for analysis.

Notably, most patients' clinical histories are in a free-form textual format, as doctors prefer such forms for conveying information conveniently, including nuances and expressions not easily captured by coding classifications [7]. However, analysing and extracting meaningful insights from text poses a challenge. Manual extraction of data from free text is time-consuming, error-prone, and impractical for studies with

* Corresponding author.
E-mail address: krzysztof.ozieranski@wum.edu.pl (K. Ozierański).

https://doi.org/10.1016/j.ijmedinf.2024.105380
Received 12 October 2023; Received in revised form 15 February 2024; Accepted 16 February 2024
Available online 19 February 2024

many participants [8].

Natural language processing (NLP) allow computers to interpret, manipulate, and comprehend human language. If utilised efficiently, NLP can unlock the full potential of electronic health record (EHR) data. There have been successful applications of NLP in cardiology and other medical fields [9,10].

We present a system designed to meet the needs of clinical researchers in cardiology. It successfully implements NLP for detecting cardiological patient cohorts from real-life EHR data, making it the first system available for medical text processing in polish. Due to our tool design tailored to the reality of the Polish healthcare system, it could be utilised on a large, national scale in the central database of e-documentation that is currently implemented.

The article discusses design considerations, the development process of the algorithm and preliminary results in comparison to data retrieval by a human. We provide a detailed summary of the challenges encountered during development, which will be valuable for future advancements in the field. Through a thorough analysis of a cohort of real patient records, both qualitatively and quantitatively, we shed light on the limitations inherent to NLP use for clinical data collection. Additionally, we explore different approaches in medical NLP and discuss future perspectives for its application in cardiology including the usage of large language models.

2. Rationale, design, scope, and general overview of the project

We aimed to automatically and accurately characterise patients hospitalised in a cardiology unit. The idea for the project stemmed from the observation that discharge reports in Poland uniformly follow a stereotyped format and contain similar textual fields, which may be a good subject for text processing. Our tool was designed to cater to the customary textual data types in Polish discharge documentation. These, among others, include (1) descriptive discharge diagnoses, (2) discharge recommendations, and (3) echocardiography reports. The desired output of the AssistMED was therefore set to provide a single hospitalisation record, uniquely identified, that includes structured data on common diagnoses, medications, and numeric echocardiographic parameters- essentially a basic cohort of patients characteristics required for any clinical research conduction. The following paragraph briefly describes AssistMED rationale and design. The technicalities of diagnosis and medication detection algorithms are further detailed in Supplementary material.

3. AssistMED algorithm design and functionality

3.1. Diagnosis detection

Descriptive diagnoses typically provide a concise representation of the diagnoses identified through the diagnostic process and any past medical history. In Poland, doctors are accustomed to including descriptive diagnoses in discharge documentation, relying on them for a quick and accurate understanding of a patient's health status in everyday clinical practice. Therefore, an efficient tool capable of analysing these text types holds excellent value in the Polish healthcare system due to its potential for broad applicability.

The knowledge base was used to achieve high natural language recognition performance levels. We created a database, proposed by our clinicians (KO, PB, PL, MG, AC, CM) of important clinical conditions and their possible expressions in medical documentation. Disease definitions- the expressions (full or modified forms) are stored in the database for guided fuzzy pattern-based search. Each expression is a sequence of words that define a disease. The database combines dictionary-based and rule-based methods, providing flexibility in disease detection. The search algorithm is implemented as two components in the spaCy pipeline (a free, open-source library for NLP that allows step-by-step model building). Some of its features include machine learning algorithms.

Our algorithm incorporates a multi-layered structure for diagnoses designed by physicians- a specific diagnosis often implies the existence of multiple broader diagnoses (Fig. 1). This hierarchical framework, consisting of up to three layers, ensures a comprehensive representation of interrelationships between diagnoses. By capturing these complex associations, our algorithm enables accurate and clinically meaningful classification of multiple conditions in cardiology- the presence of a specific diagnosis often indicates the presence of several other more general diagnoses, e.g., the diagnosis of coronary artery disease may be hinted by the presence of conditions such as: history of coronary revascularization, ischemic cardiomyopathy diagnosis, or a history of acute myocardial infarction.

The entire algorithm workflow is presented in the Central illustration.

3.2. Drugs and dose detection

Recommendations in discharge documentation typically includes a list of medications with their proposed dosing regimens and was subject to our text processing algorithm.

First, we reached a consensus within our team of physicians on which drug groups the algorithm should recognise for a high utility in clinical research. We established our own database of medicinal products that contain the desired substances. This database was created by retrieving information from the Office for Registration of Medicinal Products, Medical Devices, and Biocidal Products in Poland. However, some corrections were necessary as the Anatomical Therapeutic Chemical (ATC) coding utilised in such databases does not always reflect the clinically used categorisation of medical substances. Additional ordering and unification of the nomenclature of substances, guided by the consensus reached within our clinical team, was required. The example of hierarchical structure of medical substances in our knowledge base is depicted in Fig. 2.

Following that, we compiled a comprehensive list of the most frequently encountered patterns clinicians use to describe dosage regimens in medical documentation. The clinicians (KO, PB, PL, MG, AC, CM) were actively involved in the process and encouraged to contribute all possible expressions they encountered in their daily patient care activities and list in the order from the most to least frequently used. Subsequently, based on this information, we devised a set of rules to accurately identify key components such as a single drug dose, the number of administrations per day, and the total daily dose.

The recommendation text is first machine-translated into English, and tokenisation is performed using MED7 model [11]. The drug detection phase begins with a lookup into our knowledge database, allowing for approximate matching to account for minor spelling errors occasionally present in medical documentation. Once a drug is detected, the window for dosage detection is determined based on the presence of other detected drugs in its nearest proximity.

Rule-based patterns are then applied in a predefined order of importance to detect the dose and administration frequency of the drug. If none of the patterns yields results, machine learned Med7 algorithm is used as a last resort for retrieving this information. All identified dosages within the window are summed, and post-processing includes normalising the dose to milligrams is performed. For compound drugs containing substances from different classes, dosing detection is waived due to the lack of consistent expression patterns- we noticed a very heterogeneous way of description in medical documentation. In such cases, the algorithm indicates that the substance is part of a compound drug and that the dosage could not be detected.

3.3. Echocardiography

3.3.1. Echocardiography report

Within the echocardiography report analysis, our primary objective

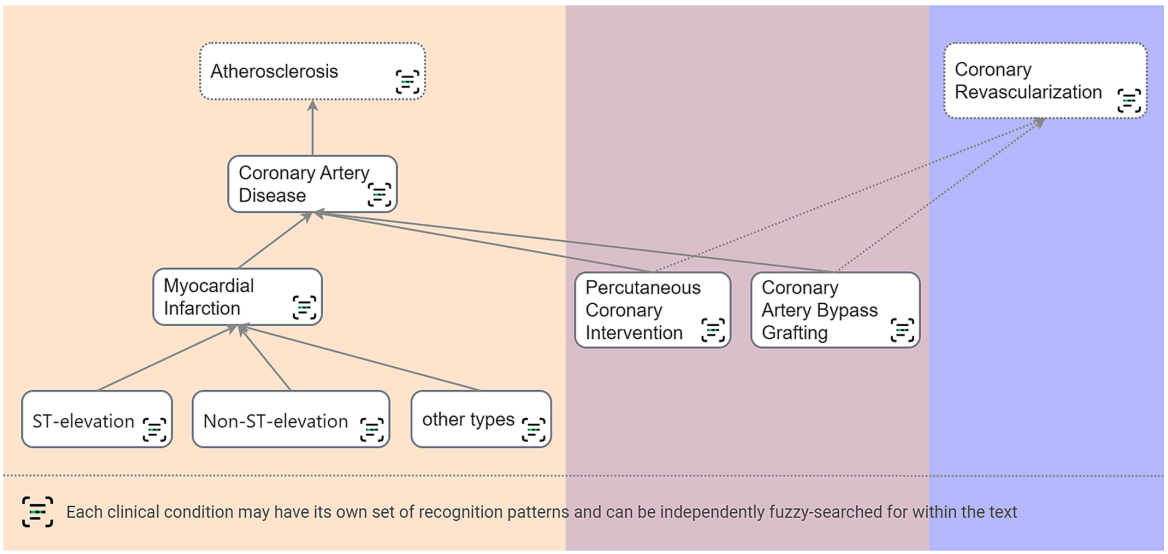


Fig. 1. The hierarchical structure of clinical conditions in the knowledge base.

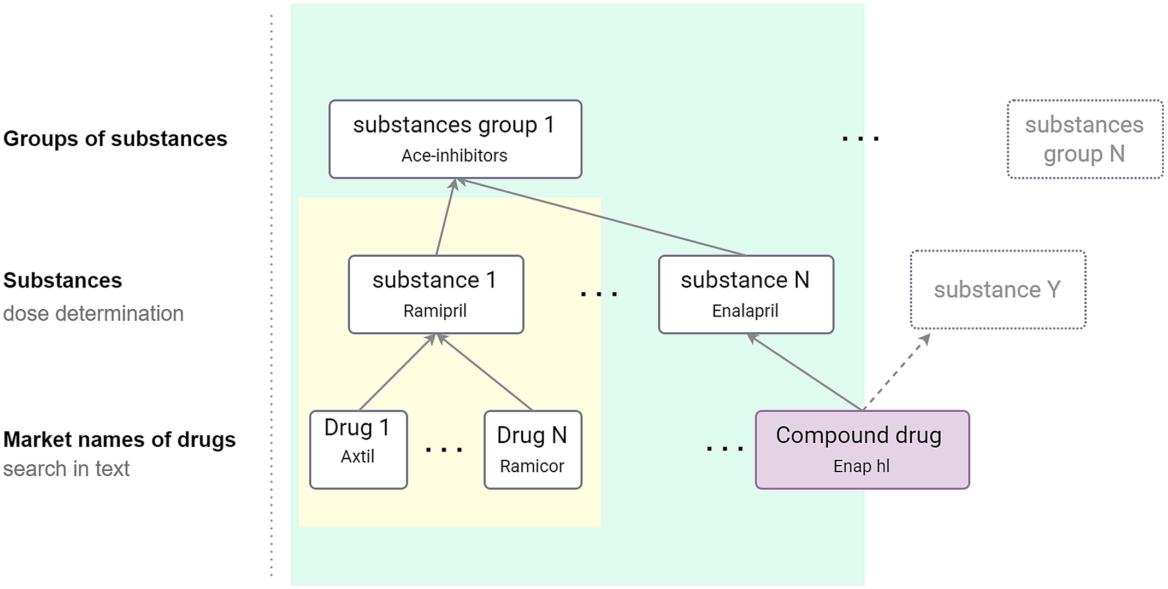


Fig. 2. The hierarchical structure of medical substances in the knowledge base.

was to ensure the reliable collection of numerical parameter values.

The clinicians suggested a list of required echocardiographic numeric parameters and possible expressions. We applied upper and lower bounds on extracted values using reference ranges. Based on this set of rules, we designed a system to detect the parameters and their values. The read parameter value is subsequently normalised to the recommended unit of measurement and checked for plausibility based on falling within the range of prespecified values for the parameter. An adequate note is included in the final output if a read value falls outside the range.

3.4. Annotation module, data analysis module

The AssistMED system includes an online-based module for annotation and review, utilised in the annotation process and qualitative analysis. It allows for simultaneous automatic and manual annotation to be reviewed and allows for the inclusion of corrections.

4. Material and methods

The study material consisted of discharge reports of a cardiology department in Warsaw, Poland (2016–2019) retrospectively extracted

in.xlsx format from the hospital EHR. Our dataset comprises 10,731 anonymised patient records, including descriptive diagnoses, discharge recommendations, and echocardiography reports. This dataset was created by approximately 70 physicians providing daily care in the clinic in that period (six of them participated in the creation of the knowledge base of the algorithm).

First, a random sample of 100 discharge reports with available echocardiography reports was selected. These reports underwent independent double-labelling by two annotators (AB, MC) using the AssistMED online-based interface (the low-right portion of the Central illustration). Annotators were blinded to the outcome of the algorithm detection. A third annotator (CM) resolved the disagreements between the annotators, and the dataset, in its settled form, served as the development dataset for optimising and internally validating the algorithm. The algorithm underwent multiple iterations of optimisation during this stage, which involved adjusting parameters, expanding our database of diagnoses expressions and medical products, incorporating new dosage patterns, and defining cutoff values for echo parameters.

To evaluate the algorithm on unseen data, our databank was once again subsetted based on two prerequisites: the diagnosis of atrial fibrillation and the presence of an echocardiography report. The cohort was manually identified by randomly evaluating records until the evaluation dataset consisted of 400 records that satisfied the specified criteria. Subsequently, the dataset was annotated by AssistMED and later verified by a single annotator who corrected the errors of the algorithm, with the task split between two annotators (AB/MC). The annotators were unaware of the design and knowledge base of the algorithm to limit bias. The selection focused on atrial fibrillation patients as this validation cohort is part of an ongoing larger project that aims to characterise this patient group in our department. Consequently, evaluating the algorithm's performance in identifying atrial fibrillation is restricted to the development dataset.

5. Statistical analysis

Our manuscript preparation aimed to follow good practices in NLP research reporting in the medical field [12].

Point estimates for sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. The F1 score was computed using the formula: $2 \times ((PPV \times sensitivity) / (PPV + sensitivity))$. Cohen's Kappa coefficient assessed inter-rater reliability between annotators and algorithm performance. The interpretation of the Kappa coefficient results is as follows: ≤ 0 indicating no agreement, 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. The above metrics were macro-averaged for a general overview of the algorithm's performance. In a qualitative analysis, total error rates for diagnoses, drugs and echo parameters were calculated as the total number of incorrectly detected parameters divided by all detected parameters.

