applied sciences

MDPI

Review

A Review of Approaches to Standardizing Medical Descriptions for Clinical Entity Recognition: Implications for Artificial Intelligence Implementation

Michał Paweł Wierzbicki, Barbara Anna Jantos and Michał Tomaszewski *

Department of Computer Science, Opole University of Technology, 45-758 Opole, Poland
* Correspondence: m.tomaszewski@po.edu.pl

Abstract: This article reviews the current state of standardization in specific areas of the medical sector in Poland, focusing on terminology and the unique context of the Polish language. The primary objective was to analyze the existing resources and examine possibilities, challenges, and opportunities associated with integrating Artificial Intelligence, particularly natural language processing methods, into the healthcare system. The additional goal of this review was to place Poland in the international context by comparing the current state of the Polish standardization of healthcare with those of selected countries with more and less developed systems. The exploration highlights the main challenges that impact integration, including the specificity of the language and challenges in transferring knowledge from other languages, lack of communication between parties, and lack of stakeholder involvement in the standardization processes. This review also presents potential solutions to the mentioned challenges and provides insights into future directions, possibilities, proposals, and recommendations for all stakeholders. The practical application of this research extends beyond Poland. Many countries with underrepresented languages face similar challenges in clinical data processing, and the advances in CER for Polish could serve as a model for implementing AI-driven solutions in these regions. By refining CER models and adapting them to diverse linguistic and healthcare contexts, this research can foster improvements in patient care, medical research, and healthcare administration on a global scale.

Keywords: clinical entity recognition; natural language processing; standardization of medical descriptions; Polish language; artificial intelligence; healthcare

check for updates

Citation: Wierzbicki, M.P.; Jantos, B.A.; Tomaszewski, M. A Review of Approaches to Standardizing Medical Descriptions for Clinical Entity Recognition: Implications for Artificial Intelligence Implementation. *Appl. Sci.* **2024**, *14*, 9903. https://doi.org/10.3390/app14219903

Academic Editor: Alexander N. Pisarchik

Received: 29 September 2024
Revised: 20 October 2024
Accepted: 25 October 2024
Published: 29 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

This article has a dual focus, primarily aiming to investigate and evaluate the state of medical description standardization in Poland. Concurrently, it seeks to review current knowledge and advancements in developing methods for recognizing medical descriptions written in Polish.

The first purpose—recognition of the current state and analysis of directions of changes in the standardization of medical descriptions in Poland—is to shed light on the existing practices of standardization in the context of medical descriptions within the Polish healthcare system. The Polish case can serve as a valuable example for healthcare systems in other countries. The medical field heavily relies on standardized terminologies, controlled vocabularies, and ontologies to ensure consistency, data exchange, and interoperability among various stakeholders. This article aims to provide a comprehensive snapshot of the current standardization practices in Poland's healthcare realm. This involves an in-depth exploration of the terminologies, ontologies, and controlled vocabularies in active use. It seeks to identify the challenges and limitations encountered in the standardization of medical descriptions, tackling issues ranging from terminological ambiguities to the necessity for regular updates and widespread adoption. This article endeavors to present and

assess potential strategies for improving the standardization of medical descriptions in the Polish healthcare landscape. This may encompass suggestions for adopting internationally recognized terminologies or crafting locally pertinent standards. A vital aspect of this article is to underscore the pivotal significance of standardization in healthcare. This is especially pertinent in clinical entity recognition (CER), where standardized terms are essential for effective data exchange, clinical decision making, and medical research.

The second significant goal of this article is to take a deep dive into the current state of knowledge about developing methodologies designed for CER in the Polish language. This part of the article aspires to scrutinize the existing knowledge base and research in recognizing medical descriptions, particularly when composed in Polish. This inquiry includes exploring the existing algorithms, techniques, and practical applications. Through this investigation, this article aims to pinpoint areas where further research and innovation are needed for effective medical description recognition in Polish and CER in general.

Many publications written solely in Polish have been published only in industry-specific journals and do not appear in indexed databases. The primary aim of this review was to compile condensed knowledge about the state of CER in Poland, as this knowledge is inaccessible to non-Polish speakers. The lack of availability of these studies in international databases, such as PubMed or Scopus, limits the exchange of experiences and findings between researchers from different countries, thereby hindering the development of global solutions for CER. Introducing research based on Polish-language sources aims to fill this gap and promote local achievements within the broader scientific community.

While some various works on Name Entity Recognition in the Polish language have been published (e.g., [1–8]), they rarely focus on medical texts. The lack of widely available research on medical texts in Polish further complicates the development of AI-based health systems. Despite their effectiveness in general applications, available Name Entity Recognition (NER) models developed for other fields are often not tailored to the specific language and medical context.

This article employed a review approach as the research method, allowing for a broad exploration of the literature on standardizing medical descriptions for CER in Poland. The review process followed several stages:

- Formulating research questions. This review began by identifying specific research questions to uncover existing work on CER in the Polish language, standardization processes, and potential applications in AI-based medical methods. The primary questions focused on are as follows:
  - What are the existing methodologies for CER in Polish medical texts?
  - How can standardization efforts in medical descriptions improve the effectiveness of CER in the Polish healthcare context?
  - What are the gaps and challenges in developing CER systems tailored to the Polish language?
  - What are the standardization directions in Polish healthcare systems?
- Source selection. The source selection process involved a combination of local and international databases to ensure the broadest possible inclusion of relevant studies: Polish industry journal databases, international databases (Google Scholar, PubMed, and Scopus), and local archives.
- Inclusion criteria. Articles and studies were selected based on specific inclusion criteria to ensure they addressed the research questions directly and contributed to understanding CER in Polish. The criteria included the following:
  - Studies specifically focused on Polish-language medical texts or methodologies for recognizing clinical entities in Polish;
  - Articles that discussed the standardization of medical descriptions in Polish, including terminologies, controlled vocabularies, and ontologies;
  - Research that explored the development of AI-based CER systems or similar natural language processing (NLP) technologies in Polish healthcare contexts;

  - Works that have been published in the last twenty years.
- Exclusion criteria. To maintain the focus on relevant research and avoid non-applicable studies, certain exclusion criteria were applied:
  - Articles that focused on general-purpose Name Entity Recognition methods without applications to medical texts or the Polish language were excluded;
  - Studies that did not explore the specific context of Polish healthcare or were only tangentially related to the standardization of medical data;
  - Publications with outdated methodologies or insufficient empirical data were excluded to maintain the review's quality.
- Data analysis and synthesis. The articles chosen were analyzed in depth to extract critical insights regarding methodologies, challenges, and results in implementing CER systems. The synthesis focused on understanding CER's technical aspects and the broader implications of medical description standardization in Poland. This analysis provided a clearer view of the existing landscape and identified areas where future research is necessary, especially in underexplored areas like Polish-language medical text annotation and the creation of annotated datasets.

This article begins with an overview of the transformation of the healthcare system in Poland, providing a background on the current state and evolution of healthcare policies and infrastructure. The following section focuses on CER and other NLP methods for the Polish language, discussing relevant studies and previous works in the area. This is followed by a comprehensive look at challenges and barriers that impede the adoption of NLP and AI technologies in Polish healthcare, highlighting issues like data availability and regulatory limitations.

The "Discussion" section presents new proposals and possibilities for overcoming these challenges for Polish and other countries' healthcare systems, including a strategy for the coming years, emphasizing the potential of Large Language Models (LLMs). It also explores future directions for research and development in the field. The "Recommendations" section provides actionable steps to foster innovation, collaboration, and data standardization, leading into the "Conclusions", which summarizes the essential findings and their implications for the future of healthcare technology.

2. Transformation of the Healthcare System in Poland: Current State

The transformation of the Polish healthcare system can be anchored in the fall of communism in 1989. For the next decade, an amplified effort was put toward moving from the Siemaszko Model to a decentralized healthcare system [9]. The Siemaszko Model [10] was a holdover system from the former system and was an example of the centralized model in which the government financed it fully and there was no private healthcare sector. In 1999, the Act of 6 February 1997 on Universal Health Insurance came into effect, effectively institutionalizing an insurance-based model [11]. The next milestone was in 2003 when the Act of 23 January 2003 on Universal Health Insurance in the National Health Fund came into effect, introducing unified rules for financing healthcare services and ensuring equal access to health services for insured individuals [12]. The year 2011 marks a step toward digitizing medical records in the form of [13].

Since 2021, Poland has operated the P1 System, the "Electronic Platform for Collection, Analysis, and Sharing of Digital Resources about Medical Events", consolidating comprehensive information on medical events [14]. In Poland, adopting the International Statistical Classification of Diseases and Related Health Problems 11th revision (ICD-11) is an ongoing process [15]. On 24 November 2023, pre-release version ICD-11 in the Polish language was released, which concluded phase one of the implementation of ICD-11. The implementation process for ICD-11 in Poland is anticipated to continue until at least 2025, possibly extending until 2026 [16]. The deadline for the whole project is 30 November 2027, meaning that ICD-11 is fully incorporated within the Polish healthcare system and adapted to international standards [17]. The P1 system aligns with the concept of

interoperability as defined in documents [18,19]. As outlined in [20], interoperability is attainable through the following:

- Unification involves utilizing the same set of norms, standards, and procedures by all parties.
- Interchangeability, signifying the ability to replace a product, service, or process without interrupting information exchange between parties, simultaneously meeting all relevant system requirements.
- Compatibility, indicating the utility of products, processes, and services that meet specified requirements without adverse side effects.

