informatics

MDPI

Case Report

Can ChatGPT Support Clinical Coding Using the ICD-10-CM/PCS?

Bernardo Nascimento Teixeira [1,2,*], Ana Leitão [3,4], Generosa Nascimento [1,5], Adalberto Campos-Fernandes [6] and Francisco Cercas [1,2]

1 Iscte-Instituto Universitário de Lisboa (ISCTE-IUL), Av. das Forças Armadas, 1649-026 Lisboa, Portugal
2 Instituto de Telecomunicações (IT), Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
3 Unidade Local de Saúde de Lisboa Ocidental, E.P.E (ULSLO), Estrada do Forte do Alto do Duque, 1449-005 Lisboa, Portugal; aleitao@ulslo.min-saude.pt
4 NOVA Medical School, Universidade Nova de Lisboa (NMS-UNL), Campo dos Mártires da Pátria 130, 1169-056 Lisboa, Portugal
5 Business Research Unit (BRU-Iscte), Av. das Forças Armadas, 1649-026 Lisboa, Portugal
6 NOVA National School of Public Health, Universidade Nova de Lisboa (ENSP-UNL), Av. Padre Cruz, 1600-407 Lisboa, Portugal
* Correspondence: bdnta@iscte-iul.pt

Abstract: Introduction: With the growing development and adoption of artificial intelligence in healthcare and across other sectors of society, various user-friendly and engaging tools to support research have emerged, such as chatbots, notably ChatGPT. Objective: To investigate the performance of ChatGPT as an assistant to medical coders using the ICD-10-CM/PCS. Methodology: We conducted a prospective exploratory study between 2023 and 2024 over 6 months. A total of 150 clinical cases coded using the ICD-10-CM/PCS, extracted from technical coding books, were systematically randomized. All cases were translated into Portuguese (the native language of the authors) and English (the native language of the ICD-10-CM/PCS). These clinical cases varied in complexity levels regarding the quantity of diagnoses and procedures, as well as the nature of the clinical information. Each case was input into the 2023 ChatGPT free version. The coding obtained from ChatGPT was analyzed by a senior medical auditor/coder and compared with the expected results. Results: Regarding the correct codes, ChatGPT's performance was higher by approximately 29 percentage points between diagnoses and procedures, with greater proficiency in diagnostic codes. The accuracy rate for codes was similar across languages, with rates of 31.0% and 31.9%. The error rate in procedure codes was substantially higher than that in diagnostic codes by almost four times. For missing information, a higher incidence was observed in diagnoses compared to procedures of slightly more than double the comparative rates. Additionally, there was a statistically significant excess of codes not related to clinical information, which was higher in procedures and nearly the same value in both languages under study. Conclusion: Given the ease of access to these tools, this investigation serves as an awareness factor, demonstrating that ChatGPT can assist the medical coder in directed research. However, it does not replace their technical validation in this process. Therefore, further developments of this tool are necessary to increase the quality and reliability of the results.

Keywords: ChatGPT; artificial intelligence; ICD-10-CM/PCS; clinical coding

Citation: Nascimento Teixeira, B.; Leitão, A.; Nascimento, G.; Campos-Fernandes, A.; Cercas, F. Can ChatGPT Support Clinical Coding Using the ICD-10-CM/PCS? *Informatics* **2024**, *11*, 84. https://doi.org/10.3390/informatics11040084

Academic Editor: Jiang Bian

Received: 12 August 2024
Revised: 26 October 2024
Accepted: 31 October 2024
Published: 7 November 2024

1. Introduction

1.1. Clinical Coding

Clinical coding is the process of converting and systematizing information found in the medical discharge report related to diagnoses and procedures. Its purpose is to

characterize hospital morbidity, providing the foundational data structure for hospital financing models and epidemiological studies [1].

The International Classification of Diseases 10th revision—Clinical Modification/Procedure Coding System (ICD-10-CM/PCS), developed by the World Health Organization, has been used in Portugal since January 2017 for the clinical coding of hospital acute care episodes (inpatient and outpatient) and categorized based on All-Patient Diagnosis Related Groups (AP-DRG-31).