Continuous variables were presented as medians and quartiles, while categorical and ordinal variables were reported as frequencies and percentages. Fisher's exact test was used to compare the frequencies of categorical variables (e.g., diagnoses, drug classes), and the Mann-Whitney *U* test was used for comparing continuous variables (e.g., echo parameter values). A p-value below 0.05 was considered statistically significant for all tests. All tests were two-tailed.

Statistical analyses and calculations were performed using Python (scipy.stats v1.9.3 and NumPy 1.22.3 libraries).

The total time required for manual annotation of the data portion was defined as the duration spent actively within the annotation module of the AssistMED web application. The time for automatic annotation was defined as the total time from the process initiation in the web application to the generation of the final report.

6. Results

6.1. Quality of manual annotations assessment

Two annotators were involved in labelling 100 records of the development dataset. Annotator 1 spent 5 h and 50 min, while Annotator 2 spent 4 h and 44 min on the task. The inter-annotator agreement demonstrated almost perfect agreement with only minor disagreements. The macro averaged Cohen's kappa coefficients were: 0.969 for diagnoses, 0.983 for drug groups, and 0.978 for echocardiographic parameters. No statistically significant differences among the analysed items indicated high agreement between annotators. Detailed comparisons between the studied items can be found in the Supplementary materials (Supplementary Tables 12–16).

6.2. The algorithm performance on the development dataset

The results achieved by the algorithm on the development dataset are presented in Supplementary Materials (Supplementary Tables 6–11). The results indicated almost perfect agreement for most of the items.

6.3. The algorithm performance on the validation dataset

The analysed sample consisted of 400 annotated health records. The total annotation time for the validation dataset was 13 h and 8 min, while the algorithm detection process took 21 min.

The analysis encompassed both quantitative and qualitative assessments of the algorithm's performance compared to manual annotation. The quantitative analysis focused on the statistical evaluation of the results generated by the algorithm. The macro averaged algorithm performance is presented in Table 1, indicating high agreement with human annotation. We noted that there was only a marginal drop in the performance on the validation dataset compared to the results achieved on the development dataset (comparison of Table 1 and Supplementary Table 6), suggesting that AssistMED managed to generalise well. The qualitative analysis involved manually examining disagreements between the algorithm and manual classifications. These disagreements were classified according to the reasons for disagreement and the type of resulting error (false positive or false negative).

6.4. Diagnoses

There were 56 analysed diagnoses entities, resulting in a total of 3,952 diagnoses to be assigned in the validation dataset, with an average of 9.9 diagnoses per patient.

The algorithm demonstrated high accuracy in correctly assigning diagnoses compared to manual annotation. The macro-averaged metrics can be found in Table 1. Supplementary materials (Supplementary Tables 1-5) present a comprehensive quantitative analysis. The algorithm incorrectly assigned 122 diagnoses, comprising 73 false positives and 49 false negatives. This corresponds to an overall error rate of 3.1 % in the diagnoses assignment task. No statistically significant differences were

Table 1
Quantitative analysis of the algorithm performance on validation dataset- macro averaged statistics.

Parameter	Diagnoses	Drug groups	Echo parameters
Sensitivity	0,972	0,991	0,981*
Specificity	0,995	0,995	0,922*
F1 score	0,924	0,983	0,988*
PPV	0,921	0,975	0,996*
NPV	0,997	0,988	0,981*
Accuracy	0,995	0,993	0,993*
Cohen's kappa	0,981	0,982	0,985*

*The ability to detect the presence of a parameter within the text- quantitative results are presented in Supplementary Table 5.

observed between classifications for scrutinised diseases.

Qualitative analysis revealed that four of the algorithm's false positives were systematically attributed to an incorrect assignment of aortic aneurysm. This issue is a flaw in the algorithm design originating from inappropriate definitions in the knowledge base and will be addressed and corrected in an upcoming release. Another reason for repeatable false positives (5 instances) was related to thyroid disease. It occurred when the algorithm assigned the presence of thyroid disease when only a thyroid goitre was present, while the annotator did not treat it as significant thyroid pathology. This misclassification was a result of the algorithm's programming is not an actual flaw. Some false positives resulted from a lack of advanced context analysis, such as when a diagnosis was only suspected (e.g. "suspicion of COPD"), or a procedure was yet to be done (e.g. "patient qualified for elective percutaneous coronary intervention"). In 27 false positives, the algorithm identified the disease correctly, while the annotator missed the diagnosis. The rest of the false positives occurred randomly.

Regarding false negatives, most were random algorithm detection problems followed by a lack of context analysis. Additionally, there were cases where the misdetection resulted from a typo, which is treated as an actual algorithm flaw, as the annotator had no problem correctly assigning the disease presence.

It is important to note that the total number of false positives and false negatives in Supplementary Table 1 (quantitative analysis) and Table 2 (qualitative analysis) does not match perfectly, as some errors in sub-diagnoses can propagate to higher-level diagnoses (hierarchical structure as presented in Fig. 1)- unfortunately, due to complexity of the task we couldn't track which specific sub-diagnosis resulted in the incorrect assignment of higher-level diagnosis.

Based on the results presented in Table 2, 56 % (37/66) of errors could be solely attributed to algorithm underperformance. Although this is speculative, if we considered these metrics, the true overall error rate in diagnosis assignment would decrease from 3.1 % to 1.7 %.

6.5. Drugs

Drug detection analysis was conducted at three degrees of detail: drug group detection, the correct substance within drug group detection, and correct dosing detection. The detailed quantitative results on the three levels are available in Supplementary Tables 2-4.

In the validation dataset, 1,813 individual substances were detected, averaging 4.5 medicines per patient. There was an almost perfect agreement between the algorithm and human classification in identifying all drug groups, as shown in Table 1 and Supplementary Table 2. There were 47 disagreements (22 false positives and 25 false negatives) at the drug group level, but no statistically significant differences were observed for any of the drug classes identified. The overall error rate for drug group identification was 2.8 %.

The qualitative analysis conducted at the substance level, as presented in Table 3, revealed 50 disagreements (28 false positives and 22 false negatives). False negatives mainly occurred due to random omissions of drugs by the algorithm (some of them stemmed from the loss of drug names from unknown reasons in the translation stage which precluded subsequent identification- this is an algorithm design flaw), with 9 cases resulting from spelling errors in the documentation that were not accounted for. False positives, accounting for 46 % of the cases, were primarily attributed to a lack of context analysis (examples detailed in Table 3), followed by incorrect human annotation (29 %). 82 % (41/50) of the errors were attributed to algorithm failure. Considering these metrics, the actual error rate for substance identification would decrease from 2.8 % to 2.3 %.

Regarding dosage detection (Table 3), there were 229 instances (12.6 % error rate) where disagreements in dosage assignment occurred. Among these, the algorithm was responsible for 133 cases (60 %), indicating incorrect dosage assignment. The remaining disagreements were attributed to complex dosing regimens, such as VKAs dependent on INR checkups or alternate dosing on certain weekdays (e.g., VKAs, digoxin) and temporary use of heparins for bridging. It should be noted that the algorithm does not attempt to recognise the dosing of compound drugs, as our early attempts were deemed challenging due to the unclear and heterogeneous descriptions in the documentation. As a result, dosing for compound drugs is left blank (substances are identified). Considering the reasons mentioned earlier, we recognise 133 disagreements (7.3 % error rate) as actual algorithm flaws, while the remaining disagreements are results of design considerations.

Table 2
Qualitative analysis of algorithm errors in diagnoses assignment- validation dataset.

Reason for misclassification	Numer of misclassifications	FP	FN
Reason for misclassification: Lack of context analysis, such as a procedure that was not yet done or a diagnosis that is still under investigation.	8	4	4
Pure algorithm detection problem, including typos in the disease name, which were correctly identified by a human annotator (value in parenthesis).	29	9	20 (4)
Incorrect interpretation by the annotator	29	27	2

Table 3
Qualitative analysis of algorithm errors in drugs assignment- validation dataset.

Reason for incorrect substance identification	Numer of misclassifications	Type of error	
		FP	FN
Algorithm detection problem, including typos in the documentation that did not preclude correct manual annotation- value in bracket	26	5	21 (9)
Lack of context analysis • Drugs recommended for future use or medication switching at a specified time in the future (both drugs, one already taken and the one recommended for switching, mentioned in the text) • Descriptions of bridging strategy with heparins in case of surgery • Drugs taken "on demand" only (e.g., captopril for high blood pressure) • Errors in documentation, such as erroneously included cautionary notes to not withdraw a medicine that is not within the regimen	15	15	–
Incorrect annotation by the annotator	9	8	1

Reason for incorrect dosing identification	Number of incorrectly identified doses
Pure algorithm detection problem	133
Complex dosing e.g., VKAs, digoxin, heparins dosed in units	68
Substance in compound drug	19
Incorrect annotation	8
Dose error in documentation that precluded correct identification by both the annotator and algorithm	1

6.6. Echo

In the dataset, there were a total of 3,771 numeric echocardiographic parameters to be detected. The macro averaged tests for parameter detection, as shown in Table 1, indicated almost perfect agreement between automatic and manual retrieval. Supplementary Table 5 provides detailed quantitative results. Notably, there were no statistically significant differences in the values detected through automated and manual retrieval for any of the analysed echo parameters.

Qualitative assessment (Table 4) identified 40 instances of inappropriate parameter detection by the algorithm. False negatives, indicating undetected parameters, occurred primarily when the values fall outside the predefined parameter range, which will be corrected in the next release. On the other hand, false positives were random occurrences.

Furthermore, there were 24 instances of disagreements regarding the parameter value. Most of these disagreements were attributed to mistakes in manual annotation or the presence of the parameter appearing twice in the echo report.

There were 65 disagreements, resulting in an error rate of 1.7 %. Among these disagreements, 53 were recognised as algorithm flaws, leading theoretically to the actual error rate of 1.40 %.

7. Discussion

This is the first in Poland and, to our knowledge of literature, a system with the broadest spectrum of data recovered automatically from EHR for research purposes simultaneously. AssistMED enables a comprehensive assessment of multiple cardiovascular and internal diseases, medications, dosing, and numeric echocardiographic parameters. It is tailored to the needs of clinical researchers in an inpatient setting and therefore has the potential for rapid application. The proposed design facilitates the acquisition of output data that is appropriately structured, allowing for rapid analysis and comprehensibility for clinical researchers.

Our hybrid approach to text processing of medical documentation combined the strengths of rule-based, dictionary-based, and machine learning techniques, resulting in improved performance and structured, interpretable output. Machine learning components of spaCy and MED7 models were integrated with our own developed advanced rule-based algorithm that includes a carefully created knowledge base by clinicians. It offers a simple way to add new diseases and drug entities to a knowledge base to retrieve various cohorts of patients, which can be used in various research initiatives. Contrary to pure machine learning methods, which require a lot of training data, this algorithm can begin functioning fast after adding new entities with their expected expressions. The process is intuitive and does not require any programming experience.

Overall, the algorithm showed high agreement with manual annotation and could generalise well, as expressed by the comparable results achieved on development and validation datasets. In the validation dataset, there were no statistically significant differences between manual and automatic retrieval for any diagnosis, drug (up to the drug dose level) or echo parameter (Supplementary Tables 1-5). This shows the great potential of our approach and proves that automatic retrieval from electronic medical documentation can be accurate and time efficient. Our rigorous qualitative analysis provides an in-depth understanding of the algorithm's performance, highlights flaws and areas where improvements are needed, and clearly shows the limitations of our and most other NLP approaches for a cohort of patient characterisation.

Table 4
Qualitative analysis of algorithm errors in echo parameters detection- validation dataset.

Reason for incorrect identification	Number of misclassifications	FP	FN
Algorithm misclassification (including unconventional nomenclature that was correctly interpreted by the annotator)	40	13	27
Incorrect annotation	1	1	–
Reason for incorrect echo parameter value identification	Number of incorrectly identified parameters		
Incorrect value detected by the algorithm	6		
Measurement present twice in the echo report- both of measurements are correct	7		
Incorrect annotation	11		

According to the systematic review on text processing in medicine [12], NLP application attempts are popular in the cardiovascular field, likely due to the need for large cohorts of patients and a higher percentage of data being unstructured than in other medical areas. Most of the attempts come from a few research groups [9] and are targeted commonly at specific cohorts of patient retrieval and classifying disease phenotypes. Notably, considering publications summarised in this review [9], our project had an extensive spectrum that simultaneously targeted the most frequently targeted goals in other projects: disease and disease subtypes detections cases, medications with dosage and cardiac measurements retrieval.

We will discuss the results of other published projects dividing them by technologies adopted for text processing in order to summarise different approaches to the tasks resolved by the AssistMED project.

7.1. Rule-based and dictionary-based methods

These approaches are based on designed dictionaries and rules for detecting specified patterns in the text. They can be rapidly applied and extract information from standardised text formats. Their limitations include: lack of flexibility, ambiguity struggles, and the requirement of regular maintenance and updates are among notable limitations.

In the study by Small et al., [13] authors utilised PennSeek tool to identify patients with trileaflet aortic stenosis and coronary artery disease through text-processing of procedural reports: echocardiography and coronary angiography. Authors, as in our approach, focused on these data types due to more stereotyped text than progress notes. The achieved results for aortic stenosis detection were: Se 0.92, Sp 0.99, PPV 0.95, NPV 0.99, F1 score 0.97 and for coronary artery disease: Se 0.99, Sp 0.94, PPV 0.97, NPV 0.94, F1 score 0.98.

