**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.