The P1 system (Figure 1) adheres to standards such as ICD-10 [21], Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [22], and Logical Observation Identifiers Names and Codes (LOINC) [23], establishing itself as an interoperable system. By embracing globally recognized standards, the P1 system facilitates the accurate coding of patient data, seamless sharing, and integration into Electronic Health Records (EHRs). Furthermore, the P1 system supports interoperability based on IHE recommendations by implementing recommended profiles [24]. In alignment with Health Level 7 (HL7) recommendations and the Digital Imaging and Communication in Medicine (DICOM) or ISO 12052 [25], the P1 system ensures interoperability and remains current with global healthcare practice. On top of that, the Polish implementation incorporates an HL7 Clinical Document Architecture (HL7 CDA), which is "a document markup standard that specifies the structure and semantics of clinical documents" for exchange between healthcare providers and patients. It defines a clinical document structure as having the following six characteristics: (1) Persistence, (2) Stewardship, (3) Potential for authentication, (4) Context, (5) Wholeness, and (6) Human readability [26]. It provides 284 templates and a dictionary of medical terms used within documentation [27].

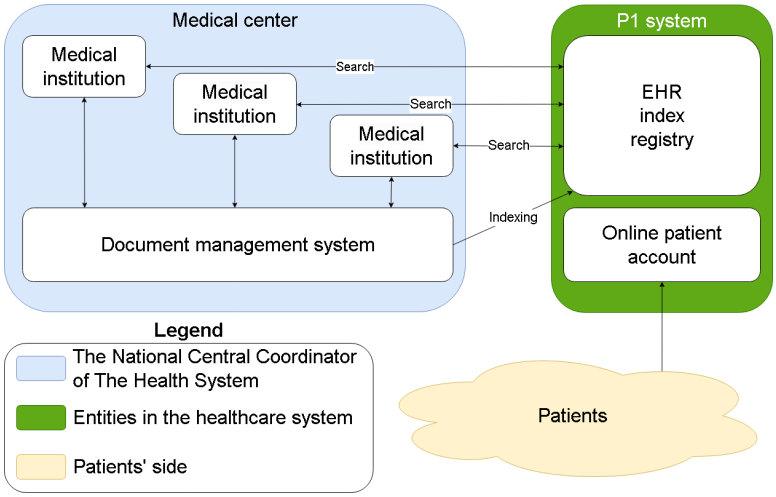

Figure 1. Data flow diagram between the medical entity and the P1 system.

In [28], the author briefly describes the roles and development of six systems in healthcare. In 2000, the Ministry of Health created the Center for Information Systems in Healthcare (Centrum Systemów Informacyjnych Ochrony Zdrowia) tasked with monitoring and constructing communication and informatics systems in the healthcare field as well as analyzing undertaken operations [28]. The main goal was to develop six complementary systems spanning the healthcare field. Those systems are [28] as follows:

- P1—Electronic Platform for Collecting, Analyzing, and Sharing Digital Resources on Medical Events—a project aimed at creating a central information system that integrates healthcare entities.
- P2—Platform for sharing medical records with private healthcare facilities (including private medical offices).

- P3—Healthcare Registry Platform—a project involving the integration of registries and the sharing of data contained in them with other information systems.
- P4—A package of small integration projects related to improving and integrating information systems in healthcare and public information systems.
- P5—Electronic platform for telemedical consultation services—a project covering hospitals and specialized centers, enabling remote medical consultations for rare and complex conditions or surgeries.
- P6—Enabling the Polish Liaison Institution to transmit data on medical events abroad—a project for the National Health Fund (Narodowy Fundusz Zdrowia) headquarters, which settles the costs of healthcare services provided to foreigners.

The system is embedded within legal frameworks that delineate the principles of its operation and define and clarify concepts that lack international definitions. Authors clarify legal matters within the terminology in [29,30]. One problem that arises from terminology disorder is that without a clear vision, there can be many interpretations of issues. For example, one of the most basic definitions is the definition of a description of diagnosis tests. Document [31] defines what can be considered a part of EHRs and which data have to be included. The value §1 enumerates what constitutes an EHR, and points 4 and 5 of this paragraph state, respectively, (4) laboratory test results along with a description; (5) a description of diagnostic tests other than those indicated in point 4.

Parties that have participated in the consultation process of this legislation have raised concerns about the ambiguity of the term "description". In the work [32], in the table entitled 'Table of comments—a reference to comments', the authors proposed clarification on this issue: "In the context of diagnostic tests, the term 'description of diagnostic tests' refers to the description of tests as defined by the regulations of the Minister of Health regarding the types, scope, and templates of medical documentation, as well as the manner of its processing. This includes a verbal description along with commentary on the test results, taking into account key information relevant to the treatment and diagnosis of the patient. On the other hand, laboratory test results consist of numerically or verbally expressed parameters" and "laboratory test results consist of parameters expressed numerically or verbally".

Each healthcare domain has its acts that provide a framework for this domain to operate within and regulate how it operates. We will take a closer look at the legal framework surrounding two particular fields—radiology and laboratory testing.

The Polish healthcare system adapts RADS standardization, but it also provides a plethora of legislation for radiology. Since this field operates with dangerous radiation, it has become one of the healthcare sector's most standardized and regulated parts. Many countries have developed good practices for the radiology field; for example, the American College of Radiology has proposed a set of good practices for the radiology field within the USA. In Poland, a similar result was obtained by [33]. It contains the list of standardized radiological procedures in the field. It also provides necessary and sufficient conditions for the proper test, including but not limited to the purpose of the procedure, details of the person carrying out the test, date of the test, list of medical conditions that said test applies to, and requirements and parameters for devices. It lists 214 radiological examinations, 202 pediatric radiological examinations, 146 radiological procedures, 58 pediatric radiological procedures, and 50 radiological procedures in cardiology. This regulation aims to standardize procedures within radiology, yet some rules are not strictly enforced. Let us examine the point "Criteria for correctly presenting examination results and their description, including criteria for properly taken X-ray images". This point appears for every procedure, yet it states that a description is sometimes not obligatory. In some cases, it states general rules and recommendations about what the said description should contain, yet again, in some cases, there is no mention of a description and all regulations regarding the image.

Standardization within the laboratory testing field is a dynamic process that has only grown in importance in the last couple of decades. Technological advancements and scien-

tific research both have heavily impacted the course of the improvement. Regulatory bodies, both those with global reach and national governance bodies and third-party agencies, frequently update norms and regulations for standards and requirements in laboratory testing. While data standardization might not be a priority, it certainly benefits those efforts. In Poland, we can point to at least two acts issued by governing bodies and one set of recommendations created by the National Chamber of Laboratory Diagnosticians. In 2016, the government issued an announcement from the Minister of Health regarding quality standards for medical diagnostic laboratories and microbiological laboratories [34]. Other acts that regard this matter and regulate the issue are included in this document. It describes procedures, requirements, and ways of operating laboratories. On 15 September 2022, the Act on Laboratory Medicine [35] was issued. It is one of Poland's most recent and up-to-date acts regarding standardization in this field. The Act specifies:

1. The principles and conditions for performing activities in laboratory medicine;
2. The principles and conditions for practicing the profession of laboratory diagnostician;
3. The principles of supervision and control of medical diagnostic laboratories;
4. The principles of organization and operation of the self-government of laboratory diagnosticians.

Document [36] is a set of recommendations from 2021 created by the National Chamber of Laboratory Diagnosticians. It outlines practical challenges that one may face in this field and how to overcome them concerning EHRs, as well as provides pragmatic solutions and good practices in that regard. It is a valuable resource that shows the perspective of healthcare practitioners. It also illustrates the gap between the legislative view of the matter and the reality that the healthcare field faces.

3. CER and Other NLP Methods for the Polish Language: Related Works

For this article, a review of CER and other NLP methods for the Polish language was conducted, with the results presented in Table 1. Papers were selected through searches of the PubMed database and Google Scholar. The search queries included the terms "machine learning" and "Polish" or "clinical entity recognition" combined with "Polish". No restrictions were applied regarding the publication date.

Before the topic of CER in the Polish language came into the picture, A. Mykowiecka and M. Marciniak [37] proposed automatic spelling correction for mammography reports. The goal of the suggested algorithm was to detect words and expressions that were written mistakenly—the types of errors that were detected were Polish diacritics, typical spelling errors, punctuation, and numbers and dates. To achieve that result, the authors created a domain dictionary that consists of words from medical reports from three Polish hospitals. This dictionary also contains the number of occurrences of specific words in the training set. The authors also prepared a second dataset, which included the Bigram model. Based on those and using minimal distance, the algorithm suggested how the word (or expression) should be written if it was detected as incorrect. The phrase was marked as incorrect if it could not be found in the domain dictionary. Even if this problem is not CER, it still addresses the issue of reports written as free speech, often with mistakes.

Another attempt to use natural language processing in Polish medical reports can be found in [38]. The authors proposed an Information Extraction (IE) system based on real data from Polish healthcare institutions. In this case, custom ontology was created because none of the existing ontologies suited the task. This ontology was then used in rule-based IE tasks. Its purpose was to recognize attributes in texts correctly. Attributes were specified based on their occurrence in the test dataset. It is highlighted that creating such a system was challenging because of Polish grammar, ambiguous terms, and negations in texts. Moreover, the rule-based approach requires a lot of manual work and depends on the quality of the texts. For the tested dataset, the authors achieved high-precision results (close to 100%) but lower recall (for one of the attributes—87.37%). However, such a method can be applied only to a very narrow domain that needs to be well described by experts.

Table 1. Review of key publications in Polish medical NLP.