In a study by van Dijk et al. [3], authors utilised text mining technique based on designed regular expressions with the use of software CTcue, version 2.0.12, to attempt LoDoCo2 trial [14] pre-screening and data collection in the entirety of EHR documentation. The accuracy of automatically extracted data was 0.88, Se. 0.806, Sp. 0.827, PPV 0.928, NPV 0.937, F1-score 0.863. The lowest accuracies were found for hypertension (62.6 %), antiplatelet therapy (68.8 %), and beta-blocker use (73.3 %) – we did not note such problems in our analysis. Researchers showed that a tool allowed to manually screen only 20.1 % of the original 92,466 patients for trial inclusion, which identified 82.4 % of the participants eventually recruited to the trial, thus showing a practical NLP solution facilitating clinical trial conduction.

The other study by Karystianis et al. [15] extracted mentions of 5 diseases, smoking status, family history status and medications from clinical notes. The project was an ambitious attempt to use less stereotypical textual data- daily clinical notes. The authors utilised a rule-based approach with own-developed dictionaries for that purpose. The overall average measures were Se 90.07, PPV 85.57, and F1-score 87.76. The comprehensive analysis of false positives revealed a lack of in-depth context awareness of the algorithm (missed negation, misclassifying allergy for a specific drug as a drug being taken by a patient) as a major obstacle. False negatives were mainly the result of unexpected shortcuts not foreseen during database design. Additionally, coronary artery disease was deemed the most challenging to identify, causing many false negatives- our group recognised this and proposed a resolution in our system's hierarchical clinical conditions structure (Fig. 1) which ended up in accurate identifications. The results were important in the NLP area because of the use of daily progress notes. The authors clearly displayed the problems of analysing clinical notes due to less predictable structure, complex context analysis and frequent jargon usage. This justifies our approach of limiting the spectrum of EHR textual data analysed in agreement with clinicians focusing on the goal of the text processing tool, which was simple, rapid and accurate data retrieval.

An NLP tool, EchoInfer, was developed to automatically extract cardiovascular structure and function data from echocardiographic

reports [16]. EchoInfer achieved results comparable to ours with an average Se of 92.21 %, PPV of 94.06 %, and F1-score of 93.12 % evaluated in a single institution.

7.2. Supervised machine learning methods

Supervised machine learning methods are flexible, adaptable to new language patterns, and can achieve high accuracy with proper training. However, they depend on large samples of labelled data to become functional, use complex models that can be difficult to interpret and are sensitive to data quality.

Weissler et al. [17] presented an elegant solution of text-processing based on multiple textual data types from EHR (progress notes, history and physical notes, discharge summaries, brief/operative reports, radiology reports, and consult notes) to recognise patients with a single condition: peripheral artery disease. All texts were processed and contributed to the training of the machine learning model. Based on the cut-off selected by the authors, the algorithm achieved Se 90 %, Sp 62 %, and PPV of 74 %; therefore, the solution needs further development due to many false positives. Despite this, the approach is exciting as it goes beyond the classification of specific text fields in the EHR aggregating multiple textual data types, which was a limitation of ours and cited works previously discussed.

7.3. Deep learning methods: Large Language Models (LLMs)- the future of the field?

The development of the transformer deep learning models (https://arxiv.org/abs/1706.03762v5) marked the beginning of a rapid acceleration in the field of NLP with large language models (LLMs) and, therefore, must be discussed. These models excel in language comprehension and context awareness, and thanks to their pre-training on vast datasets from the internet, they contain a broad knowledge base across various domains, including medicine. Context awareness may enable them to be effectively applied for tasks such as diagnosis and drug detection in less stereotypical texts of medical documentation such as daily clinical notes. This may help in correctly identifying drug intake status (whether it is: taken, discontinued, halted, considered for introduction or allergic to) or disease status (confirmed, included in the differential diagnosis but not yet confirmed or excluded). Our algorithm solely incorporated simple negation detection, limiting its capabilities in this matter and the high accuracy of our approach is attributed, in part, to the processing of more organised textual data types in EHR.

LLMs pre-trained specifically on scientific (PubMed abstracts and full articles- BioBERT [18]) and real-world medical data (MedPalm 2-https://arxiv.org/pdf/2212.13138.pdf, Clinical BERT-https://arxiv.org/abs/1904.03323, MedBERT [19]) are being released. These models may perform better in certain clinical domain goals and can be fine-tuned for specific tasks.

With fine-tuning, LLMs already proved to be capable of effectively performing tasks such as parameter value retrieval from MRI reports, as demonstrated by Singh et al. [20], which is a task similar to that attempted in our echo parameters retrieval. The highest achieved macro averaged F1 score was 0.957. The authors developed the model with just 370 human annotations by fine-tuning the BERT-LARGE model. The authors performed simulations that indicated that even smaller datasets would suffice.

The practical example cited show the potency of LLMs for NLP in medicine – having language understanding and background knowledge may allow for the limitation of the number of annotations of model training. Traditionally, a primary constraint for utilising machine learning in medical NLP was the requirement of large annotated corpora lacking in medicine and days, or perhaps weeks, of computation required to pre-train a model. These obstacles might be overcome soon. However, it is essential to note that fine-tuning (or even costly and lengthy pre-training) on clinical data and extensive validation will be necessary for such purposes in the clinical domain. Training of LLMs still requires significant computational resources and can be computationally expensive. Pre-training on large datasets and fine-tuning for specific tasks often requires high-performance computing infrastructure and substantial time and cost investments. This can limit the accessibility of LLMs for organisations or researchers with limited resources- another advantage of tools based on rule-based, dictionary-based and mixed models (such as AssistMED).

One of the advantages of the methods such as those utilised in our project are predictable results and complete control of the algorithm behaviour by the administrator, which are desired in the research domain- the source of errors can be easily tracked, explained and corrected as shown in our qualitative analysis. LLMs are known for their "black box" nature, meaning it can be challenging to understand how they arrive at their decisions or predictions. This lack of interpretability and explainability may be a concern in clinical research.

Furthermore, although LLMs excel in language translation, most of their training data was in English. This might compromise the results of medical data categorisation for other less-represented languages, including Polish.

Nevertheless, our group will likely focus on exploring using LLMs in the future as the most promising technology. By incorporating rule-based or dictionary-based components into the development, tuning, and validation of LLMs, hybrid NLP models, as AssistMED project, may offer a complementary approach to enhance the accuracy, robustness, and domain-specific performance of LLMs. We believe our current work may prove valuable for developing and validating similar tools based on LLMs in Poland, ultimately accelerating the field's growth.

7.4. Limitations

The main limitation of our algorithm is that it has only been so far validated in a single tertiary cardiology medical centre- its applicability to other centres is still to be determined. Therefore, external validation in other cardiac units is needed. Cooperation with other polish centers is being established. Although our results are preliminary and require further validation on a larger scale, it is worth noting that our dataset is comparatively sizable to most of the other published studies.

One might argue that excluding other textual data types, such as progress notes, is a limitation. However, we contend that our approach is valid as it capitalises on natural tendencies ingrained in the everyday clinical practice (extensive usage of specific textual data types) of hospitalists in Poland and as we showed allows the acquisition of high-quality data. Furthermore, the quality of other data types, such as progress notes, varies significantly among local centres and physicians, as there is no standard for their creation. This poses challenges for accurate data acquisition, a concern supported by the experience of other groups [3,15].

8. Conclusions

The utilisation of NLP in clinical research shows significant promise. The AssistMED project is ambitious, aiming to provide clinical researchers with automatic, broad, and detailed patient characteristics for cardiology research. Our work has demonstrated a hybrid approach combining rule-based, dictionary-based and machine-learning techniques. We showed that the task extends beyond NLP and requires careful clinically guided hierarchy design and decision-making for results processing, highlighting the crucial collaboration between IT specialists and clinicians during the early stages of system design. This enabled us to achieve good model efficacy and generate output data tailored explicitly to clinical researchers' needs. We conclude that time is ideal for ambitious, large-scale projects aimed at comprehensive cohorts retrieval from EHRs. Our experience may direct other groups in future developments of NLP tools designed to accelerate clinical research.

Summary table

- A well-performing, natural language processing algorithm (NLP) for automatic, vast and detailed cardiology and internal medicine cohort characterisation from polish electronic health records (EHR) is presented.
- Acquired data is presented in clinically meaningful structure from the clinical researcher perspective and ready for further analysis.
- Through an extensive algorithm validation, limitations of NLP-based data acquisition from EHR for research purposes are discussed.
- Potential solutions to obstacles and future perspectives on NLP usage for clinical research facilitation are drawn.

CRediT authorship contribution statement

Cezary Maciejewski: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Krzysztof Ozierański:** Methodology, Supervision, Writing – review & editing, Conceptualization, Writing – original draft. **Adam Barwiołek:** Conceptualization, Formal analysis, Software, Writing – original draft. **Mikołaj Basza:** Conceptualization, Methodology, Writing – original draft. **Aleksandra Bożym:** Data curation, Writing – original draft, Investigation. **Michalina Ciurla:** Data curation, Investigation, Writing – original draft. **Maciej Janusz Krajsman:** Conceptualization, Methodology, Writing – review & editing. **Magdalena Maciejewska:** Conceptualization, Writing – review & editing. **Piotr Lodziński:** Conceptualization, Supervision, Writing – review & editing. **Grzegorz Opolski:** Conceptualization, Writing – review & editing. **Marcin Grabowski:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Andrzej Cacko:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Validation, Writing – original draft, Writing – review & editing. **Paweł Balsam:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Special acknowledgements to Eliza Konstanciuk and Anna Kula from the Technology Transfer Office of the Medical University of Warsaw for their exceptional support in the project implementation on the administrative front.

Special thanks to Tomasz Pluta and Wioletta Wąsowska-Filipek from Codifive for their valuable contribution during the interface design and programming phases.

Statements and Declarations: The authors have no relevant financial or non-financial interests to disclose.

Funding: This research was supported by the non-commercial research grant: Innovation Incubator 4.0, Medical University of Warsaw (Grant number: 1MF/FS249/ZW/CTT/EK/14); Ministry of Science and Higher Education.

Ethics approval: Due to the observational character of the study, involving anonymised patients' data the ethics committee approval was not required.

Informed consent: non-applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2024.105380.

References

[1] H. Hemingway, F.W. Asselbergs, J. Danesh, R. Dobson, N. Maniadakis, A. Maggioni, et al., Big data from electronic health records for early and late translational cardiovascular research: challenges and potential, Eur. Heart J. 39 (16) (2018) 1481–1495.
[2] T.M. Maddox, N.M. Albert, W.B. Borden, L.H. Curtis, T.B. Ferguson, D.P. Kao, et al., The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association, Circulation 135 (14) (2017) e826–e857.
[3] W.B. van Dijk, A.T.L. Fiolet, E. Schuit, A. Sammani, T.K.J. Groenhof, R. van der Graaf, et al., Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study, J. Clin. Epidemiol. 132 (2021) 97–105.
[4] M.R. Cowie, J.I. Blomster, L.H. Curtis, S. Duclaux, I. Ford, F. Fritz, et al., Electronic health records to facilitate clinical research, Clin. Res. Cardiol. 106 (1) (2017) 1–9.
[5] E. Sumi, S. Teramukai, K. Yamamoto, M. Satoh, K. Yamanaka, M. Yokode, The correlation between the number of eligible patients in routine clinical practice and the low recruitment level in clinical trials: a retrospective study using electronic medical records, Trials 14 (2013) 426.
[6] R. Farmer, R. Mathur, K. Bhaskaran, S.V. Eastwood, N. Chaturvedi, L. Smeeth, Promises and pitfalls of electronic health record analysis, Diabetologia 61 (6) (2018) 1241–1248.
[7] C. Lovis, R.H. Baud, P. Planche, Power of expression in the electronic patient record: structured data or narrative text? Int. J. Med. Inf. 58–59 (2000) 101–110.
[8] D.O. Klein, R. Rennenberg, R. Gans, R. Enting, R. Koopmans, M.H. Prins, Limited external reproducibility restricts the use of medical record review for benchmarking, BMJ Open Qual. 8 (2) (2019) e000564.
[9] M. Reading Turchioe, A. Volodarskiy, J. Pathak, D.N. Wright, J.E. Tcheng, D. Slotwiner, Systematic review of current natural language processing methods and applications in cardiology, Heart 108 (12) (2022) 909–916.
[10] E. Ford, J.A. Carroll, H.E. Smith, D. Scott, J.A. Cassell, Extracting information from the text of electronic medical records to improve case detection: a systematic review, J. Am. Med. Inform. Assoc. 23 (5) (2016) 1007–1015.
[11] A. Kormilitzin, N. Vaci, Q. Liu, A. Nevado-Holgado, Med7: A transferable clinical natural language processing model for electronic health records, Artif. Intell. Med. 118 (2021) 102086.
[12] S. Sheikhalishahi, R. Miotto, J.T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani, Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review, JMIR Med. Inform. 7 (2) (2019) e12239.
[13] A.M. Small, D.H. Kiss, Y. Zlatsin, D.L. Birtwell, H. Williams, M.A. Guerraty, et al., Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease, J. Biomed. Inform. 72 (2017) 77–84.
[14] S.M. Nidorf, A.T.L. Fiolet, A. Mosterd, J.W. Eikelboom, A. Schut, T.S.J. Opstal, et al., Colchicine in Patients with Chronic Coronary Disease, N. Engl. J. Med. 383 (19) (2020) 1838–1847.
[15] G. Karystianis, A. Dehghan, A. Kovacevic, J.A. Keane, G. Nenadic, Using local lexicalized rules to identify heart disease risk factors in clinical notes, J. Biomed. Inform. 58 (2015) S183–S188.
[16] C. Nath, M.S. Albaghdadi, S.R. Jonnalagadda, A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports, PLoS One 11 (4) (2016) e0153749.
[17] E.H. Weissler, J. Zhang, S. Lippmann, S. Rusincovitch, R. Henao, W.S. Jones, Use of Natural Language Processing to Improve Identification of Patients With Peripheral Artery Disease, Circ. Cardiovasc. Interv. 13 (10) (2020) e009447.
[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[19] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, npj Digital Med. (2021), 4(1):86.
[20] P. Singh, J. Haimovich, C. Reeder, S. Khurshid, E.S. Lau, J.W. Cunningham, et al., One Clinician Is All You Need-Cardiac Magnetic Resonance Imaging Measurement Extraction: Deep Learning Algorithm Development, JMIR Med. Inform. 10 (9) (2022) e38178.