The ICD-10-CM provides alphanumeric codes (ranging from three to seven characters) for diagnoses and classifying diseases, signs, symptoms, and external causes of injuries or diseases. These are organized into 21 and 22 chapters in the 2019 and 2023 versions, respectively, structured into categories, subcategories, and codes. The ICD-10-PCS allows the classification of procedures performed on different parts of the body with alphanumeric codes (seven characters), organized in tables from which the code is constructed. Each code is unique and specific to a particular technique and device. The ICD-10-CM/PCS contains over 69,000 diagnostic codes and over 78,000 procedure codes, with official conventions and guidelines regulating clinical coding, updated one to two times per year with the introduction of new codes and the revision or discontinuation of existing codes.

In Portugal, clinical coding is exclusively carried out by physicians with specific certification. The Central Administration of the Portuguese Health System ensures the standardization and regulation of clinical coding rules for acute inpatients settings according to ICD-10-CM/PCS conventions and guidelines. It is also responsible for ensuring that medical coders are up to date.

The digitization of clinical records, combined with AI, personalization, and automation, improves patient care by optimizing data management and enabling better decision-making. It enhances the monitoring of key indicators such as quality, cost, and efficiency, fostering a sustainable healthcare system. Electronic medical records are valuable tools that support personalized care and streamline organizational processes, ultimately improving healthcare outcomes [2–5].

Since 2017, clinical coding in Portugal has been performed through the Hospital Morbidity Information System (SIMH) by consulting and collecting clinical information from the Electronic Medical Records (EMRs) database. Both systems are centrally provided by the Ministry of Health. The SIMH allows online coding/auditing and is supplied with all ICD-10-CM/PCS coding documentation, conventions, guidelines, and interactive alerts, always updated with last ICD versions. According to the Portuguese Health Statistics report of 2021, there were 110 public hospitals (total: 240 hospitals) and an estimated resident population of 10,343,066 inhabitants. In that year, there were approximately 1.1 million hospitalizations (70% in public hospitals), corresponding to 687,941 coded inpatient settings episodes [6,7].

Accurate ICD-10-CM/PCS coding is crucial for avoiding errors in morbidity and mortality statistics, reimbursements processes, and automated decision support. It is complex and challenging work. Therefore, it is important to provide this activity with innovative tools that increase the effectiveness and efficiency of the process, as well as the reliability of the results.

1.2. Clinical Coding and Artificial Intelligence

Artificial intelligence (AI) and natural language processing are highly competitive in various fields, including healthcare. Concerning clinical coding, the reliance on an internationally accepted standard and the digitized format of the foundational data for diagnoses and procedures provide the minimum requirements for developing AI-driven solutions aimed at improving this task. Several technology companies have already created semi-automated clinical coding systems, including Deloitte, Optum, and Capita [8,9]

Dong et al. (2022) lists the main challenges to automated coding development: unstructured documentation, long and incomplete texts, synonyms and regionalisms used

in clinical reports, and new versions of the ICD-10-CM/PCS (a complex and dynamic code system) [8].

Every day witnesses the introduction of new and more robust technologies, including developments using AI. The increasing popularity of chatbots, computational programs simulating human interactions using Machine Learning, and natural language processing, allowing us to understand and generate responses, is notable. ChatGPT, developed by OpenAI, serves as the motivation for this research. This AI-based development demonstrates powerful functions in task understanding, instruction translation, and multilingual automatic translation, among others. ChatGPT operates on a generative pre-trained transformer (GPT) neural network AI architecture [10]. An initial free version was released in 2022, surpassing 100 million monthly active users by January 2023. ChatGPT integrates multiple technologies, including deep learning, unsupervised learning, instruction fine-tuning, multitask learning, and context and reinforcement learning. It is based on a generative pre-trained transformer (GPT) model that has been iteratively updated in several versions: GPT-1, GPT-2, GPT-3, and GPT-4 [11]. Kosinski (2023) analyzed the responses of ChatGPT-4, concluding that its performance is equivalent to that of 7-year-old children, suggesting improvements in language skills of models that were previously thought to be exclusively human [12]. Vaishya et al. (2023) highlight that the content of the response, human-like interactivity, and proper justification enhance the attractiveness and trust in using this tool [13].