Publication	Task	Dataset	Approach	Results	Issues	Dataset Size (Number of Records)
A. Mykowiecka and M. Marciniak (2007) [37]	Spelling correction for mammography reports	Medical reports from 3 Polish hospitals	Domain dictionary with Bigram model	Detected misspellings, suggested corrections	Polish diacritics, spelling, punctuation, and numbers/dates	1400
A. Mykowiecka and M. Marciniak (2009) [38]	Information Extraction (IE) system for healthcare texts	Real data from Polish healthcare institutions	Rule-based IE, custom ontology	High precision (close to 100%), recall for one attribute: 87.37%	Polish grammar, ambiguity, negations, domain-specific texts	1473
Dobrakowski et al. (2021) [39]	Grouping medical records for post-visit recommendations	Free-text notes from ~100,000 clinical visits	GloVe embeddings, k-means, hierarchical clustering	Grouped similar visits for treatment suggestions	Narrow domain, free-text records	100,000
Anetta et al. (2021) [40]	General entity recognition for medical texts	23M+ Polish words from cardiology reports (pl_ehr_cardio)	PolDeepNer2, spaCy, Spark NLP	Exceeded expectations for PolDeepNer2	Limited categories, generic tools, narrow application in medical context	50,000
Anetta et al. (2022) [41]	ICD-10 diagnosis prediction from EHRs	50,465 EHRs from cardiology (Leszek Giec Upper Silesian Medical Centre)	RoBERTa, BERT (Polish/ multilingual)	Accuracy: 68.79% (admission), 78.64% (discharge)	Class imbalance, variation in diagnosis	50,465
Obuchowski et al. (2023) [42]	Information Extraction from radiology reports	Radiology reports (University Clinical Centre, Gdansk)	Polish-roberta, distilroberta, longformer, HerBERT, mLuke	mLuke achieved best F1 score	Lack of large annotated datasets, need for domain-specific models	1200
M. Marciniak et al. (2011) [43]	Creating annotated medical corpus	460 hospital discharge reports (diabetes patients)	TEI P5 guidelines	348,461 tokens, 3 levels of annotation (semantic, morphological, and tokens)	Manual correction process ongoing	460
Marcińczuk et al. (2019) [44]	Review of NER tools and corpora for Polish language	Multiple corpora (NKJP, KPWr, CEN, PST)	PolDeepNer (Bi-GRU with CRF), KPWr n82 NER	Recognizes nested entities and 82 entity types	General and domain-specific NER tools for Polish NLP	Not given

In work [39], the authors present explainable machine learning for unsupervised models that divide medical records into groups. The used corpus consists of free-text notes written in Polish from about 100,000 clinical visits. This task aimed to group visits in clusters based on post-visit descriptions made by a doctor and then—as an application of achieved results—use information about similar visits to propose further treatment. The authors did not use any general model for the Polish language because of the narrow domain. Instead, they created embeddings for medical concepts with the GloVe algorithm. Those embeddings were used in vector representations of visits, so clustering using k-means and hierarchical clustering with Ward's method for merging clusters was possible. The authors chose to find similarities between visits, so they did not need to conduct classic named entity recognition tasks. However, they had to bother with Polish clinical records.

In the CER field, in publication [40], the authors discuss if a general entity recognition system could be used in the medical domain for the Polish language. They conducted experiments using a custom dataset containing more than 23,000,000 Polish words from cardiology reports from the Medical University of Silesia in Katowice, Poland, called pl_ehr_cardio. The number of detected entities needed to be limited and mapped because of the use of generic tools such as PolDeepNer2, spaCy, or Spark NLP entity_recognizer_md. The final categories were person, organization, location, and miscellaneous. Even though

the tools were not designed for CER tasks, the authors claim that they met their expectations and even exceeded them for PolDeepNer2. However, the suggested categories are not tailored to the medical context, so the application of such an approach is very narrow.

Other studies on cardiological records were conducted by the authors in [41]. They aimed to predict the ICD-10 diagnosis category based on Electronic Health Records. In their experiment, the dataset was obtained from the Asseco Medical Management Solutions (Asseco Poland S.A., Rzeszow, Poland) EHR system, including data from patients from the third Department of Cardiology, Leszek Giec Upper Silesian Medical Centre of the Medical University of Silesia in Katowice (GCM), Poland, and consists of 50,465 records. They were written as free text but with four sections—1. admission, 2. physical, 3. discharge, and 4. recommendation. The authors trained the models RoBERTa and BERT on Polish and multilingual data transformer language. It was necessary to handle class imbalance and variation in diagnosis. It was possible to obtain optimistic results. For models using admission data, accuracy in the best case equals 68.79%, and for models based on discharge data, the best accuracy was 78.64%. The authors concluded that systems able to attribute with ICD code automatically could benefit healthcare systems; however, its accuracy still needs improvement.

Applying machine learning to the medical field in specific languages is limited because of the lack of adequately annotated datasets. Therefore, CER for the Polish language is still a relatively uncovered topic. This problem is also highlighted in [42], where the authors proposed a model for information extraction. They used a dataset created from radiology reports written narratively and collected at the University Clinical Centre in Gdansk, Poland. Annotations were performed by radiologists and consisted of 44 tags for observation in chest X-rays. The reports were anonymized and preprocessed. In experiments, the authors implemented Polish-roberta-base-v2, Polish-distilroberta, Polish longformer, HerBERT, and mLuke language models, with mLuke obtaining the best F1 score. Further work could focus on training domain-specific language models with larger datasets to improve general results. Similarly, the authors in [39] had to annotate their dataset themselves to apply machine learning methods, as no ready, well-annotated, and publicly available dataset was available. By comparison, the English language dataset, which includes 20,000 annotated patient records, has been available since 2009 [45].

Natural language processing needs a high-quality annotated corpus. In [43], the authors pick up the gauntlet of creating a corpus for the Polish medical domain following the TEI P5 guidelines—as recommended for biomedical corpora. They used 460 hospital discharge reports of diabetic patients, each one 1.5–2.5 pages long. Those texts are mostly written as free speech with diagnosis in Latin. As a result, the corpus consists of 348,461 tokens and contains three annotation levels—semantic, morphological tags, and tokens. The corpus was manually corrected when publishing the mentioned article, and the authors could not find new information.

Apart from the domain approach, studies are also conducted in a general scope. In [44], the authors review named entity recognition tools for the Polish language and available corpora. The authors list four available corpora: The National Corpus of Polish (NKJP), Corpus of Wrocław University of Technology (KPWr), Corpus of Economic News (CEN), and Polish Spatial Texts (PST), where only CEN is a domain-specific one. In this article, specialized tools for NER can be found. One of them is PolDeepNer [46]—a system created with NKJP corpus that can recognize nested entities such as a person's name (forename and surname), organization name, geographical name, place name (district, settlement, region, country, and bloc), data, and time. This tool used a Bi-directional Long Short-Term Memory and Bi-directional Gated Recurrent Units (Bi-GRU) with a CRF layer. It is a part of the actively developed project CLARIN-PL as well as the KPWr n82 NER [47] model, which was built on the KPwr corpus and can recognize 82 types of entities. The results of this work may also help apply NLP methods in medicine.

4. Challenges and Barriers

The Ministry of Health in Poland presents standards as obligatory practices and procedures. Non-compliance with the law may result in sanctions. On the other hand, recommendations or guidelines are not strictly enforced. This leads to further differentiation in many fields within healthcare institutions. One of the most demanding challenges is implementing uniform standards across the entire health system. While this task seems straightforward, the underlying differences in approaches and circumstances at the local level might hinder the process. Each institution has its problems and faces different challenges. The most significant challenges and limitations related to standardization in Polish healthcare, determined based on the conducted literature review, are presented in Table 2.

Table 2. Key publications on the challenges and barriers of standardizing medical descriptions for the Polish language.

Publication	Focus	Findings	Challenges and Barriers
Stasiewska et al. (2022) [48]	Standardization in urinalysis laboratory tests	Analyzed urinalysis parameters like clarity, color, pH, etc. Identified multiple names, reference intervals, inconsistent data presentation, and manual entry errors.	Issues with multiple names/reference intervals, inconsistent presentation, and lack of method information.
Kilańska (2017) [49]	Opinions on Electronic Health Records (EHRs) in Poland	Surveyed 322 nurses about EHRs. Found uncertainty about EHRs' impact on work efficiency, patient safety, and overall benefits.	Uncertainty about EHR benefits, minimal consultative input from stakeholders.
Kautsch et al. (2016) [50]	Cooperation and communication in standardization development	Discussed the need for cooperation among governing bodies and effective communication. Identified issues with overlooked areas and process delays in Poland.	Problems with coordination among stakeholders, slow progress, and communication gaps.
Zabielska et al. (2016) [51]	Comparison of Polish and English medical descriptions	Compared Polish and English medical Case Descriptions. Found similar structures in the Introduction, Case Description, and Discussion.	Challenges in adapting English-based solutions to Polish, but noted structural similarities.
Instytut Komunikacji Specjalistycznej i Interkulturowej (2016) [52]	Analysis of English and Polish case report structures	Compared case report structures in English and Polish. Found repeatable schemas and similar grammatical structures, though English reports were more structured.	Differences in structural coherence and detail between English and Polish reports.

For example, in [48], the authors analyzed the standards of one of the most common laboratory tests—urinalysis. They generally call for standardization in the laboratory testing field to increase reliability and the correct interpretation of results. Urinalysis consists of multiple parameters—some are expressed in numerical form while others are descriptive. The authors analyzed whether or not the tests meet and fulfill formal criteria; for example, physicochemical parameters: clarity, color, specific gravity, pH, protein, glucose, ketone bodies, bilirubin, urobilinogen, nitrites, blood, and leukocytes. The analysis included the conformity of parameter names, how the results were presented, and the presence and ranges of reference intervals.

The authors found that multiple names or reference intervals were used for most parameters. A separate source of inconsistencies is manual data entry. The authors identified misspellings and grammar errors that made readability and the understanding of results more challenging. Another case was identified as the improper usage of values. For example, the descriptive method was used instead for effects with reference intervals. Another issue is that reference intervals can be presented using different rounding and number of decimal places. This also applies to the result values—some were presented using one decimal place, while the reference interval was presented with two decimal places. Intervals themselves were also presented in inconsistent ways. For example, the ref-

erence interval for erythrocytes was given using "0–3" or "<3", which means the same thing. Overall, the authors identified eleven different reference intervals for erythrocytes. Abbreviations and terms too general to describe the findings were also noted as factors that increased ambiguity. The authors also noted that information about the methods used was not included.