Shortcut list and NLP dictionary

Se-: sensitivity
Sp-: specificity
PPV-: positive predictive value
NPV-: negative predictive value
Tokenization-: separating a piece of text into smaller units (tokens)- in this scenario, individual words (e.g. The sentence: "This is aorta." is split into 4 individual **tokens:** "This", "is", "aorta", ".")
Truncation-: cutting off a portion of text (e.g. individual word) to a specified maximum length.
Stemming –: reducing the words to their base or root form, known as the stem (e.g. words: "**aort**ic" and "**aort**a" have a common stem: "aort")
dictionary-based text processing-: text analytics technique based on look-up whether the phrase in the text appears in the dictionary
rule-based methods text processing-: a more advanced technique in which predefined linguistic rules are used to analyze and process textual data allowing for greater flexibility

Fuzzy searching-: a technique used to find approximate matches for a given query or search term. It allows for the retrieval of words or words sequences similar to, but not necessarily identical to, the search term

Levenshtein algorithm-: mathematical formula to calculate the similarity between words/phrases. Allows for matching similar words and is a part of a **fuzzy searching** technique

6.3. Publikacja 3

ORIGINAL ARTICLE

Practical use case of natural language processing for observational clinical research data retrieval from electronic health records: AssistMED project

Cezary Maciejewski[1,2,3], Krzysztof Ozierański[1], Mikołaj Basza[4], Adam Barwiołek[5], Michalina Ciurla[1], Aleksandra Bożym[1], Maciej J. Krajsman[3], Piotr Lodziński[1], Grzegorz Opolski[1], Marcin Grabowski[1], Andrzej Cacko[1,3], Paweł Balsam[1]

1 First Department of Cardiology, Medical University of Warsaw, Warszawa, Poland
2 Doctoral School, Medical University of Warsaw, Warszawa, Poland
3 Department of Medical Informatics and Telemedicine, Medical University of Warsaw, Warszawa, Poland
4 Medical University of Silesia in Katowice, Katowice, Poland
5 Codifive sp. z o.o., Warszawa, Poland

KEY WORDS

anticoagulation, atrial fibrillation, electronic health records, natural language processing, text mining

ABSTRACT

INTRODUCTION Electronic health records (EHRs) contain data valuable for clinical research. However, they are in textual format and require manual encoding to databases, which is a lengthy and costly process. Natural language processing (NLP) is a computational technique that allows for text analysis.

OBJECTIVES Our study aimed to demonstrate a practical use case of NLP for a large retrospective study cohort characterization and comparison with human retrieval.

PATIENTS AND METHODS Anonymized discharge documentation of 10314 patients from a cardiology tertiary care department was analyzed for inclusion in the CRAFT registry (Multicenter Experience in Atrial Fibrillation Patients Treated with Oral Anticoagulants; NCT02987062). Extensive clinical characteristics regarding concomitant diseases, medications, daily drug dosages, and echocardiography were collected manually and through NLP.

RESULTS There were 3030 and 3029 patients identified by human and NLP-based approaches, respectively, reflecting 99.93% accuracy of NLP in detecting AF. Comprehensive baseline patient characteristics by NLP was faster than human analysis (3 h and 15 min vs 71 h and 12 min). The calculated CHA_2DS_2VASc and HAS-BLED scores based on both methods did not differ (human vs NLP; median [interquartile range], 3 [2–5] vs 3 [2–5]; $P = 0.74$ and 1 [1–2] vs 1 [1–2]; $P = 0.63$, respectively). For most data, an almost perfect agreement between NLP- and human-retrieved characteristics was found; daily dosage identification was the least accurate NLP feature. Similar conclusions on cohort characteristics would be made; however, daily dosage detection for some drug groups would require additional human validation in the NLP-based cohort.

CONCLUSIONS NLP utilization in EHRs may accelerate data acquisition and provide accurate information for retrospective studies.

Correspondence to:
Krzysztof Ozierański, MD, PhD, First Department of Cardiology, Medical University of Warsaw, ul. Banacha 1a, 02-097 Warszawa, Poland, phone: +48 22 599 19 58, email: krzysztof.ozieranski@wum.edu.pl

Received: November 9, 2023.
Revision accepted: March 13, 2024.
Published online: March 19, 2024.
Pol Arch Intern Med. 2024; xx: 16704
doi:10.20452/pamw.16704

INTRODUCTION Data acquisition for observational clinical studies is a time-consuming and costly process that requires medical expertise. Despite a widespread adoption of electronic health records (EHRs) in Poland, this process is not significantly accelerated, because only a fraction of data collected in the EHRs are structured data (eg, billing codes, such as the *International Classification of Diseases, 10th Revision* [ICD-10] classifications, laboratory examinations), and the rest

WHAT IS NEW?

Natural language processing (NLP) has become popular with a release of large language models, such as ChatGPT. One of the exciting areas of NLP use in medicine is automated clinical research data retrieval from electronic health records. This could allow for time-efficient and accurate clinical research based on large, real-world data cohorts, thus bringing innovation to current methodologies. We present the AssistMED project, during which we developed tools facilitating automated, comprehensive retrieval of patient characteristics (clinical condition, drugs, echocardiographic parameters) through NLP. The project is the first of this type in Poland and provides one of the most comprehensive NLP solutions in the literature, aiding clinical research. We validated the AssistMED results by comparing them to human data retrieval in a large, real-world cohort of patients with atrial fibrillation for an observational registry. Furthermore, we discussed limitations and future perspectives for evolving NLP techniques in clinical research.

consist of unstructured data, for example, textual reports. Despite known limitations of Polish and other health care systems,[1-3] billing codes are widely used for large population studies in cardiology and internal medicine due to a lack of alternative solutions for time-efficient large population data collection.

Natural language processing (NLP) technologies allow computers to interpret, manipulate, and comprehend human language. A public release of ChatGPT, an NLP-based chatbot, sparked interest in text processing among researchers, showing a potential for scientific process facilitation. Globally, a lot of effort has been made to develop tools utilizing NLP to unlock the potential of these textual data from EHRs. Most of these efforts were traditionally focused on English due to availability of datasets and more advanced text-processing tools for this language. The availability of good-quality EHR-derived data is vital for clinical research progress, and is essential for conducting meaningful projects that utilize artificial intelligence in medicine. This fits in with the objectives of the European Health Data Space regulation proposal[4] aiming for better health care data availability for patients, researchers, and industry in the European Union; therefore, progress in efficient NLP application in cardiology is desired.

Our manuscript displays a practical use case of NLP in the AssistMED project, during which we developed a set of tools based on NLP for EHR data analysis in Polish. We demonstrate how NLP techniques can be used for continuation of the CRAFT registry (Multicenter Experience in Atrial Fibrillation Patients Treated with Oral Anticoagulants; NCT02987062), with fully automatic data retrieval and manual validation by humans. Herein, we present a complete workflow of data acquisition with the AssistMED tool, briefly describe the design of our solution, and provide an engaging but concise discussion on the NLP of medical documentation for research data acquisition.

PATIENTS AND METHODS **Rationale for and functioning of AssistMED** The idea of the project stemmed from an observation that discharge reports in Poland follow a typical format and contain equivalent textual fields. The data types for analysis are, therefore, 1) descriptive discharge diagnoses, 2) discharge recommendations, and 3) echocardiography reports (if present). Such data are stored in the EHR system in an organized form, and may be acquired by a clinical researcher in cooperation with the hospital information technology department. Data acquisition requires a legal analysis and consent of the institutional executive and data protection offices. Our study followed both these steps. The National Health Fund central electronic documentation platform could be a data source for even larger-scale applications in Poland.

Our algorithm can receive data in Excel spreadsheet format (.xlsx) (data examples in Polish are presented in Supplementary material, *Table S1*). Anonymized data can be uploaded to an online or offline computer application, and data analysis can be initiated. Currently, the algorithm can detect 72 clinical conditions related to cardiology and internal medicine, medications from 22 drug classes, and 15 numeric echocardiographic parameters. New diagnoses and medications can be added. The analyzed data are presented in an Excel spreadsheet, and a basic statistical report can be provided. They generally represent basic clinical characteristics of a cohort of patients required for any clinical research.

Algorithm implementation The algorithm itself utilizes NLP for entity recognition. For diagnosis detection, we systematically established our database of possible expressions related to each clinical condition in Polish, as proposed by clinicians (KO, PB, PL, MG, AC, and CM). The diagnoses are appropriately structured, recognizing that a specific diagnosis often signifies the presence of another, more general disease (eg, a history of coronary artery bypass grafting indicates the presence of coronary artery disease or a diagnosis of carotid artery disease means that a patient has atherosclerosis). The NLP techniques we adopted allow for flexibility in recognition, which means a condition is recognized despite possible minor typos or different word ordering due to advanced calculations of similarity to examples in the database.

We created our database of medicinal products available on the Polish market for medicine detection. This database was created by retrieving information from the Office for Registration of Medicinal Products, Medical Devices, and Biocidal Products in Poland. However, some corrections were necessary, as the Anatomical Therapeutic Chemical Classification System utilized in such databases does not always reflect the clinically used categorization of medical substances. Additional ordering and unification of the terminology on substances, guided by a consensus reached

within our clinical team, was required. Data are structured and can be retrieved at 3 levels of detail: drug class, active substance, and dosing. For dose detection, our clinicians (KO, PB, PL, MG, AC, and CM) developed a list of the most common patterns of drug dosing description in discharge recommendations. Based on this, we designed our dosage detection rules, supported with a publicly-available machine-learning MED7 module for better accuracy. Dosing detection is not possible for compound drugs.

As for echocardiographic analysis, the clinicians suggested a list of required echocardiographic numeric parameters and possible expressions. We applied upper and lower boundaries on the extracted values using reference ranges. Based on this set of rules, we designed a system to detect the parameters and their values. The read parameter value is subsequently normalized to a universal unit of measurement, and checked for plausibility based on falling within the range of clinically possible values for the parameter. An adequate note is included in the final output, if a read value falls outside the range.

The algorithm knowledge base, rules, and parameters were iteratively modified on a random sample of 200 manually annotated records from our cardiology department to reach its current performance. A technical description of the implementation of the AssistMED algorithms has been documented and published in research literature.[5]

Patients The study material consisted of anonymized documentation of 10 314 consecutive patients discharged from a single tertiary cardiology center between January 1, 2016 and July 15, 2019. For patients hospitalized more than once (n = 2598), only the latest hospitalization was considered for analysis, to retrieve the most current available data. Therefore, we had 7716 individual patient records available for analysis.

The main inclusion criteria of the retrospective CRAFT registry were a diagnosis of atrial fibrillation (AF) and anticoagulation with oral anticoagulants (OACs); therefore, we focused on comprehensive characterization of patients with AF diagnosis.

The entire available database (7716 records) was subjected to an automatic NLP-based analysis by the AssistMED algorithms to extract data on clinical conditions, medications with dosage, and echocardiographic parameters if echocardiography was performed. Each annotation item suggested by the algorithm was subsequently reviewed by a human (patients with AF diagnosis confirmed by human assessment had all their characteristics manually verified; patients without confirmed AF were not verified for other medical conditions, drugs etc.). Inaccurate suggestions of the algorithm were corrected through a single human verification in a convenient data reviewing and correction module (Supplementary material, *Figure S1*), with the task split between 2 annotators who had to accept, deny, or correct each algorithm suggestion (diagnosis, drug, dose) or add potential missing data.

Additionally, we conducted a separate analysis of a sample of 100 records randomly selected from the entire dataset of 7716 patients (with and without AF diagnosis), in which both annotators (AB, MC) independently analyzed data without suggestions from the AssistMED analysis. The results from the annotators were compared to check the interannotator agreement. Data classification by the annotators was also compared to automatic classification, to see if there are any differences when the annotators were blinded to the algorithm suggestions.

The annotators did not participate in the algorithm design and knowledge base preparation to limit annotation bias.

Statistical analysis The Cohen κ coefficient assessed inter-rater reliability between the annotators and the AssistMED algorithm performance in comparison with human judgment. Importantly, for numeric variables (echo parameter values, daily dosage of medications), a complete agreement between the parameter values of 2 compared classifications was treated as an agreement for the κ coefficient calculation (unavailable data or minor differences in values between the 2 classifications were treated as disagreements). The interpretation of the κ coefficient results was as follows: values equal to or below 0 indicated no agreement, those between 0.01 and 0.2 denoted none to slight agreement, between 0.21 and 0.4 poor agreement, between 0.41 and 0.6 moderate agreement, between 0.61 and 0.8 substantial agreement, and between 0.81 and 1 almost perfect agreement. Of note, the *P* value below 0.05 for the Cohen κ coefficient indicated that the degree of agreement (Cohen κ value) was significant.