Frontline representatives also highlighted concerns about a lack of communication throughout the process and minimal or nonexistent consultative input. They assert that their involvement in shaping the policy was confined to minor details and did not influence the core issues they had identified. Additionally, the absence of a prototyping phase in the project became evident when practical issues surfaced following the policy's implementation.

Similar interviews were conducted in Poland. In [49], the author surveyed nurses and asked them about their opinions about EHRs and the benefits stemming from their usage. Overall, 322 respondents took part in the survey. Most responses indicated that respondents are unsure about the impact of standardization in nursing. In total, 45.9% were uncertain whether or not EHRs would make their work easier, 52.1% were unsure about patient safety, and 47.1% were uncertain whether EHRs are safer than paper documentation.

Another issue that might play a part in the success of developing and implementing standards is the cooperation of governing bodies. It is vital that all involved parties are on the same side, and the flow of information allows for effective communication. Delimiting responsibilities for each step of the process and a means of validating progress is crucial for the success of the process. In Poland, these problems appeared, causing some areas to be overlooked and slowing down the whole process [50].

An essential aspect of improving the quality of standards and earning the trust of medical practitioners is to have them involved in the process. They are the frontline of the medical field and encounter problems that others may overlook. By the same token, government bodies have to align expectations with their budget. The result of negotiations can be unsatisfactory for both sides, yet it is a step toward better standardization.

Integrating natural language processing into the medical workflow is extensive and challenging. While the availability of English language resources can be advantageous, these resources may not sufficiently cater to the specificity of specific languages. Adapting foreign solutions to other languages is not always straightforward and might require significant effort, particularly when the solutions were not designed with other languages in mind. This highlights the necessity for language-adaptive, domain-adaptive, and task-adaptive methods for clinical NLP.

A significant concern in addressing this issue is the lack of native language resources. The limited number of datasets, publications, and knowledge cannot be expanded overnight. These challenges, coupled with regulations and privacy acts that do not necessarily address the research aspects of sharing medical documentation, create a complex environment for developing NLP solutions.

The English language provides more structured and standardized medical documentation to serve as a reference point for other languages. In [51], the authors chose to compare Polish and English medical descriptions using a corpus created from medical journals (20 cases) and assessed the coherence of Polish texts. The analysis focused on three essential parts: the Introduction, Case Description, and Discussion. The Introduction provides information about the patient and ailments; the Case Description includes the history, examination, and treatment parts; and the last part presents the therapy results and potential recommendations. The authors concluded that in both English and Polish, the structures of medical Case Descriptions are similar in every analyzed part.

Similarly, the authors in [52], conducting an analysis comparing the structures of English and Polish case reports, found repeatable structure schemas in both languages. Although English reports are more structured and coherent, the rhetorical structures in both languages are comparable, utilizing the same grammatical structures.

5. Discussion

5.1. New Proposals and Possibilities

The comprehensive standardization approach necessitates establishing systems encompassing all facets of healthcare. Scientific and academic associations often play an active role in this standardization process. Emphasizing the role of medical practitioners and healthcare institutions in developing modern healthcare would make these processes more coherent and focused on improving the healthcare field. By bringing them to the forefront, experts can provide more clarity into the matter, allowing them to provide suggestions and oversee them holistically. Moreover, it would be mainly their responsibility to create and curate annotation guidelines, which would make data on a national level more consistent. It would also help transfer knowledge from other languages with more curated datasets and solutions. Adopting uniform standards would greatly enhance knowledge transfer efforts, as the data would follow the same scheme or blueprint. Effective feedback and evaluation protocols are essential to stay synchronized with the ever-changing world. These procedures provide a mechanism to correct and oversee ongoing processes while fostering trust, involvement, and cooperation.

A handful of identified challenges can be addressed by implementing standards, even more so international, well-established standards. To avoid numerous possibilities of describing a lab test result or dosage of medication, one can implement UCUM or SNOMED units of measure [53]. Some concrete proposals have already been presented; a notable example is the Polish Society of Laboratory Diagnostics, which, in 2022, introduced the proposed recommendations outlined in [54]. These recommendations center around the development of the System of Standardization of Information in Laboratory Diagnostics (SSILD), aiming to facilitate unambiguous terminology and codes based on the LOINC standardization (Figure 2). Precisely, this project aspires to achieve the following:

- The creation of a unified system of terms and codes: this involves establishing a comprehensive, unambiguous, and up-to-date system that ideally attains international acceptance.
- The extension of standardization to analyses, materials, methods, and the expression of results: ensuring uniformity across these key dimensions contributes to the effectiveness of the standardization effort.
- Certainty in reference intervals: implementing standards that provide certainty in defining reference intervals is crucial for the accurate and consistent interpretation of results.
- Access to the information needed for results interpretation: the standardization efforts aim to facilitate seamless access to the information necessary to interpret laboratory diagnostic results.

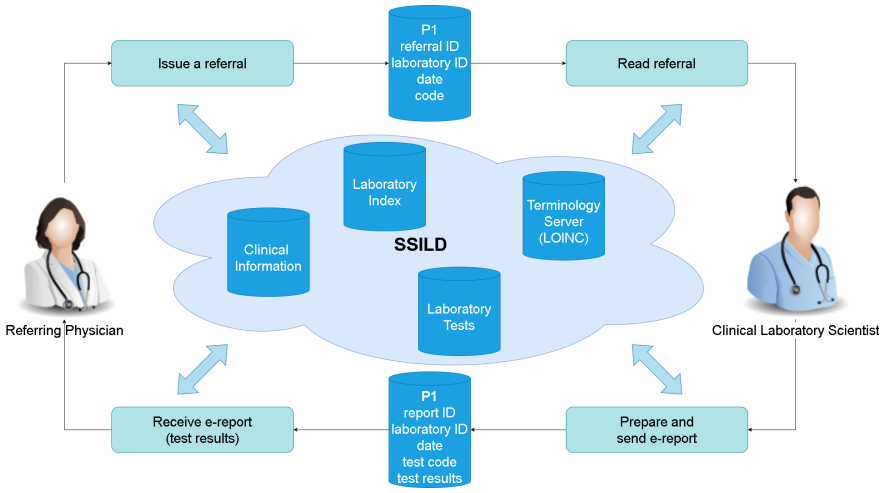

Figure 2. The conception of a new System of Standardization of Information in Laboratory Diagnostics (SSILD) [54].

An intriguing initiative called "Zdrowy pomysł" ("Healthy Idea") [55] was introduced by the Polish industry magazine "Menadżer Zdrowia" ("Health Manager"). This initiative is dedicated to showcasing innovative ideas to address and resolve healthcare-related challenges. The range of presented ideas is diverse, from niche domain standardization to developing robust solutions that traverse entire healthcare systems. The initiative has published several proposed solutions addressing security, laboratory standardization, and asthma procedures. "Healthy Idea" is a compelling indication of the increasing interest within the healthcare sector in standardized rules and integrating new technologies. Its voice resonates as a clear signal to other stakeholders, expressing a desire to actively participate in creating innovative solutions within the healthcare domain.

The authors of [49] put forth a compelling long-term strategy to boost compliance with standardized solutions. Their proposition advocates incorporating such systems into academic curricula, exposing students to these solutions, and fostering familiarity with them. This approach aims to cultivate a deeper understanding of the presented solutions and enhance the decision-making process among future professionals. Recognizing the indispensability of training for the successful implementation of EHR systems and standardization, the authors emphasize the importance of instilling this knowledge early on.

One of the main benefits of further adaptation to international standards is enhanced interoperability with European and global systems. It ensures that any compatible system can understand and process medical data. Furthermore, data analysis for a vast population would be standardized, which would help to create a bigger picture of health-related issues and contribute to global public health efforts. As a consequence, patient care can greatly benefit from such initiatives. The exchange and availability of medical records can be crucial in effective cross-border healthcare. Adhering to global standards may also, by extension, provide compliance with legal regulations, such as the GDPR (General Data Protection Regulation).

Involvement in the standardization processes, while presenting challenges, is viewed as a valuable opportunity. This active engagement imparts understanding and confidence in new projects and establishes a bridge between challenges and opportunities. Crucially, the authors stress the necessity of including representatives from all parties that would utilize such a holistic system. This inclusive approach is pivotal for ensuring compliance and effective system performance. Beyond compliance, it provides stakeholders with a platform to contribute insights and articulate their specific needs, thus enriching the development and implementation of these systems.

To forge cutting-edge solutions and sustainable systems, a robust financial foundation is imperative for their inception and continued maintenance. Publication [50] outlines several key opportunities in this regard:

- National public investment and European funds. Tapping into national public investment and European funds offers a substantial financial base for the development and longevity of these solutions.
- The integration of existing systems. The prospect of achieving integration among existing systems presents an opportunity to enhance efficiency and coherence within healthcare infrastructure.
- The development of regional and central platforms for patients. Investing in creating regional and central patient platforms stands as a pivotal opportunity, fostering improved accessibility and coordination.
- The transnational movement of EU patients. Facilitating the movement of patients across borders within the EU is a strategic opportunity that can lead to collaborative advancements in healthcare.
- The development of information and communication technologies. Embracing the continual development of Information and Communication Technologies (ICT) is highlighted as a key avenue for progress.

The support from government bodies and the EU is critical to the success of these initiatives. Notably, each opportunity, while contributing to immediate advancements,

also holds long-term benefits for all stakeholders involved. As the EU contributes to the financing of these systems, it concurrently stipulates the necessity for their implementation with a focus on interoperability with other national and European solutions.

5.2. Strategy for Coming Years in Poland

In response to the extensive challenges and the pursuit of newfound opportunities, researchers have formulated a comprehensive plan for the years ahead [56]. In "Strategia Centrum e-Zdrowia na lata 2023–2027" ("Strategy for the E-Health Centre in Poland for the period 2023–2027") outlined the primary challenges anticipated in the coming years, categorized into four dimensions:

- Financial challenges: developing and sustaining systems entail significant costs that demand careful budgetary considerations.
- Human resources: the availability of specialized personnel is crucial for effectively executing projects.
- Technological evolution: the ever-changing landscape of new technologies necessitates constant adaptation to ensure systems remain aligned with the latest advancements.
- Legal framework: the regulatory framework within which systems operate can limit the ability to implement certain solutions.

To address these challenges, an all-encompassing strategy plan was devised, focusing on five strategic areas pivotal for the development of a modern, technology-driven healthcare sector centered around the patient:

- Services. Facilitating patients to manage their health-related events in one centralized location, covering diverse aspects of healthcare.
- Stakeholders. Strengthening collaboration among all relevant parties, including patients, healthcare workers, and governing bodies, to ensure engagement and the achievement of common shared goals.
- Development. Emphasizing steady progress, enhancing security, and fostering innovation as critical components leading to success.
- Resources. Acknowledging the importance of recruiting and retaining skilled healthcare professionals possessing the necessary expertise to fulfill defined goals.
- Processes. Implementing agile methodologies and a responsive approach to system management to ensure the timely and budget-conscious delivery of projects.

Similar initiatives are underway globally, and the HACID (Hybrid Artificial Collective Intelligence for Decision Making) project stands out on the European scene. Financed primarily by the EU, the HACID aims to create a decision support system that combines collective human intelligence and AI [57]. The project addresses two main goals: overcoming knowledge fragmentation in medicine and supporting policymakers. There is a clear commitment to collaborating with the healthcare and AI sectors to implement technology-driven solutions that enhance interoperability across healthcare fields. Access to knowledge and resource sharing between researchers is a key component of this project, making its impact even more pronounced. Knowledge structurization is envisioned by creating a decision support system (HACID-DSS) that would organize domain knowledge in graphs (Domain Knowledge Graph—DKG). Subsequently, experts would engage in discussions for cases within each domain and create a Case Knowledge Graph (CDG) as a subset of the DKG. The HACID aims to aggregate solutions provided by experts, expanding them through AI suggestions.

5.3. Polish Healthcare System in International Context

Since admission to the European Union, Poland has greatly benefited from various opportunities and initiatives [58]. Along with Poland, Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Slovakia, and Slovenia were also admitted. Later on, Bulgaria and Romania joined, and then Croatia.

While each country exists in a specific context, joining the international union has allowed new members to learn from experience and with the help of long-standing members. An interesting case study in Estonia. As a former communist country, the starting point was very similar, in terms of the healthcare system, to Poland. The crucial role in the digitization of Estonia was played by its first Prime Minister—Mr Mart Laar (from 1992–1994 and 1999–2002) [59]. He envisioned a digitized, modern country and laid the foundation for various technological initiatives, including e-documentation, e-banking, and e-health. Between 2003 and 2005, the project preparation went underway, resulting in the Estonian Nationwide Health Information System (EHIS) becoming operational in 2008. One of the parts of this system is the EHR system, which is compatible with international standards, integrates a variety of services, and securely provides access to medical information.

Croatia underwent similar processes but began them later than Poland or Estonia. Ref. [60] summarizes the period from 2011 to 2021 of the "Declaration of eHealth". In this period, Croatia succeeded in health informatics and its incorporation into healthcare systems, and it created an environment for further efforts in the form of the adaptation of existing systems to meet the interoperability requirements of the European Union. At the same time, challenges persist in the form of incomplete implementations of informatics systems and overall compliance with only five statements out of nineteen. Croatia also outlined a strategy from 2021 to 2027 [61]. It outlines five goals that holistically regard healthcare and its surroundings. The goals range from developing and improving the existing systems, promoting a healthy lifestyle, and making the healthcare sector a desirable workplace.

In 2022, members of the European Union adopted the Declaration on European Digital Rights and Principles and the Digital Decade Policy Programme 2030. In May of 2024, the results of the monitoring framework for the "Digital Decade target of 100% of EU" [62], the second yearly report, were published. The report's goal was to evaluate e-health systems' performance in the EU by four indicators (and twelve sub-indicators) measuring the accessibility of health data, security of the data and access to data, service delivery, and user engagement. Overall, in the ranking, Poland was placed 5th, Croatia at 13th, and Estonia 3rd.

Even more so, the EU provides solutions and legal frameworks for creating strategies, which must be compatible with EU requirements at the top level. For example, the main problem with standards in Poland is medical entities' level of informatics systems. This hinders interoperability; what is more, medical descriptions are not evenly regulated within medical domains, leaving much freedom for medical practitioners in narrative texts. The strategy for upcoming years addresses those concerns, aiming to increase the convergence of medical systems, promote interoperability, and implement and adhere to international standards, guidelines, and regulations [56].

Moreover, the WHO, the organization responsible for ICD editions, released the "Regional digital health action plan for the WHO European Region 2023–2030" that outlines a strategy for the acceleration and convergence of healthcare digitization processes [63]. It outlines four strategic goals that highlight the most significant focus areas, from setting universal standards and promoting digitization in healthcare to involving stakeholders and strengthening international interoperability. To align with this strategy's goals, Poland has to increase the interoperability of health systems and adjust them to WHO guidelines.

Agents can benefit significantly from transferring knowledge from other members in networks of connected entities on international levels. Adapting international standards not only helped to provide established ontologies and made knowledge more accessible, but it also allowed for the more straightforward adaptation of systematic solutions across various domains. For example, the Polish system P1 is quite similar to the Estonian counterpart in structure and functionality. A combination of blueprints and ideas made it possible for Poland to catch up to established members of the EU. Poland can be studied as an example for newer members, like Croatia, and Estonia was an example and inspiration for Poland.

5.4. Large Language Models

The introduction of transformers [64] in 2017 unlocked hidden potential for AI by enabling the efficient parallelization of computation and the use of attention mechanisms, allowing for the reference of distant dependencies within data. Those aspects are crucial for performing NLP tasks, as the input sequence can be arbitrarily long, with dependencies spanning arbitrarily large distances. Scaling transformer models gave birth to Large Language Models, a solution that is capable of processing immense amounts of data and understanding the underlying meaning behind them. LLMs are capable of performing a magnitude of tasks, finding their way in a variety of different domains.

The authors in [65] deliminate medical areas that incorporated LLMs, providing examples of the solutions implemented. For medical question answering, BioMistral [66] and Med-PaLM 2 [67] are prime examples, or for radiology, Radiology-Llama2 [68]. In particular, entity recognition is available to implement such models effectively. NER and CER both experienced significant momentum in academic research since 2017–2018. According to PubMed search results, publications for CER increased from 423 in 2016 to 720 in 2023, peaking in 2022 with 786 [69]. Similarly, publications for NER increased from 255 in 2016 to 406 in 2023, peaking in 2022 with 416 [70]. Ref. [65] showcases ClinicalT5 [71] and KeBioLM [72] for the NER domain. As for CER, effort has been put toward language-specific solutions; for example, for the Arabic language, [73] adapts two existing models to suit said language, [74] evaluates models' performance for the French language, and similarly [75] benchmarks models for the English language. For the Polish language specifically, [44] provides an overview of suitable solutions. For some languages, resources are sparse and limited. To remedy this, the authors in [76] propose a Cross-Lingual Name Entity Recognition method that transfers knowledge from high-resource to low-resource languages.

5.5. Limitations and Future Research Directions

This article faced several limitations and evident risks of bias that could have impacted its findings and recommendations. One notable limitation was the reliance on a relatively narrow range of sources, especially publications available in Polish, which were not always indexed in international databases like PubMed or Scopus. This limited the scope of the review and introduced the risk of selection bias, as the studies reviewed might not have fully represented the global state of research in CER and medical standardization. The scarcity of English-language studies on the subject may have hindered cross-country comparisons and the generalizability of the findings to other healthcare systems.

The lack of widely available, well-annotated datasets for Polish-language medical texts was another significant limitation that affected the conclusions drawn about CER's applicability in Poland. This limitation hampered the development of AI models and introduced a bias, as the performance of CER systems tested on limited data may not have reflected their real-world potential in diverse clinical settings. Furthermore, this article's emphasis on the Polish context may have obscured broader international challenges, limiting its applicability to other healthcare systems that faced different regulatory or linguistic issues.

Additionally, the methodology used in this review, although suitable for exploring broad trends, lacked the rigorous assessment of quality and bias that systematic reviews typically provide. This could have led to the inclusion of studies with varying degrees of reliability, potentially affecting the accuracy of the conclusions.

The findings of this study could guide further activities in similar linguistic and healthcare contexts. However, the evident risk of bias arising from the selection of non-indexed, localized research emphasizes the need for more collaborative, international studies that could address these limitations.

Future research on CER in Polish should prioritize several key directions. The creation of comprehensive and well-annotated datasets is essential. Collaboration among healthcare professionals, linguists, and AI experts will help ensure that these datasets capture the complexities of Polish clinical texts, supporting the development of more accurate machine

learning models. Partnerships with healthcare institutions are crucial for accessing clinical records, enabling the curation of larger, more diverse datasets that reflect real-world clinical scenarios.

Another primary focus is the standardization of annotation guidelines. Consistent labeling rules will enhance the quality and interoperability of datasets, facilitating more effective machine learning applications. Moreover, open data initiatives that ensure patient privacy while sharing these annotated datasets should be encouraged to accelerate research.

International collaboration also presents a significant opportunity. Engaging with global research communities facing similar challenges in other languages can provide valuable insights and shared strategies, helping Polish CER models progress more efficiently. Lastly, educational campaigns are needed to raise awareness about the importance of annotated datasets for AI-driven medical solutions, which can help secure broader support for future research initiatives.