Continuous variables were presented as median and interquartile range (IQR), while categorical variables were reported as frequency and percentage. The Fisher exact test was used to compare the frequency of categorical variables (eg, diagnosis, drug class), and the Wilcoxon signed-rank test was used for comparing continuous variables (eg, daily drug dosage, echo parameter values). For categorical variables (diagnosis, drug class), point estimates for sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with human annotation as reference were calculated.

A *P* value below 0.05 was considered significant for all tests. All tests were 2-tailed.

Statistical analyses and calculations were performed using Python (scipy. stats v1.9.3, pyirr v0.84.1.2, and NumPy 1.22.3 libraries, Python Software Foundation, Wilmington, Delaware, United States).

Total time required for manual annotation of the data portion was defined as the time spent actively within the annotation module of the AssistMED web application. The time for automatic annotation was defined as the total time

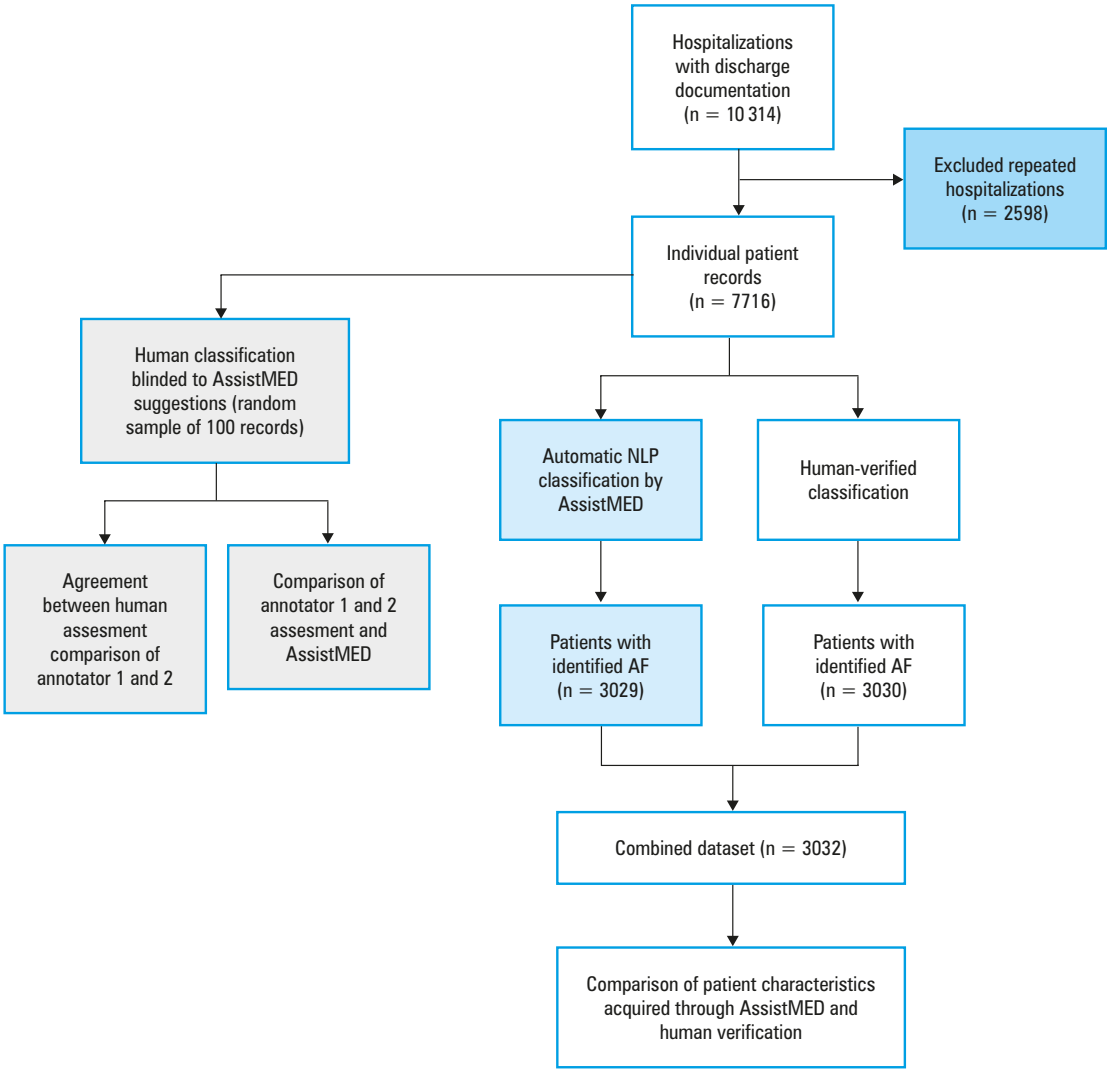


FIGURE 1 Flowchart presenting the study design
Abbreviations: AF, atrial fibrillation; NLP, natural language processing

from the process initiation in the web application to generation of the final report.

Ethics statement Due to an observational design of the study, involving anonymized patient data, neither the ethics committee's approval, nor the patients' informed consent were required.

RESULTS In FIGURE 1 we resent patient flow in the study. First, a separate analysis was performed in a sample of the same 100 records annotated independently by 2 annotators. Its results indicated an almost perfect agreement for most items (moderate agreement for peripheral artery disease and cardiac resynchronization therapy), signifying equivalency of the 2 annotators' judgment. Detailed comparisons between the studied items can be found in Supplementary material, *Tables S2–S5*.

In the entire available dataset, the algorithm identified 3029 and the annotators 3030 patients with AF. The accuracy, sensitivity, and specificity of the automatic AF identification were 99.93%, 99.9%, and 99.96%, respectively. The algorithm falsely identified 2 AF diagnoses and missed 3 AF diagnoses. OACs were taken by 2601 of these patients according to the AssistMED, and 2624 according to the annotators. The algorithm's accuracy in identifying patients fulfilling both principal CRAFT registry inclusion criteria (presence of AF and OAC prescription) was 99.5%, sensitivity was 98.8%, and specificity 99.8%, indicating high agreement. This signifies that the AssistMED-identified cohort would consist of largely the same patients if used for the CRAFT registry as the human-identified cohort.[6]

There were 3032 individual patient records in the combined dataset available for comparison (5 AF records inaccurately identified by the AssistMED, that is, 2 false positives and 3 false negatives included in all further analyses). All analyses were performed on dependent samples;

TABLE 1 Diagnosis detection (n = 3032), AssistMED performance in comparison with human annotators

Diagnosis	Cases detected by a human, n (%)	Cases detected by AssistMED, n (%)	Sensitivity	Specificity	PPV	NPV	*P* value (Fisher test)	Cohen κ	*P* value (Cohen κ)
Heart failure	1393 (45.96)	1389 (45.83)	1	1	1	1	0.94	0.99	<0.001
Hypertension	2173 (71.69)	2179 (71.89)	1	0.99	1	1	0.89	0.99	<0.001
Poorly controlled hypertension	16 (0.53)	18 (0.59)	0.94	1	0.83	1	0.86	0.88	<0.001
Diabetes	893 (29.46)	895 (29.53)	1	1	1	1	0.98	1	<0.001
Diabetes and glycemic disorders	1092 (36.03)	1078 (35.57)	0.984	1	1	0.99	0.73	0.98	<0.001
Ischemic stroke	250 (8.25)	250 (8.25)	0.99	1	0.99	1	1	0.99	<0.001
Ischemic stroke or TIA	303 (10)	302 (9.96)	0.99	1	0.99	1	>0.99	0.99	<0.001
Ischemic stroke or TIA or systemic embolism	310 (10.23)	322 (10.62)	0.99	1	0.95	1	0.64	0.97	<0.001
Atherosclerosis (any evidence)	1401 (46.22)	1416 (46.72)	1	1	0.99	1	0.72	0.98	<0.001
Carotid artery disease	113 (3.73)	109 (3.6)	0.94	1	0.97	1	0.84	0.95	<0.001
PCI	653 (21.54)	646 (21.31)	0.99	1	1	1	0.85	0.99	<0.001
CABG	227 (7.49)	229 (7.56)	1	1	0.99	1	0.96	1	<0.001
STEMI	122 (4.03)	102 (3.37)	0.84	1	1	0.99	0.2	0.91	<0.001
NSTEMI	282 (9.3)	278 (9.17)	0.99	1	1	1	0.89	0.99	<0.001
MI (any type)	745 (24.58)	742 (24.48)	1	1	1	1	0.95	1	<0.001
Coronary artery disease	1348 (44.47)	1350 (44.54)	1	1	1	1	0.98	1	<0.001
Gastrointestinal bleeding	123 (4.06)	119 (3.93)	0.97	1	1	1	0.84	0.98	<0.001
Intracranial bleeding	31 (1.02)	30 (0.99)	0.94	1	0.97	1	>0.99	0.95	<0.001
Labile INR	12 (0.4)	12 (0.4)	1	1	1	1	1	1	<0.001
Liver disease	12 (0.4)	9 (0.3)	0.75	1	1	1	0.66	0.86	<0.001
Chronic kidney disease	872 (28.77)	873 (28.8)	0.99	1	0.99	1	0.98	0.99	<0.001
Alcoholism	27 (0.89)	27 (0.89)	0.96	1	0.96	1	1	0.96	<0.001

Abbreviations: CABG, coronary artery bypass graft; INR, international normalized ratio; MI, myocardial infarction; NPV, negative predictive value; NSTEMI, non–ST-segment elevation myocardial infarction; PCI, percutaneous coronary intervention; PPV, positive predictive value; STEMI, ST-segment elevation myocardial infarction, TIA, transient ischemic attack

thus, similar records in human- and AssistMED--identified cohorts were required. Baseline characteristics regarding concomitant diseases, drugs, and echocardiographic parameters for both cohorts were determined per the identification method.

The major aspects of the baseline characteristics of automated and manual retrieval are presented in TABLES 1 to 3 (the rest of the available data are provided in Supplementary material, *Tables S6–S7*).

For diagnosis detection, in most cases, there was an almost perfect agreement between the AssistMED and the annotators; the lowest (moderate) agreement was identified for type 1 diabetes, a history of valvuloplasty, and a history of systemic embolism.

For medication detection, there was an almost perfect agreement in the drug group and active substance identification. Accurate identification of dosage proved most challenging (TABLE 3; Supplementary material, *Table S7*), as the algorithm failed to detect the dosage more frequently than the annotators (reflected as more missing dosage data). Most drugs showed a substantial agreement in dose detection (the lowest, that is, slight agreement for antiplatelet dose detection was identified). The disagreements were primarily attributed to a lack of dose detection (more missing data on dosage for the automatic data retrieval method; Supplementary material, *Table S7*), and not to incorrect dosage recognition. This was reflected by low Cohen κ (low agreement), but no significant differences in identified dosages (the paired Wilcoxon test omits cases with no available dosage). Such a situation was detected for antiplatelets and acetylsalicylic acid (TABLE 3; Supplementary material, *Table S7*). However, for vitamin K antagonists there was a significant difference in dose detection.

For echocardiography, we found an almost perfect agreement for all detected parameters.

The calculated CHA_2DS_2VASc and HAS-BLED scores based on both classifications (human vs algorithm) yielded the following results: median (IQR), 3 (2–5) vs 3 (2–5); $P = 0.74$ and 1 (1–2) vs 1 (1–2); $P = 0.63$, respectively, indicating equivalent

TABLE 2 Drug groups, active substances, and dosage detection (n = 3032), AssistMED performance in comparison with human annotators