These conclusions are expanded upon by recommendations presented in the next section, which can be generalized to health systems in other countries.

6. Recommendations

Incorporating machine learning techniques within the medical field for specific languages is confronted with challenges primarily from the shortage of well-annotated datasets. This deficiency becomes particularly evident when examining CER within the context of the Polish language, where exploration and advancements in this area remain relatively uncharted. The scarcity of meticulously annotated datasets acts as a significant impediment, impeding the progress of developing and training machine learning models that are finely tuned for the nuanced intricacies of clinical text analysis in Polish.

The absence of adequately labeled datasets hinders the evolution of tailored machine learning models. It presents a formidable obstacle to advancing research and applications within the medical domain for the Polish language. This deficiency accentuates the pressing need for collective endeavors to curate comprehensive and accurately labeled datasets. These efforts are essential for overcoming the existing limitations and propelling the development of robust machine learning solutions specifically designed for CER in Polish.

Proposals to address the challenges posed by the scarcity of well-annotated datasets in utilizing machine learning techniques for CER in the Polish language include ten positions:

1. Collaborative dataset curation: collaboration between healthcare professionals, linguists, and machine learning experts should be facilitated to collectively curate comprehensive datasets collectively, ensuring an accurate representation of the intricacies of clinical language in Polish.
2. The engagement of healthcare institutions: Partnerships with healthcare institutions should be established to gain access to clinical records and texts, fostering the creation of large and diverse datasets. Collaboration with medical professionals will enhance the authenticity and relevance of the annotated data.
3. Community involvement and crowdsourcing: Community involvement and the utilization of crowdsourcing platforms should be encouraged to facilitate the annotation process. Engaging the broader community can expedite the dataset creation, ensuring a diverse range of clinical contexts is covered.
4. Government support and funding: Advocacy for government support and funding to promote research initiatives focused on building annotated datasets for clinical text analysis in Polish should be undertaken. Financial backing can incentivize researchers and organizations to contribute to developing these datasets.
5. Standardized annotation guidelines: Standardized annotation guidelines for CER in the Polish language should be developed. Clear and consistent guidelines will enhance the quality and interoperability of annotated datasets, facilitating the training of machine learning models.
6. Cross-disciplinary workshops and training programs: Workshops and training programs that bring together healthcare, linguistics, and AI experts to share knowledge

and foster cross-disciplinary collaboration should be organized. This can lead to a better understanding of the unique challenges posed by clinical language in Polish.
7. Open data initiatives: Adopting open data initiatives within the medical and research communities, promoting the sharing of annotated datasets, should be encouraged. Open data initiatives can accelerate the development of machine learning models by providing a shared resource for experimentation and validation. In this regard, data anonymization and ensuring patients' privacy is essential.
8. Continuous evaluation and feedback mechanisms: Mechanisms for continuous evaluation and feedback on existing datasets should be established. This iterative process ensures that datasets evolve to meet the changing needs of machine learning models, maintaining relevance and accuracy over time.
9. International collaboration: Collaboration with international research communities working on similar challenges in other languages should be fostered. Shared experiences and best practices can inform the development of strategies specific to CER in the Polish language.
10. Education and awareness campaigns: Educational campaigns should be conducted to raise awareness among healthcare professionals, researchers, and the general public about the importance of annotated datasets in advancing machine learning applications in the medical domain for the Polish language. This can help garner support and participation in dataset curation efforts.

Addressing these proposals requires a collaborative approach involving healthcare professionals, linguists, and machine learning experts to ensure the creation of datasets that accurately capture the intricacies of clinical language in Polish. This concerted effort would contribute to refining existing models and pave the way for the innovation and implementation of cutting-edge machine learning solutions, ultimately fostering advancements in medical research and applications within the Polish-speaking community. By recognizing and actively addressing the scarcity of annotated data, stakeholders can collectively contribute to developing a foundation for successfully integrating machine learning techniques in clinical text analysis for the Polish language, thus bridging the current gap and unlocking the full potential of these technologies in healthcare.

7. Conclusions

This article reviews two intricately linked areas: the standardization of medical descriptions in Poland and the utilization of machine learning techniques for CER. By fulfilling these dual objectives, this article contributed to a richer understanding of the present landscape and prospects in these critical domains, ultimately enhancing the management of healthcare data and the comprehension of clinical texts within the Polish context. Moreover, the analysis of the Polish case can serve as a valuable example for healthcare systems in other countries.

Standardization efforts contribute to the improved performance of CER systems by ensuring consistent, high-quality data, enhancing interoperability, facilitating data integration, and reducing errors and development costs. These benefits are essential for accurately and efficiently recognizing clinical entities, ultimately supporting better patient care, research, and healthcare data analysis.

Clinical entity recognition is a critical component of medical text analysis with wide-ranging applications that impact patient care, research, healthcare administration, and information retrieval. CER facilitates the conversion of unstructured text into structured data, enabling more informed decision making and advancing healthcare. CER is vital for automating and streamlining tasks such as patient record management, clinical trials, and diagnostic assistance, which all depend on the precise and rapid extraction of crucial medical information from textual data.

The research presented in this article highlights the pressing need for developing robust CER models tailored to Polish-language medical texts. Despite the widespread application of CER in English, there is a clear gap in resources for the Polish language.

In this article, we cited examples of CER in Polish, including systems for correcting doctors' notes, extracting information from medical records, and segmenting patient visits. These tools can significantly impact the healthcare system by reducing administrative burdens on healthcare professionals, improving the accuracy of medical records, and ultimately enhancing patient outcomes through better data-driven decision making.

Crucially, the practical application of this research extends beyond Poland. Many countries with underrepresented languages face similar challenges in clinical data processing, and the advances in CER for Polish could serve as a model for implementing AI-driven solutions in these regions. By refining CER models and adapting them to diverse linguistic and healthcare contexts, the research can foster improvements in patient care, medical research, and healthcare administration on a global scale.

This work underscores that the standardization of medical language, alongside the development of machine learning tools for clinical text analysis, has the potential to benefit not only Polish healthcare but also healthcare systems in other countries with similar linguistic challenges. The approaches and technologies discussed in this article could be used to improve interoperability, data integration, and the overall efficiency of clinical information systems across borders. Therefore, this research lays the groundwork for further international collaboration and innovation in medical AI applications.

Author Contributions: Conceptualization, M.T.; formal analysis, M.T.; investigation, M.P.W. and B.A.J.; methodology, M.T.; project administration, M.P.W.; supervision, M.T.; validation, M.T. and B.A.J.; visualization, M.T. and M.P.W.; writing—original draft, M.P.W. and B.A.J.; writing—review and editing, M.T. and B.A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Wróblewska, A.; Sydow, M. DEBORA: Dependency-Based Method for Extracting Entity-Relationship Triples from Open-Domain Texts in Polish. In *Proceedings of the Foundations of Intelligent Systems;* Chen, L., Felfernig, A., Liu, J., Raś, Z.W., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 155–161.
2. Pohl, A. Knowledge-based named entity recognition in Polish. In Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, Kraków, Poland, 8–11 September 2013; Ganzha, M., Maciaszek, L., Paprzycki, M., Eds.; pp. 145–151. Available online: https://annals-csis.org/proceedings/2013/pliks/300.pdf (accessed on 24 October 2024).
3. Marcinczuk, M.; Kocon, J.; Gawor, M. Recognition of named entities for Polish—Comparison of deep learning and conditional random fields approaches. In Proceedings of the PolEval 2018 Workshop, Warsaw, Poland, 19 October 2018; pp. 77–92.
4. Wróbel, K.; Smywiński-Pohl, A. KNER: Named Entity Recognition for Polish. In Proceedings of the PolEval 2018 Workshop, Warsaw, Poland, 19 October 2018; pp. 101–108.
5. Borchmann, Ł.; Gretkowski, A.; Gralinski, F. Approaching nested named entity recognition with parallel LSTM-CRFs. In Proceedings of the PolEval 2018 Workshop, Warsaw, Poland, 19 October 2018; pp. 63–73.
6. Dadas, S. Combining Neural and Knowledge-Based Approaches to Named Entity Recognition in Polish. In *Artificial Intelligence and Soft Computing*; Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J.M., Eds.; Springer: Cham, Switzerland, 2019; pp. 39–50.
7. Denisiuk, A.; Ganzha, M.; Wasielewska-Michniewska, K.; Paprzycki, M. Feature Extraction for Polish Language Named Entities Recognition in Intelligent Office Assistant. In Proceedings of the Annual Hawaii International Conference on System Sciences, HICSS, Maui, HI, USA, 3–7 January 2022; Bui, T.X., Ed.; pp. 1320–1329.
8. Sopyla, K.; Drozda, P.; Ropiak, K.; Witkowska, U.; Sieniewicz, M.; Jankowski, S. Named Entity Recognition Model for Polish Books. In *Intelligent Information and Database Systems*; Nguyen, N.T., Chbeir, R., Manolopoulos, Y., Fujita, H., Hong, T.P., Nguyen, L.M., Wojtkiewicz, K., Eds.; Springer: Singapore, 2024; pp. 147–158.
9. Lachowska, H. Transformacja Systemu Ochrony Zdrowia w Polsce a Efektywność Ekonomiczna śWiadczenia Usług Medycznych przez Szpitale [Transformation of the Healthcare System in Poland and the Economic Efficiency of Medical Services Provided by Hospitals]. Available online: https://weif.uniwersytetradom.pl/wp-content/uploads/sites/15/2021/11/WE_doktorat_LachowskaH_autoreferat.pdf (accessed on 12 October 2024).