Drug class	Cases detected by a human, n (%)	Cases detected by AssistMED, n (%)	Sensitivity	Specificity	PPV	NPV	*P* value (Fisher test)	Cohen κ (drug group agreement)	*P* value (Cohen κ, drug group agreement)	Cohen κ (active substance agreement)	*P* value (Cohen κ, active substance agreement	Cohen κ (active substance and dose agreement)	*P* value (Cohen κ, active substance and dose agreement)
ACEI	1574 (51.91)	1564 (51.58)	0.99	0.997	0.997	0.990	0.82	0.987	<0.001	0.986	<0.001	0.907	<0.001
Amiodarone	220 (7.26)	221 (7.29)	0.991	0.999	0.986	0.999	>0.99	0.988	<0.001	0.988	<0.001	0.604	<0.001
Antiarrhythmic 1c	144 (4.75)	132 (4.35)	0.91	1	0.992	0.996	0.5	0.947	<0.001	0.947	<0.001	0.699	<0.001
ASA	390 (12.86)	405 (13.36)	0.997	0.994	0.960	1	0.59	0.975	<0.001	0.975	<0.001	0.271	<0.001
β-Blocker	2504 (82.59)	2474 (81.6)	0.986	0.989	0.998	0.935	0.33	0.953	<0.001	0.953	<0.001	0.906	<0.001
CCBs (dihydropyridine)	724 (23.88)	724 (23.88)	0.986	0.996	0.986	0.996	1	0.982	<0.001	0.983	<0.001	0.897	<0.001
CCBs (non--dihydropyridine)	20 (0.66)	20 (0.66)	1	1	1	1	1	1	<0.001	1	<0.001	0.933	<0.001
Digoxin	297 (9.8)	298 (9.83)	0.997	0.999	0.993	1	>0.99	0.994	<0.001	0.994	<0.001	0.656	<0.001
SGLT2i	18 (0.59)	19 (0.63)	1	1	0.947	1	>0.99	0.973	<0.001	0.973	<0.001	0.336	<0.003
Gliptin	34 (1.12)	32 (1.06)	0.941	1	1	0.999	0.9	0.969	<0.001	0.969	<0.001	0.801	<0.001
GLP-1 agonist	2 (0.07)	3 (0.1)	1	1	0.667	1	>0.99	0.8	<0.001	0.8	<0.001	0.571	<0.03
Metformin	498 (16.43)	493 (16.26)	0.990	1	1	0.998	0.89	0.994	<0.001	0.994	<0.001	0.840	<0.001
MRA	827 (27.28)	810 (26.72)	0.976	0.999	0.996	0.991	0.64	0.981	<0.001	0.981	<0.001	0.902	<0.001
NSAID	5 (0.17)	6 (0.2)	0.8	0.999	0.667	1	>0.99	0.727	<0.001	0.727	<0.001	0.282	<0.11
Heparin	245 (8.08)	283 (9.33)	0.992	0.986	0.859	0.999	0.09	0.913	<0.001	0.901	<0.001	0.448	<0.001
NOAC	1879 (61.97)	1860 (61.35)	0.988	0.997	0.998	0.981	0.63	0.983	<0.001	0.983	<0.001	0.743	<0.001
VKA	747 (24.64)	753 (24.84)	0.992	0.995	0.984	0.997	0.88	0.984	<0.001	0.981	<0.001	0.282	<0.001
Antiplatelet	336 (11.08)	340 (11.21)	1	0.999	0.988	1	0.9	0.993	<0.001	0.993	<0.001	0.062	<0.001
ARB	590 (19.46)	572 (18.87)	0.968	1	0.998	0.992	0.58	0.979	<0.001	0.979	<0.001	0.84	<0.001
Sotalol	53 (1.75)	57 (1.88)	1	0.999	0.930	1	0.77	0.963	<0.001	0.963	<0.001	0.784	<0.001
Statin	1931 (63.69)	1911 (63.03)	0.982	0.987	0.993	0.97	0.61	0.966	<0.001	0.966	<0.001	0.92	<0.001
Sulfonylurea	239 (7.88)	222 (7.32)	0.925	1	0.996	0.994	0.44	0.955	<0.001	0.955	<0.001	0.858	<0.001

Abbreviations: ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin-II receptor blocker; ASA, acetylsalicylic acid; CCB, calcium channel blocker; GLP-1, glucagon-like peptide-1 agonist; MRA, mineralocorticoid receptor antagonist; NSAID, nonsteroidal anti-inflammatory drug; NOAC, non–vitamin K antagonist oral anticoagulant; SGLT2i, sodium glucose cotransporter 2 inhibitor; VKA, vitamin K antagonist; others, see TABLE 1

TABLE 3 Echocardiographic parameters (n = 3032), AssistMED performance in comparison with human annotators

Echocardiographic parameter	Cases detected by a human, n (%)	Median (IQR) (human)	Cases detected by AssistMED, n (%)	Median (IQR) (AssistMED)	Cohen κ (full agreement in parameter value)	*P* value (Cohen κ)	*P* value (Wilcoxon test)
AVA	126 (11.55)	1.56 (1.1–1.8)	126 (11.55)	1.6 (1.1–1.8)	0.96	<0.001	0.69
AVAi	73 (6.69)	0.73 (0.4–0.9)	67 (6.14)	0.73 (0.41–0.9)	0.92	<0.001	>0.99
AcT	556 (50.96)	102 (92–120)	557 (51.05)	102 (92–120)	0.99	<0.001	>0.99
Ao	908 (83.23)	3.5 (3.2–3.8)	905 (82.95)	3.5 (3.2–3.8)	0.99	<0.001	>0.99
IVS	939 (86.07)	1.2 (1.1–1.3)	941 (86.25)	1.2 (1.1–1.3)	0.99	<0.001	>0.99
LA	930 (85.24)	4.7 (4.2–5.1)	928 (85.06)	4.7 (4.2–5.1)	0.98	<0.001	0.95
LAA	602 (55.18)	30 (26–36)	599 (54.9)	30 (26–35.6)	0.96	<0.001	0.74
LVDD	948 (86.89)	5.1 (4.6–5.7)	950 (87.08)	5.1 (4.6–5.7)	0.99	<0.001	0.95
LVEF	1009 (92.48)	54 (44–60)	1010 (92.58)	54 (44–60)	0.99	<0.001	0.95
PWD	917 (84.05)	1.1 (1–1.2)	914 (83.78)	1.1 (1–1.2)	0.99	<0.001	0.95
RAA	508 (46.56)	26 (22–31)	509 (46.65)	26 (22–31)	0.97	<0.001	0.86
RV	915 (83.87)	3 (2.8–3.3)	911 (83.5)	3 (2.8–3.3)	0.97	<0.001	0.95
SPAP	167 (15.31)	46 (39–55)	170 (15.58)	46 (39–55)	0.97	<0.001	>0.99
TAPSE	563 (51.6)	21 (18–24)	561 (51.42)	21 (18–24)	0.98	<0.001	>0.99
TRPG	590 (54.08)	27 (22–35)	599 (54.9)	27 (22–35)	0.97	<0.001	0.93

Abbreviations: AVA, aortic valve area; AVAi, indexed aortic valve area; AcT, pulmonary acceleration time; Ao, aortic diameter; IQR, interquartile range; IVS, interventricular septum diameter; LA, left atrial anteroposterior diameter; LAA, left atrial area; LVDD, left ventricular diastolic diameter; LVEF, left ventricular ejection fraction; PWD, posterior wall diameter; RAA, right atrial area; RV, right ventricle; SPAP, estimated systolic pulmonary arterial pressure; TAPSE, tricuspid annulus plane systolic excursion; TRPG, tricuspid regurgitation pressure gradient

assessment of thrombosis and bleeding risk for both methods.

A complete automatic detection took 3 hours and 15 minutes (about 6.5 min per 100 records), while human verification of the algorith work took 71 hours and 12 minutes (about 2 h and 22 min per 100 records). The analysis of 100 records blinded to the algorithm indications revealed that the first annotator spent 5 hours and 50 minutes on the task, while the second spent 4 hours and 44 minutes. Therefore, the mean manual-only annotation time was 5 hours and 17 minutes per 100 patient records analyzed. Based on this, the estimated time of the human-only database collection would take 159 hours. This signifies that automated retrieval was 20 times faster than human annotation, and 50 times faster than fully manual retrieval. Human verification of the algorithm suggestions was 2.2 times more rapid than fully manual retrieval.

The separate sample of 100 records annotated blinded to the AssistMED classification was additionally evaluated. The results achieved for this dataset by the algorithm and the annotators were compared and are presented in Supplementary material, *Tables S8–S15*. The results indicated an almost perfect agreement between automatic and manual analysis for diagnosis, medications (reduced agreement in the case of similar drug dose identification as in the main cohort), and echocardiographic parameters, indicating that working on the algorithm suggestions did not bias judgment of the annotators.

DISCUSSION The results indicated that NLP-based cohort acquisition yielded a cohort highly similar to that retrieved by human annotators. Unsurprisingly, automatic detection was more rapid. Our findings indicate that utilization of NLP may enable a comprehensive assessment of multiple cardiovascular and internal diseases, medications, dosing, and numeric echocardiographic parameters. Dosage detection was, unsurprisingly, the most challenging for the algorithm as the most detailed feature. This algorithm behavior was documented in our prior publication,[5] which qualitatively described sources of such errors.

To present the results in the context of conventionally and widely used automatic cohort retrieval methodologies, we may compare them to the accuracy of the ICD-10 diagnostic codes from the National Health Fund database. Disease characterization of the cohort utilizing the AssistMED was more accurate than the analysis of administrative ICD-10 codes, according to our previous study.[2] In that study, comparing manually gathered data in the CRAFT registry and the ICD-10–based data, we demonstrated sensitivity of 83% for AF detection, 82% for heart failure, 89% for hypertension, and 69% for thromboembolic events, to mention a few. Specificity was generally decent but varied depending on the condition (32% for hypertension and 40% for atherosclerosis). Ultimately, all these inaccuracies translated in the final cohort of patient baseline characteristics looking substantially different than that observed in the CRAFT registry, with significant differences in estimated CHA_2DS_2VASc and HAS-BLED scores.[7]

TABLE 4 Advantages and disadvantages of current natural language programming (NLP) approaches

NLP approach	Advantages	Disadvantages
Rule-based and dictionary-based	• Full control of the algorithm performance • Predictable results • Large datasets of annotated data not needed at the project onset but for validation of the results • Possibility to design a decently performing NLP tool categorizing multiple conditions at a time • Low development costs	• Less favorable accuracy at other institutions • Portability concerns • Health care professional involvement is often needed during development • Unsatisfactory results in disorganized text data, eg, progress notes • Favorable results in selected textual data types, not the entirety of electronic health records • Pertinent only to the language the tool is developed in
Supervised machine learning	• Usually portable • Can be used at other institutions with good results • Predictable results • Ability to analyze various textual data types at once • Can be language-agnostic • Potentially widely applicable	• Requirement of large high-quality annotated datasets for training at the project onset • Risk of overtraining the model during development • A model too accustomed to training data, which in the end performs badly on new data • High development costs • Practical use cases in the literature demonstrate identification of 1 or a few conditions at a time due to discussed limitations
Deep-learning, large language models	• Potentially allows for accurate identification in less organized texts, such as progress notes • Lower amounts of annotated data required than in conventional machine-learning due to pretraining on vast amounts of openly available data • Solution developed with these techniques will potentially be applicable in different languages	• Black-box approach, ie, lack of understanding why a certain decision was made • Lack of predictability • Model hallucinations • High development costs

According to a systematic review on text processing in medicine,[8] NLP application attempts are popular in the cardiovascular field, likely due to a need for large cohorts of patients and a higher percentage of data being unstructured than in other medical specialties. The AssistMED project is one of the first in Poland, with a vast spectrum of data recovered simultaneously, as compared with other studies, and with a large validation cohort.[9] Output data are tailored to the needs of clinical researchers in an inpatient setting and, therefore, have a potential for broader application. The proposed design facilitates acquisition of output data that are appropriately structured, allowing for rapid analysis and comprehensibility for clinical researchers. Furthermore, an automatic summary statistics generation module may facilitate the initial stage of a research project, for example, providing information on the number of patients with specific characteristics in the past to ascertain future study recruitment.

The following section will briefly describe various text-processing solutions used in electronic medical documentation from a technical perspective and will provide examples of vital research problems they solve.

Landscape of text processing of medical documentation for clinical data retrieval

Rule-based and dictionary-based algorithms The simplest methods of text processing from a technical perspective are rule-based and dictionary-based algorithms that enable detection of specified patterns in the presented text. Their development includes establishing a database of terms / expressions / patterns in the data that need to be recognized. This process often requires cooperation of text-processing experts. An advantage of this approach is its predictability, that is, errors are easily tracked, and the algorithm can be gradually improved. Its main drawbacks are lack of flexibility (even a typo in the text can make it unrecognizable to the algorithm) and generalizability issues (developed dictionaries and patterns are most pertinent to the data for which the algorithm was developed, and thus may perform unsatisfactory at other institutions and are only applicable to the same language). Despite these limitations, there are multiple successful examples in the literature.

One of the largest and most clinically sound example is a study by van Dijk et al.[3] The authors utilized a text mining technique based on designed regular expressions for the entire EHR documentation during LoDoCo2 trial (Low Dose Colchicine for Secondary Prevention of Cardiovascular Disease)[10] prescreening and data collection phases. Mean accuracy, sensitivity, and specificity of the automatically extracted data were 88%, 81%, and 83%, respectively. The lowest accuracy was found for hypertension (62.6%), antiplatelet

therapy (68.8%), and β-blocker use (73.3%). Despite these limitations, the tool allowed for manual screening of only 20.1% of the original 92 466 patients for the trial inclusion, resulting in 82.4% of the final participants being recruited through this prescreening method, which was a remarkably time-efficient solution.

Another study by Karystianis et al[11] extracted mentions of 5 diseases, smoking status, family history, and medications from clinical notes. The project was an ambitious attempt to use less stereotypical textual data, that is, daily clinical notes. The authors utilized a rule-based approach with dictionaries developed especially for that purpose. Mean average sensitivity reached 90%. Errors were predominantly due to a lack of context analysis and unforeseen shortcuts frequently used in clinical notes. Additionally, coronary artery disease was deemed the most challenging to identify, causing many undetected cases. A diagnosis of coronary artery disease, even if not stated directly, is hinted by other diagnoses, such as a history of coronary artery bypass grafting, percutaneous coronary intervention, or myocardial infarction. Taking this into account, such clinically meaningful hierarchies have been implemented in the AssistMED tool. The authors discussed the problems of analyzing clinical notes, such as their less predictable structure, complex context analysis, and frequent jargon usage.

An NLP tool, EchoInfer, was developed to automatically extract cardiovascular structure and function data from echocardiographic reports.[12] EchoInfer achieved comparable results with an average sensitivity of 92.21% evaluated at a single institution.

Supervised machine learning Supervised machine learning methods are flexible, adaptable to new language patterns, and can achieve high accuracy with proper training. Supervised learning means that a computer needs to know the final answer, that is, whether a particular diagnosis / drug / echocardiography parameter is associated with an individual patient. In this technique, the text is preprocessed and then treated as a signal for machine learning. The computer learns to associate certain constellations of text chunks with the presence or absence of a disease. An advantage of the approach is that once the algorithm is trained, it usually works well at other institutions, meaning it is portable. However, this technique requires large samples of high-quality, labeled data to become functional. Examples available in the literature identify a few diseases at a time.

Weissler et al[13] presented an elegant solution of text-processing based on various textual data types from EHRs to recognize patients with peripheral artery disease. A majority of the texts available in EHRs contributed to training of the machine learning model (including progress notes, consults, etc.). Based on a cutoff selected by the authors, the algorithm achieved a sensitivity of 90% and specificity of 62%. The approach is exciting, as it goes beyond classifying specific text fields in the EHR and aggregates all available textual data.

Deep learning methods: large language models Large language models (LLMs) excel in language comprehension and context awareness, and thanks to their pretraining with openly available online data, they also include medical information. Pretraining means that such models can perform decently in out-of-the-box tasks. Fine-tuning is required for further improvement. Here, a big step forward is that the fine-tuning does not need so much annotated data, which is the main limitation of machine-learning approaches. Additionally, context awareness and acknowledgement of shortcuts may enable them to be effectively applied for tasks such as diagnosis and drug detection in heterogeneous data, such as daily clinical notes. This may help correctly identify drug intake status (whether taken, discontinued, halted, considered for introduction, or allergic to) or disease status (confirmed, included in differential diagnosis but not yet confirmed, or excluded). Our algorithm solely incorporated simple negation detection, limiting its capabilities in this matter, and high accuracy of our approach is partly attributed to processing of more organized textual data types in EHRs.

Even ChatGPT shows a potential for proposing ICD-10 codes from provided pieces of anonymized medical documentation and categorizing findings from echocardiographic reports, which may aid in structuring medical data. There are also LLMs pretrained on real-world medical data (MedPalm 2, https://arxiv.org/pdf/2212.13138.pdf, Clinical BERT, https://arxiv.org/abs/1904.03323, MedBERT).[14] MedPalm 2, for example, achieved a passing rate on the United States Medical Licensing Examination.[15]

With fine-tuning, LLMs could effectively perform tasks such as parameter value retrieval from magnetic resonance imaging reports, as demonstrated by Singh et al,[14] which is a task similar to that attempted in our echocardiographic parameters retrieval. Notably, the authors developed the model with just 370 human annotations by fine-tuning the BERT-LARGE model, thus overcoming a typical limitation of such projects utilizing artificial intelligence, that is, the requirement for large datasets with high-quality annotated data.

There are, however, other limitations of LLMs, for example, hallucinations. This is a well-known problem of ChatGPT, as the chatbot tends to give invented, false information to provide any answer to a question at hand, which is undesirable in research data acquisition. Furthermore, although LLMs excel in language translation, most of their training data are in English. This might compromise the results of medical data understanding and categorization for other less-represented

languages, including Polish. Suwała et al[16] already reported unsatisfactory performance of ChatGPT at the Polish board certification examination in internal medicine, despite being able to pass multiple other international exams, for example the European Exam in Core Cardiology.[17]

In summary, so far, dictionary-based and rule-based approaches have been the most popular in the medical field, but more advanced text-processing techniques are being adopted. **TABLE 4** summarizes subjective advantages and disadvantages of the discussed NLP approaches. The approaches used to develop the AssistMED tool could be described as hybrid, as we utilized dictionaries, rules, and some machine learning to achieve the presented results. Therefore, more portability can be expected than in traditional dictionary- and rule-based algorithms, although this requires future testing on data from other institutions.

Study limitations Despite optimistic results, our approach has significant limitations that must be addressed.

First, the AssistMED algorithm has only been validated in a single tertiary cardiology center, and its applicability to data from other centers is still to be determined.

Second, we limited our analysis to specific textual data types of the discharge documentation, not the entire textual data available in the EHRs. However, choosing these specific textual data types takes advantage of natural tendencies ingrained in everyday clinical practice in Poland, that is, listing clinical diagnoses descriptively and including medication dosing in discharge recommendations. These data types are more homogeneous and more straightforward to analyze than disorganized progress notes, and processing such texts brought satisfactory results. As discussed before, analysis of other data types is challenging, and published results indicate relatively moderate retrieval accuracy,[11,18] despite demonstrated research utility of such data.

Third, the AssistMED algorithm lacks advanced context analysis, its only capability is detecting negation in the diagnosis recognition module. This is largely the reason for inaccuracies reflected in the results. The source of errors includes detection of a condition that is not yet established (eg, "patient qualified for elective PCI [percutaneous coronary intervention] scheduled on..."—there was no PCI yet, but the algorithm recognized the condition), significant typos that precluded identification, identification of a drug which is not taken (eg, "please do not take bisoprolol"—the patient is not taking bisoprolol), and random algorithm errors. As discussed, LLMs are most likely to resolve such problems, as they require advanced language understanding.

Conclusions NLP tools implemented in the AssistMED project could quickly and accurately characterize patients with AF, as compared with human-based retrieval. Further improvements will likely arise by adopting LLMs for similar tasks.

SUPPLEMENTARY MATERIAL

Supplementary material is available at www.mp.pl/paim.

ARTICLE INFORMATION

ACKNOWLEDGEMENTS We especially acknowledge the help of Eliza Konstanciuk and Anna Kula from the Technology Transfer Office of the Medical University of Warsaw for their exceptional support in the project implementation on the administrative side. Special thanks also go to Tomasz Pluta and Wioletta Wąsowska-Filipek from Codifive sp. z o.o. for their valuable contribution during the project's interface design and programming phase.

FUNDING This research was supported by a noncommercial research grant Innovation Incubator 4.0 from the Medical University of Warsaw awarded by the Ministry of Science and Higher Education (to AC).

CONTRIBUTION STATEMENT Study conception and design: CM, MB, ABarwiołek, MJK, AC, GO, MG, and PB. Material preparation, data collection and analysis: CM, MK, ABarwiołek, MC, ABożym, PL, KO, and PB. Manuscript preparation: CM, ABarwiołek, and MB. All authors commented on previous versions of the manuscript as well as read and approved the final manuscript.

HOW TO CITE Maciejewski C, Ozierański K, Basza M, et al. Practical use case of natural language processing for observational clinical research data retrieval from electronic health records: AssistMED project. Pol Arch Intern Med. 2024; XX: 16704. doi:10.20452/pamw.16704

REFERENCES

1 Köpcke F, Trinczek B, Majeed RW, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak. 2013; 13: 37.

2 Maciejewski C, Ozierański K, Basza M, et al. Administrative data in cardiovascular research: a comparison of Polish National Health Fund and CRAFT Registry Data. Int J Environ Res Public Health. 2022; 19: 11964.

3 Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013; 46: 830-836.

4 Gill SK, Karwath A, Uh HW, et al. Artificial intelligence to enhance clinical value across the spectrum of cardiovascular healthcare. Eur Heart J. 2023; 44: 713-725.

5 Maciejewski C, Ozierański K, Barwiołek A, et al. 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; 185: 105380.

6 Balsam P, Gawalko M, Peller M, et al. Clinical characteristics and thromboembolic risk of atrial fibrillation patients with and without congestive heart failure. Results from the CRATF study. Medicine (Baltimore). 2018; 97: e13074.

7 Balsam P, Tyminska A, Ozieranski K, et al. Randomized controlled clinical trials versus real-life atrial fibrillation patients treated with oral anticoagulants. Do we treat the same patients? Cardiol J. 2018; 27: 590-599.

8 Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019; 7: e12239.

9 Reading Turchioe M, Volodarskiy A, Pathak J, et al. Systematic review of current natural language processing methods and applications in cardiology. Heart. 2022; 108: 909-916.

10 Nidorf SM, Fiolet ATL, Mosterd A, et al. Colchicine in patients with chronic coronary disease. N Eng J Med. 2020; 383: 1838-1847.

11 Karystianis G, Dehghan A, Kovacevic A, et al. Using local lexicalized rules to identify heart disease risk factors in clinical notes. J Biomed Inform. 2015; 58: S183-S188.

12 Nath C, Albaghdadi MS, Jonnalagadda SR. A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS One. 2016; 11: e0153749.

13 Weissler EH, Zhang J, Lippmann S, et al. Use of natural language processing to improve identification of patients with peripheral artery disease. Circ Cardiovasc Interv. 2020; 13: e009447.

14 Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021; 4: 86.

15 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023; 620: 172-180.

16 Suwała S, Szulc P, Dudek A, et al. ChatGPT fails the Polish board certification examination in internal medicine: artificial intelligence still has much to learn. Pol Arch Intern Med. 2023; 133: 16608.

17 Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023; 4: 279-281.

18 van Dijk WB, Fiolet ATL, Schuit E, et al. Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study. J Clin Epidemiol. 2021; 132: 97-105.

7. Podsumowanie i wnioski

Przedstawiony cykl prac dostarcza kompleksowego spojrzenia na współczesne metody szybkiego pozyskiwania danych do badań w dziedzinie kardiologii, z podkreśleniem roli technik procesowania tekstu w elektronicznej dokumentacji medycznej. Wypracowane w toku prac rozwiązania oparte o NLP są 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.

W publikacji pierwszej przybliżono porównanie danych opartych na analizie danych w dokumentacji medycznej przez człowieka oraz danych administracyjnych z NFZ, na przykładzie dużej kohorty pacjentów, z migotaniem przedsionków. Wyniki wskazały na istotne różnice w charakterystyce grupy pacjentów wg. kodów rozliczeniowych w porównaniu do analizy przez człowieka.

W publikacji drugiej zaprezentowane zostało autorskie rozwiązanie wytworzone w toku badań własnych- narzędzie „AssistMED" wykorzystujące techniki procesowania języka naturalnego w celu analizy określonych typów danych opisowych z EDM. Zweryfikowano dokładność wypracowanego narzędzia w porównaniu do podwójnej analizy danych przez człowieka. Przeprowadzono analizę ilościową i jakościową błędów popełnianych przez algorytm w celu precyzyjnego scharakteryzowania ograniczeń narzędzi opartych o NLP w kontekście automatycznego pozyskiwania danych.

W publikacji trzeciej zaprezentowano praktyczne użycie narzędzia „AssistMED", w celu pozyskania kompleksowej charakterystyki (rozpoznania chorobowe, stosowane leki, liczbowe parametry echokardiograficzne) dużej, retrospektywnej kohorty chorych z migotaniem przedsionków z EDM.

Wszystkie trzy publikacje prezentują jednoczasowo wiele jednostek chorobowych, a dane prezentowane są w analogiczny sposób, pozwalający na wygodne porównania wyników w poszczególnych manuskryptach. Zaprezentowano obiecujące wyniki automatycznego pozyskania danych klinicznych z wykorzystaniem NLP w oparciu o dane tekstowe z EDM. Jest to istotne, w kontekście dostępności ogromnych zasobów danych, do tego typu analiz. Na podstawie przedstawionego cyklu powiązanych tematycznie publikacji, można sformułować następujące wnioski:

- 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.

8. Oświadczenia wszystkich współautorów publikacji określające indywidualny wkład.

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

Tytuł publikacji: Administrative Data in Cardiovascular Research-A Comparison of Polish National Health Fund and CRAFT Registry Data. Int J Environ Res Public Health. 2022 Sep 22;19(19):11964. doi: 10.3390/ijerph191911964. PMID: 36231265; PMCID: PMC9565600.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabeli - opracowanie artykułu - rewizja manuskryptu	86%	06/05/2024
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	13/05/2024 dr hab. n. med. Krzysztof Ozierański Lekarz specjalista kardiologii 3168969
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
4.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	17/05/24 dr hab. n. med. [illegible] FESC [illegible] chorób wewnętrznych CARDIOLOG [illegible] specjalista EHRA 1551970
5.	Dr hab. n. med. Andrzej Śliwczyński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
6.	Dr n. med. Leszek Kraj	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	dr n. med. Leszek Kraj lekarz specjalista [illegible] ASYSTENT 20/05/2024
7.	Mgr. inż. Maciej Janusz Krajsman	- krytyczna analiza wyników - rewizja manuskryptu	1%	Zakład Informatyki Medycznej Pomorski Uniwersytet Medyczny 20/05/2024
8.	Jefte Prado Paulino	- krytyczna analiza wyników - rewizja manuskryptu	1%	mgr inż. Maciej Janusz Krajsman
9.	Dr hab. n. med. Agata Tymińska	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	A Tymińska 20/05/2024
10.	Prof. dr hab. n. med. Grzegorz Opolski	- krytyczna analiza wyników - rewizja manuskryptu	1%	Grzegorz Opolski 21/05/2024
11.	Dr hab. n. med. Andrzej Cacko	- krytyczna analiza wyników - rewizja manuskryptu	1%	17/05/2024 Andrzej Cacko
12.	Prof. dr hab. n. med. Marcin Grabowski	- krytyczna analiza wyników - rewizja manuskryptu	1%	20/05/2024
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	KIEROWNIK Oddział Kliniczny Elektrokardiologii I Klinika Kardiologii Centralny Szpital Kliniczny UCK WUM prof. dr hab. n. med. Paweł Balsam

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

Tytuł publikacji: Administrative Data in Cardiovascular Research-A Comparison of Polish National Health Fund and CRAFT Registry Data. Int J Environ Res Public Health. 2022 Sep 22;19(19):11964. doi: 10.3390/ijerph191911964. PMID: 36231265; PMCID: PMC9565600.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabeli - opracowanie artykułu - rewizja manuskryptu	86%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
4.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
5.	Dr hab. n. med. Andrzej Śliwczyński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	Signed by / Podpisano przez: Andrzej Marek Śliwczyński Date / Data: 2024-05-23 08:46
6.	Dr n. med. Leszek Kraj	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
7.	Mgr. inż. Maciej Janusz Krajsman	- krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Jefte Prado Paulino	- krytyczna analiza wyników - rewizja manuskryptu	1%	
9.	Dr hab. n. med. Agata Tymińska	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Grzegorz Opolski	- krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Dr hab. n. med. Andrzej Cacko	- krytyczna analiza wyników - rewizja manuskryptu	1%	
12.	Prof. dr hab. n. med. Marcin Grabowski	- krytyczna analiza wyników - rewizja manuskryptu	1%	
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