10. Paszkowska, M. System ochrony zdrowia w Polsce—Zmiana modelu [Healthcare System in Poland—A Change of Model]. *Probl. Zarz.* **2017**, *15*, 25–41. [CrossRef]
11. Marshal of the Sejm. Ustawa z Dnia 6 lutego 1997 r. o Powszechnym Ubezpieczeniu Zdrowotnym [Act of February 6, 1997 on Universal Health Insurance]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu19970280153 (accessed on 12 October 2024).
12. Marshal of the Sejm. Ustawa z Dnia 23 Stycznia 2003 r. o Powszechnym Ubezpieczeniu w Narodowym Funduszu [Act of January 23, 2003 on Universal Insurance in the National Fund]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu20030450391 (accessed on 12 October 2024).
13. Marshal of the Sejm. Ustawa z Dnia 28 Kwietnia 2011 r. o Systemie Informacji w Ochronie Zdrowia [Act of April 28, 2011 on the Health Information System]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu20111130657 (accessed on 12 October 2024).
14. Sejm of the Republic of Poland. Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z Dnia 26 Marca 2020 r. w Sprawie Ogłoszenia Jednolitego Tekstu Ustawy o Systemie Informacji w Ochronie Zdrowia [Announcement of the Marshal of the Sejm of the Republic of Poland dated March 26 2020 Regarding the Publication of the Consolidated Text of the Act on the Health Information System]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20200000702 (accessed on 9 November 2023).
15. Ministry of Health. Ruszają Prace nad Opracowaniem Polskiej Wersji Klasyfikacji ICD-11 [The Work on Developing the Polish Version of the ICD-11 Classification is underway]. Available online: https://www.zdrowie.gov.pl/aktualnosc-3879-ruszaja_prace_nad_opracowaniem_polskiej.html (accessed on 9 November 2023).
16. Majewicz, P.I. ICD-11 implikacje w badaniach i praktyce psychologiczno-pedagogicznej [ICD-11 implications in psychological and educational research and practice]. *Lub. Rocz. Pedagog.* **2023**, *42*, 123–139. [CrossRef]
17. Komitet Monitorujący Program Fundusze Europejskie dla Rozwoju Społecznego. Załącznik nr 2 do Uchwały nr 60 Komitetu Monitorującego Program Fundusze Europejskie dla Rozwoju Społecznego 2021–2027 [Annex No. 2 to Resolution No. 60 of the Monitoring Committee of the European Funds for Social Development Program 2021–2027], 2023. Available online: https://www.rozwojspoleczny.gov.pl/media/126188/Wsparcie_wdroZenia_jedenastej_rewizji_MiEdzynarodowej_Klasyfikacji_Chorob_ICD-11.pdf (accessed on 13 October 2024).
18. The European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions European Interoperability Framework—Implementation Strategy. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:134:FIN (accessed on 12 November 2023).
19. Sejm of the Republic of Poland. Ustawa z Dnia 17 Lutego 2005 r. o Informatyzacji Działalności Podmiotów Realizujących Zadania Publiczne [The Act of February 17 2005 on the Informatization of Entities Carrying Out Public Tasks]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu20050640565 (accessed on 9 November 2023).
20. The Council of Ministers. Rozporządzenie Rady Ministrów z Dnia 12 Kwietnia 2012 r. w Sprawie Krajowych Ram Interoperacyjności, Minimalnych Wymagań dla Rejestrów Publicznych i Wymiany Informacji w Postaci Elektronicznej oraz Minimalnych Wymagań dla Systemów Teleinformatycznych [Regulation of the Council of Ministers of April 12 2012 on National Interoperability Framework, Minimum Requirements for Public Registers and Electronic Information Exchange, as Well as Minimum Requirements for Teleinformation Systems]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu20120000526 (accessed on 9 November 2023).
21. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*. World Health Organization: Geneva, Switzerland, 2004. Available online: https://iris.who.int/handle/10665/42980?search-result=true&query=icd-10&scope=&rpp=10&sort_by=score&order=desc (accessed on 24 October 2024).
22. International Health Terminology Standards Development Organisation. SNOMED International | Delivering SNOMED CT. Available online: https://www.snomed.org/value-proposition (accessed on 9 November 2023).
23. Regenstrief Institute. LOINC: Logical Observation Identifiers Names and Codes. Available online: https://loinc.org (accessed on 24 October 2024).
24. Bourquard, K.; Parisot, C. eHealth Interoperability in Poland. Report on profile recommendations for e-Referral and exchange of medical documentation (P1/Increment 2 & 3) Final Report IHE Services. Available online: https://ezdrowie.gov.pl/pobierz/ihe-profile-recommendations-report-v-1-5-english-version (accessed on 9 November 2023).
25. International Organization for Standardization. ISO 12052:2021: Health informatics—Digital imaging and communication in medicine (DICOM)—Part 1: Introduction and overview. Available online: https://www.iso.org/standard/72941.html (accessed on 24 October 2024).
26. Health Level Seven International. HL7 Standards Product Brief—CDA® Release 2 | HL7 International. Available online: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7 (accessed on 15 November 2023).
27. e-Health Center Team. Wszystkie Szablony: Plcda- Zestawienie DECOR (Elementy Danych, Kody, OIDy i Reguły) [All Templates: Plcda- DECOR Overview (Data Elements, Codes, OIDs, and Rules)]. Available online: https://www.cez.gov.pl/HL7POL-1.3.2/plcda-html-1.3.2/plcda-html/rules.html (accessed on 11 November 2023).
28. Kielar, M. Trendy Informatyzacji Szpitali w Polsce [Informatization Trends in Hospitals in Poland]. Available online: https://ucyfrowienie.pl/wp-content/uploads/2013/03/opm_artykul_2013_03_37944.pdf (accessed on 19 November 2023).

29. Spalek, D. Opis Badań Diagnostycznych w EDM—Proste to RODO Ochrona Danych w Medycynie [Description of Diagnostic Tests in EDM—Simple GDPR Data Protection in Medicine]. Available online: https://prostetorodo.pl/2020/04/16/opis-badan-diagnostycznych-w-edm/ (accessed on 20 November 2023).
30. Książkowski, M. Opis Badań Diagnostycznych w Elektronicznej Dokumentacji Medycznej—MZ Precyzuje Projekt Rozporządzenia [Description of Diagnostic Tests in Electronic Medical Documentation—Ministry of Health Specifies the Draft Regulation]. Available online: https://pulsmedycyny.pl/opis-badan-diagnostycznych-w-elektronicznej-dokumentacji-medycznej-mz-precyzuje-projekt-rozporzadzenia-971847 (accessed on 20 November 2023).
31. Sejm of the Republic of Poland. Ustawa z dnia 28 kwietnia 2011 r. o systemie informacji w ochronie zdrowia [The Act of April 28 2011 on the Health Information System]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20230001851 (accessed on 12 November 2023).
32. Minister of Health. Projekt Rozporządzenia Ministra Zdrowia Zmieniający Rozporządzenie w Sprawie RodzajóW Elektronicznej Dokumentacji Medycznej [Draft Regulation of the Minister of Health Amending the Regulation on Types of Electronic Medical Documentation]. Available online: https://legislacja.rcl.gov.pl/projekt/12321750/katalog/12591108#12591108 (accessed on 12 November 2023).
33. Minister of Health. Obwieszczenie Ministra Zdrowia w Sprawie OgłOszenia Wykazu Wzorcowych Procedur Radiologicznych z Zakresu Radiologii—Diagnostyki Obrazowej i Radiologii Zabiegowej [The Announcement of the Minister of Health Regarding the Publication of the List Of Standard Radiological Procedures in the Field of Diagnostic Imaging and Interventional Radiology]. Available online: https://dziennikmz.mz.gov.pl/DUM_MZ/2015/78/akt.pdf (accessed on 12 November 2023).
34. Minister of Health. Obwieszczenie Ministra Zdrowia z dnia 13 września 2016 r. w Sprawie Ogłoszenia Jednolitego Tekstu Rozporządzenia Ministra Zdrowia w Sprawie Standardów Jakości dla Medycznych Laboratoriów Diagnostycznych i Mikrobiologicznych [The Announcement of the Minister of Health Dated September 13 2016 Regarding the Publication of the Consolidated Text of the Regulation of the Minister of Health on Quality Standards for Medical Diagnostic and Microbiological Laboratories]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20160001665 (accessed on 12 November 2023).
35. Marshal of the Sejm. Obwieszczenie Marszałka Sejmu Rzeczypospolitej Polskiej z Dnia 4 Września 2023 r. w Sprawie Ogłoszenia Jednolitego Tekstu Ustawy o Medycynie Laboratoryjnej [The Announcement of the Marshal of the Sejm of the Republic of Poland Dated September 4 2023 Regarding the Publication of the Consolidated Text of the Law on Laboratory Medicine]. Available online: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20230002125 (accessed on 12 November 2023).
36. Legal Team of the National Chamber of Laboratory Diagnosticians, e-Health Center Team. Elektroniczna Dokumentacja Medyczna w Medycznych Laboratoriach Diagnostycznych [Electronic Medical Documentation in Medical Diagnostic Laboratories]. Available online: https://kidl.org.pl/get-file/4190_kidl-rekomendacje-edm-2021-nowela-faqfnla.pdf (accessed on 12 November 2023).
37. Mykowiecka, A.; Marciniak, M. Domain–Driven Automatic Spelling Correction for Mammography Reports. *Adv. Soft Comput.* **2007**, *35*, 521–530. [CrossRef]
38. Mykowiecka, A.; Marciniak, M.; Kupść, A. Rule-based information extraction from patients' clinical data. *J. Biomed. Inform.* **2009**, *42*, 923–936. [CrossRef] [PubMed]
39. Dobrakowski, A.G.; Mykowiecka, A.; Marciniak, M.; Jaworski, W.; Biecek, P. Interpretable segmentation of medical free-text records based on word embeddings. *J. Intell. Inf. Syst.* **2021**, *57*, 447–465. [CrossRef]
40. Anetta, K.; Arslan, M. Transferability of General Polish NER to Electronic Health Records. In Proceedings of the Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, Karlova Studánka, Czech Republic, 10–12 December 2021; pp. 151–159.
41. Anetta, K.; Horak, A.; Wojakowski, W.; Wita, K.; Jadczyk, T. Deep Learning Analysis of Polish Electronic Health Records for Diagnosis Prediction in Patients with Cardiovascular Diseases. *J. Pers. Med.* **2022**, *12*, 869. [CrossRef]
42. Obuchowski, A.; Klaudel, B.; Jasik, P. Information Extraction from Polish Radiology Reports Using Language Models. Available online: https://aclanthology.org/2023.bsnlp-1.14.pdf (accessed on 2 November 2023).
43. Marciniak, M.; Mykowiecka, A. Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. In *Proceedings of BioNLP 2011 Workshop*; Association for Computational Linguistics: Portland, OR, USA, 2011; pp. 92–100.
44. Marcińczuk, M.; Wawer, A. Named entity recognition for Polish. *Pozn. Stud. Contemp. Linguist.* **2019**, *55*, 239–269. [CrossRef]
45. Roberts, A.; Gaizauskas, R.; Hepple, M.; Demetriou, G.; Guo, Y.; Roberts, I.; Setzer, A. Building a semantically annotated corpus of clinical texts. *J. Biomed. Inform.* **2009**, *42*, 950–966. [CrossRef]
46. Common Language Resources & Technology Infrastructure. PolDeepNer. Available online: https://github.com/CLARIN-PL/PolDeepNer (accessed on 16 November 2023).
47. Marcińczuk, M. KPWr n82 NER Model (on Polish RoBERTa Base). Available online: https://clarin-pl.eu/dspace/handle/11321/743 (accessed on 16 November 2023).
48. Stasiewska, A.; Kamińska, J.; Bil-Lula, I.; Jankowski, M.; Ćwiklińska, A. Analysis of urinalysis laboratory report forms in medicallaboratories. *Diagn. Lab.* **2022**, *58*, 74–85. [CrossRef]
49. Kilańska, D.M. Elektroniczny rekord pacjenta w opinii pielęgniarek. Implikacje do dydaktyki—Wykorzystanie narzędzi IT w nauczaniu klasyfikacji ICNP® [Electronic Patient Record in the Opinion of Nurses. Implications for Didactics—Utilizing IT Tools in Teaching the ICNP® Classification]. *Probl. Pielęgniarstwa* **2017**, *25*, 69–76. [CrossRef]
50. Kautsch, M.; Lichoń, M.; Matuszak, N. Development of Publicly Funded eHealth in Poland: Barriers and Opportunities. *Econ. Sociol.* **2016**, *9*, 28–40. [CrossRef]