Tytuł publikacji: Administrative Data in Cardiovascular Research-A Comparison of Polish National Health Fund and CRAFT Registry Data. Int J Environ Res Public Health. 2022 Sep 22;19(19):11964. doi: 10.3390/ijerph191911964. PMID: 36231265; PMCID: PMC9565600.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabeli - opracowanie artykułu - rewizja manuskryptu	86%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
4.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
5.	Dr hab. n. med. Andrzej Śliwczyński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
6.	Dr n. med. Leszek Kraj	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
7.	Mgr. inż. Maciej Janusz Krajsman	- krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Jefte Prado Paulino	- krytyczna analiza wyników - rewizja manuskryptu	1%	25.05.2024
9.	Dr hab. n. med. Agata Tymińska	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Grzegorz Opolski	- krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Dr hab. n. med. Andrzej Cacko	- krytyczna analiza wyników - rewizja manuskryptu	1%	
12.	Prof. dr hab. n. med. Marcin Grabowski	- krytyczna analiza wyników - rewizja manuskryptu	1%	
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

Tytuł publikacji: Administrative Data in Cardiovascular Research-A Comparison of Polish National Health Fund and CRAFT Registry Data. Int J Environ Res Public Health. 2022 Sep 22;19(19):11964. doi: 10.3390/ijerph191911964. PMID: 36231265; PMCID: PMC9565600.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabeli - opracowanie artykułu - rewizja manuskryptu	86%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
4.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
5.	Dr hab. n. med. Andrzej Śliwczyński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
6.	Dr n. med. Leszek Kraj	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
7.	Mgr. inż. Maciej Janusz Krajsman	- krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Jefte Prado Paulino	- krytyczna analiza wyników - rewizja manuskryptu	1%	
9.	Dr hab. n. med. Agata Tymińska	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Grzegorz Opolski	- krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Dr hab. n. med. Andrzej Cacko	- krytyczna analiza wyników - rewizja manuskryptu	1%	
12.	Prof. dr hab. n. med. Marcin Grabowski	- krytyczna analiza wyników - rewizja manuskryptu	1%	
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

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. doi: 10.1016/j.ijmedinf.2024.105380. Epub 2024 Feb 19. PMID: 38447318.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - opracowanie artykułu - rewizja manuskryptu	81%	08/05/24
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	13/05/2024 dr hab. n. med. Krzysztof Ozierański Lekarz specjalista kardiologii
3.	Lic. Adam Barwiołek	- opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	2%	
4.	Lek. Mikołaj Basza	- stworzenie projektu badania - opracowanie metodologii - krytyczna analiza wyników - rewizja manuskryptu	2%	
5.	Aleksandra Bożym	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	17/05/24 Aleksandra Bożym
6.	Michalina Ciurla	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	17/05/24 ASYSTENT Zakład Informatyki Medycznej i Telemedycyny
7.	Mgr. Inż. Maciej Janusz Krajsman	- analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	1%	Maciej Janusz Krajsman
8.	Dr. n. med. Magdalena Maciejewska	- gromadzenie danych - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	1%	
9.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	17/05/24 Specjalista chorób wewnętrznych KARDIOLOGIA
10.	Prof. dr hab. n. med. Grzegorz Opolski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	1551970 Grzegorz Opolski 20/05/2024 21/05/2024
11.	Prof. dr hab. n. med. Marcin Grabowski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
12.	Dr hab. n. med. Andrzej Cacko	- stworzenie projektu badania - opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	2%	17/05/2024 Andrzej Cacko
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	KIEROWNIK Oddział Kliniczny Elektrokardiologii I Klinika Kardiologii Centralny Szpital Kliniczny UCK WUM prof. dr hab. n. med. Paweł Balsam

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

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. doi: 10.1016/j.ijmedinf.2024.105380. Epub 2024 Feb 19. PMID: 38447318.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - opracowanie artykułu - rewizja manuskryptu	81%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lic. Adam Barwiołek	- opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	2%	
4.	Lek. Mikołaj Basza	- stworzenie projektu badania - opracowanie metodologii - krytyczna analiza wyników - rewizja manuskryptu	2%	
5.	Aleksandra Bożym	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
6.	Michalina Ciurla	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
7.	Mgr. inż. Maciej Janusz Krajsman	- analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Dr. n. med. Magdalena Maciejewska	- gromadzenie danych - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	1%	
9.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Grzegorz Opolski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Prof. dr hab. n. med. Marcin Grabowski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
12.	Dr hab. n. med. Andrzej Cacko	- stworzenie projektu badania - opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	2%	
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

PODPIS ZAUFANY
ADAM PATRYK
BARWIOŁEK
23.05.2024 07:22:04 [GMT+2]
Dokument podpisany elektronicznie podpisem zaufanym

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

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. doi: 10.1016/j.ijmedinf.2024.105380. Epub 2024 Feb 19. PMID: 38447318.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - opracowanie artykułu - rewizja manuskryptu	81%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lic. Adam Barwiołek	- opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	2%	
4.	Lek. Mikołaj Basza	- stworzenie projektu badania - opracowanie metodologii - krytyczna analiza wyników - rewizja manuskryptu	2%	
5.	Aleksandra Bożym	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
6.	Michalina Ciurla	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
7.	Mgr. inż. Maciej Janusz Krajsman	- analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Dr. n. med. Magdalena Maciejewska	- gromadzenie danych - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	1%	
9.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Grzegorz Opolski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Prof. dr hab. n. med. Marcin Grabowski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
12.	Dr hab. n. med. Andrzej Cacko	- stworzenie projektu badania - opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	2%	
13.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

PODPIS ZAUFANY
MIKOŁAJ
BASZA
26.05.2024 19:18:24 [GMT+2]
Dokument podpisany elektronicznie podpisem zaufanym

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

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. doi: 10.20452/pamw.16704. Epub ahead of print. PMID: 38501989.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - opracowanie artykułu - rewizja manuskryptu	82%	06/05/2024
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	13/05/2024 dr hab. n. med. Krzysztof Ozierański Lekarz specjalista kardiologii
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - opracowanie metodologii - krytyczna analiza wyników - rewizja manuskryptu	2%	
4.	Lic. Adam Barwiołek	- opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	2%	
5.	Michalina Ciurla	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	17/05/24
6.	Aleksandra Bożym	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	17/05/24 Aleksandra Bożym ASYSTENT Zakład Informatyki Medycznej i Telemedycyny
7.	Mgr. Inż. Maciej Janusz Krajsman	- analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	1%	2024/05/17 mgr inż. Maciej Janusz Krajsman
8.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	17/05/24 dr hab. n. med. Piotr Lodziński Specjalista kardiologii 15519
9.	Prof. dr hab. n. med. Grzegorz Opolski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	21/05/2024 Grzegorz Opolski
10.	Prof. dr hab. n. med. Marcin Grabowski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	20/05/2024
11.	Dr hab. n. med. Andrzej Cacko	- stworzenie projektu badania - opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	2%	17/05/2024 Andrzej Cacko
12.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	KIEROWNIK Oddział Kliniczny Elektrokardiologii I Klinika Kardiologii Centralny Szpital Kliniczny UCK WUM Prof. dr hab. n. med. Paweł Balsam

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

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. doi: 10.20452/pamw.16704. Epub ahead of print. PMID: 38501989.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - opracowanie artykułu - rewizja manuskryptu	82%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - opracowanie metodologii - krytyczna analiza wyników - rewizja manuskryptu	2%	
4.	Lic. Adam Barwiołek	- opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	2%	
5.	Michalina Ciurla	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
6.	Aleksandra Bożym	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
7.	Mgr. inż. Maciej Janusz Krajsman	- analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
9.	Prof. dr hab. n. med. Grzegorz Opolski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Marcin Grabowski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Dr hab. n. med. Andrzej Cacko	- stworzenie projektu badania - opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	2%	
12.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

PODPIS ZAUFANY
ADAM PATRYK
BARWIOŁEK
23.05.2024 07:27:13 [GMT+2]
Dokument podpisany elektronicznie podpisem zaufanym

Oświadczenie o współautorstwie w publikacji oraz wyrażenie zgody na wykorzystanie pracy jako części rozprawy doktorskiej lek. Cezarego Maciejewskiego

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. doi: 10.20452/pamw.16704. Epub ahead of print. PMID: 38501989.

Lp.	Współautor	Rodzaj wkładu merytorycznego	Procentowy wkład w publikację	Data, podpis
1.	Lek. Cezary Maciejewski	- stworzenie projektu badania - opracowanie metodologii - gromadzenie danych - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - opracowanie artykułu - rewizja manuskryptu	82%	
2.	Dr hab. n. med. Krzysztof Ozierański	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
3.	Lek. Mikołaj Basza	- stworzenie projektu badania - opracowanie metodologii - krytyczna analiza wyników - rewizja manuskryptu	2%	
4.	Lic. Adam Barwiołek	- opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - opracowanie rycin i tabel - rewizja manuskryptu	2%	
5.	Michalina Ciurla	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
6.	Aleksandra Bożym	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	2%	
7.	Mgr. inż. Maciej Janusz Krajsman	- analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	1%	
8.	Dr hab. n. med. Piotr Lodziński	- gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu	1%	
9.	Prof. dr hab. n. med. Grzegorz Opolski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
10.	Prof. dr hab. n. med. Marcin Grabowski	- stworzenie projektu badania - krytyczna analiza wyników - rewizja manuskryptu	1%	
11.	Dr hab. n. med. Andrzej Cacko	- stworzenie projektu badania - opracowanie metodologii - analiza statystyczna - krytyczna analiza wyników - rewizja manuskryptu	2%	
12.	Prof. dr hab. n. med. Paweł Balsam	- stworzenie projektu badania - gromadzenie danych - krytyczna analiza wyników - rewizja manuskryptu - nadzór merytoryczny	2%	

9. Bibliografia

1. Masic I., M. Miokovic, and B. Muhamedagic Evidence based medicine - new approaches and challenges. Acta Inform Med. 2008; 16: 219-225.
2. Jones D.S. and S.H. Podolsky The history and fate of the gold standard. Lancet. 2015; 385: 1502-1503.
3. Sibbald B. and M. Roland Understanding controlled trials. Why are randomised controlled trials important? BMJ. 1998; 316: 201.
4. Guyatt G.H., A.D. Oxman, G.E. Vist, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008; 336: 924-926.
5. Harrer M., P. Cuijpers, L.K.J. Schuurmans, et al. Evaluation of randomized controlled trials: a primer and tutorial for mental health researchers. Trials. 2023; 24: 562.
6. Balsam P., K. Ozieranski, A. Tyminska, et al. Comparison of clinical characteristics of real-life atrial fibrillation patients treated with vitamin K antagonists, dabigatran, and rivaroxaban: results from the CRAFT study. Kardiol Pol. 2018; 76: 889-898.
7. Fernainy P., A.A. Cohen, E. Murray, et al. Rethinking the pros and cons of randomized controlled trials and observational studies in the era of big data and advanced methods: a panel discussion. BMC Proceedings. 2024; 18: 1.
8. Mc Cord K.A. and L.G. Hemkens Using electronic health records for clinical trials: Where do we stand and where can we go? Cmaj. 2019; 191: E128-e133.
9. Cowie M.R., J.I. Blomster, L.H. Curtis, et al. Electronic health records to facilitate clinical research. Clinical research in cardiology : official journal of the German Cardiac Society. 2017; 106: 1-9.
10. Wang Y., L. Kung, and T.A. Byrd Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change. 2018; 126: 3-13.
11. Quach S., C. Blais, and H. Quan Administrative data have high variation in validity for recording heart failure. Can J Cardiol. 2010; 26: 306-312.
12. Kaspar M., G. Fette, G. Güder, et al. Underestimated prevalence of heart failure in hospital inpatients: a comparison of ICD codes and discharge letter information. Clinical research in cardiology : official journal of the German Cardiac Society. 2018; 107: 778-787.
13. Hersh W.R., M.G. Weiner, P.J. Embi, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical care. 2013; 51: S30-S37.
14. Parrinello C.M., K.N. Seidl-Rathkopf, A.B. Bourla, et al. Comparison of Structured Versus Abstracted Comorbidities Using Oncology EHR Data from Cancer Patients in the Flatiron Health Network. Value in Health. 2018; 21: S14.
15. Li R., B. Hu, F. Liu, et al. Detection of Bleeding Events in Electronic Health Record Notes Using Convolutional Neural Network Models Enhanced With Recurrent Neural Network Autoencoders: Deep Learning Approach. JMIR Med Inform. 2019; 7: e10788.
16. Patterson O.V., M.S. Freiberg, M. Skanderson, et al. Unlocking echocardiogram measurements for heart disease research through natural language processing. BMC Cardiovascular Disorders. 2017; 17: 151.
17. Mandhan S., *Numerical Atrribute Extraction from Clinical Texts*. 2015.
18. Elkin P., S. Mullin, C. Crowner, et al. Abstract 14929: Strokes Prevented: Biosurveillance of NVAF Patient Cohorts CHA₂DS₂-VASc and HAS-BLED Scores Using Natural Language Processing and SNOMED CT. Circulation. 2017; 136: A14929-A14929.