51. Zabielska, M.; Żelazowska, M., Spójność tekstu specjalistycznego a obraz pacjenta w medycznym opisie przypadku (kontynuacja badania) [The coherence of specialized text and the patient's image in the medical case description (continuation of the study)]. In *Spójność Tekstu Specjalistycznego: 2*; Wydawnictwo Naukowe Instytutu Komunikacji Specjalistycznej i Interkulturowej Uniwersytet Warszawski: Warsaw, Poland, 2016; Volume 34, pp. 112–124.
52. Żelazowska, M.; Zabielska, M. Opis przypadku jako tekst specjalistyczny w dyskursie medycznym—Przegląd badań [The case report as a specialised text in medical discourse —Review of the literature]. *Lingwist. Stosow. Appl. Linguist. Angew. Linguist.* **2016**, *16*, 165–190.
53. Hansen, A.S.; Melgara, M.; Abad, I.; Thorp, J. The Experience of Selecting the Code Systems for the Development of the epSOS Master Value Catalogue (MVC). European Commission, DG SANTE, 2013. Available online: https://health.ec.europa.eu/document/download/42149fff-d855-4f92-9d62-471a18726876_en (accessed on 10 October 2023).
54. Kozierkiewicz, A.; Goralewska, A.; Jankowska-Kulawy, A.; Judycki, T.; Klimaszewska-Łata, J.; Kołtan, S.; Kopeć, B.; Koper-Lenkiewicz, O.M.; Łangowski, K.; Osuch, A.; et al. Standardization of information in laboratory diagnostics. Proposal of the Polish Society of Laboratory Diagnostics. *Diagn. Lab.* **2022**, *58*, 21–30. [CrossRef]
55. Lurka, K. Zdrowy Pomysł, Menedżer Zdrowia [Healthy Idea, Health Manager], 2022. Available online: https://www.termedia.pl/mz/Masz-Zdrowy-pomysl-zglos-go,46359.html (accessed on 21 December 2023).
56. e-Health Center Team. Strategia Centrum e-Zdrowia na Lata 2023–2027 [Strategy of the e-Health Center for the years 2023-2027]. Available online: https://cez.gov.pl/pl/page/o-nas/aktualnosci/strategia-centrum-e-zdrowia-na-lata-2023-2027 (accessed on 21 December 2023).
57. Consiglio Nazionale Delle Ricerche. Hybrid Human Artificial Collective Intelligence in Open-Ended Decision Making | HACID Project | Fact Sheet | HORIZON. Available online: https://cordis.europa.eu/project/id/101070588 (accessed on 4 December 2023).
58. (PISM), P.I.S.M. 20 lat Polski w Unii Europejskiej [20 Years of Poland in the European Union], 2024. Available online: https://www.pism.pl/webroot/upload/files/Raport/Raport%20PISM%2020%20lat%20Polski%20w%20Unii%20Europejskiej.pdf (accessed on 13 October 2024).
59. Janek, M.; Peeter, R.; Dirk, D.; Gunnar, P. Ten Years of the e-Health System in Estonia. In Proceedings of the 3rd International Workshop on (Meta)Modelling for Healthcare Systems, MMHS 2018, Bergen, Norway, 13 June 2018. Available online: https://ceur-ws.org/Vol-2336/MMHS2018_invited.pdf (accessed on 13 October 2024).
60. Josipa, K.; Hrvoje, B.; Silvije, B.B.M.; Silvije, V. Declaration on eHealth—10 years later. *ResearchGate* **2022**, *28*, 14–33.
61. Ministry of Health, Republic of Croatia. National Health Development Plan 2021–2027 [Croatian], 2022. Available online: https://zdravlje.gov.hr/UserDocsImages/2022%20Objave/Nacionalni%20plan%20razvoja%20zdravstva%202021.-2027..pdf (accessed on 13 October 2024).
62. European Commission. Digital Decade 2024: EHealth Indicator Study Report, 2024. Available online: https://digital-strategy.ec.europa.eu/en/library/digital-decade-2024-ehealth-indicator-study (accessed on 13 October 2024).
63. World Health Organization. Regional Digital Health Action Plan for the WHO European Region 2023–2030. 2022. Available online: https://iris.who.int/bitstream/handle/10665/360950/72wd05e-DigitalHealth-220529.pdf (accessed on 13 October 2024).
64. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In Proceedings of the 2017 Conference on Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Volume 30.
65. Nazi, Z.A.; Peng, W. Large Language Models in Healthcare and Medical Domain: A Review. *Informatics* **2024**, *11*, 57. [CrossRef]
66. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.A.; Rouvier, M.; Dufour, R. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv* **2024**, arXiv:2402.10373.
67. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. Towards expert-level medical question answering with large language models. *arXiv* **2023**, arXiv:2305.09617.
68. Liu, Z.; Li, Y.; Shu, P.; Zhong, A.; Yang, L.; Ju, C.; Wu, Z.; Ma, C.; Luo, J.; Chen, C.; et al. Radiology-llama2: Best-in-class large language model for radiology. *arXiv* **2023**, arXiv:2309.06419.
69. PubMed Search for "(CER) OR (Clinical Entity Recognition)" (2016–2024). Search Yielded Number of Articles Related to Clinical Entity Recognition by Year. Online Resource. Available online: https://pubmed.ncbi.nlm.nih.gov/?term=(CER)%20OR%20(Clinical%20Entity%20recognition)&filter=years.2016-2024&timeline=expanded (accessed on 26 September 2024).
70. PubMed Search for "(NER) OR (Named Entity Recognition) OR (Name Entity Recognition)" (2016–2024). Search Yielded Number of Articles Related to Named Entity Recognition by Year. Online Resource. Available online: https://pubmed.ncbi.nlm.nih.gov/?term=(NER)%20OR%20(Named%20Entity%20recognition)%20OR%20(Name%20Entity%20recognition)&timeline=expanded (accessed on 26 September 2024).
71. Lu, Q.; Dou, D.; Nguyen, T. ClinicalT5: A generative language model for clinical text. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 6–10 December 2022; pp. 5436–5443.
72. Yuan, Z.; Liu, Y.; Tan, C.; Huang, S.; Huang, F. Improving biomedical pretrained language models with knowledge. *arXiv* **2021**, arXiv:2104.10344.
73. Kocaman, V.; Mellah, Y.; Haq, H.; Talby, D. Automated De-Identification of Arabic Medical Records. In *Proceedings of ArabicNLP 2023*; Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., et al., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 33–40. [CrossRef]

74. Bannour, N.; Servan, C.; Névéol, A.; Tannier, X. A Benchmark Evaluation of Clinical Named Entity Recognition in French. *arXiv* **2024**, arXiv:2403.19726.
75. Li, J.; Wei, Q.; Ghiasvand, O.; Chen, M.; Lobanov, V.; Weng, C.; Xu, H. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 235. [CrossRef]
76. Shi, K.; Chen, G.; Gu, J.; Qian, L.; Zhou, G. Cross-Lingual Name Entity Recognition from Clinical Text Using Mixed Language Query. In *Health Information Processing: CHIP 2023: Communications in Computer and Information Science*; Xu, H., Chen, Q., Lin, H., Wu, F., Liu, L., Tang, B., Hao, T., Huang, Z., Eds.; Springer: Singapore, 2024; pp. 3–21.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.