In theory, given the growing amount of digitally collected medical data, natural language processing tools could assist doctors in decision-making, substantially improving the quality and efficiency of healthcare [14]. Johnson et al. (2023) tested the accuracy of clinical information returned by ChatGPT, finding that it is not entirely reliable, and accuracy decreases with complexity [14]. Gilson et al. (2023) evaluated ChatGPT in the medical domain, concluding that its average performance is equivalent to a third-year medical student [15]. Vaishya et al. (2023) enumerate advantages of this innovative technology in health, such as in literature research, title and keyword generation, rewriting texts for clarity, generating feedback, searching databases like PubMed, Scopus, and Google, and providing useful references across multiple research contexts [13]. In healthcare specifically, it stands out in summarizing patient clinical data, quickly retrieving clinical information in any healthcare context and any public health institution in the country, monitoring disease progression, and identifying diagnoses, allergies, treatments, complementary test results, among other details, from descriptive clinical notes. On the other hand, the authors express concerns about ChatGPT, such as the lack of specialized medical knowledge, potential bias, limitations in understanding complex clinical concepts, irregularities in content that may produce biased or harmful content, outdated medical knowledge and literature, the inability to provide medical advice and evaluate medical images, among others.

Given the prominence and accessibility of a free version of ChatGPT with reasonable performance, allowing users to gain confidence in the responses, the research question that prompted this work arises: Can ChatGPT support clinical coders in their clinical practice?

2. Objectives

It is estimated that the use of AI in the healthcare sector can alleviate approximately 40% of the working hours dedicated to healthcare, thereby reducing the pressure observed in the healthcare system in different countries. This is a result of the shortage of healthcare professionals, high rates of burnout among these workers, and the high demand for healthcare by the elderly [16,17].

Taking this assumption into account, the main objective of this study was to validate the applicability of AI through ChatGPT, as an auxiliary tool for medical coders to improve the performance of the clinical coding process. As secondary objectives, we aimed

to compare the accuracy of ChatGPT in coding clinical cases presented in different languages (Portuguese and English).

In this sense, we intended to assess whether ChatGPT can be a tool that provides quality, safety, and accuracy as support for clinical coding using the ICD-10-CM/PCS. The 2023 ChatGPT free version, based on the GPT-4 architecture developed by OpenAI, was utilized.

3. Methodology

We conducted a prospective exploratory study between 2023 and 2024 over 6 months, including the selection of texts for evaluation based on the literature, the stratification of structures and evaluation models for the responses, and the production of the analysis.

A total of 150 practical examples of clinical coding using the ICD-10-CM/PCS were randomly systematized. These examples were extracted from technical books [18,19] and translated into Portuguese and English. They encompass information of varying complexity levels in terms of quantity and nature, involving the number and combination of described diagnoses and procedures. In the context of this study, these examples are considered an approximation to the concept of episodes and are henceforth referred to as such.

The episodes were organized in a case notebook, and a certified medical coder and auditor reviewed the clinical coding. The ICD version from 2019, corresponding to the year of ChatGPT's update at the time of this study, was considered. In this validation phase, information accuracy, text clarity, and response quality were verified.

The episodes are in free-text format, including letters, numbers, and special characters, resembling the natural language used in this type of document.

Prompts are essential in influencing the outcomes of ChatGPT. In this study, syntactically well-constructed and unambiguous statements were used, mitigating the introduction of entropy in information processing, thus excluding this variable from this study's scope. The entirety of the information contained within the prompts was treated as convertible input, producing valuable output in a unary relation to the instruction, with due consideration to the subject under analysis. Clear and specific prompts facilitate the generation of responses by AI models that are more accurate, focused, and aligned with the intended objectives. Furthermore, effective prompt engineering ensures that generative AI models provide the most relevant, precise, and actionable insights, thereby enhancing support for clinical decision-making processes [20]. The clear instructions associated with the statement were as follows: 'The expected output is the coding in ICD-10-CM/PCS in code format.'.

The differentiated evaluation in both languages is justified as it aims to assess potential performance differences in Portuguese and in the native language of the ICD-10 (English).

The information contained in each medical discharge report is structured into three groups: Group 1 and Group 2, each with exclusive diagnoses or procedures, respectively; and Group 3, with both diagnoses and procedures.

The previous subdivision allows for a distinction to be made in the analysis, introducing the concept of complexity based on the processing object. Complexity was defined by the number of codes present in the expected clinical coding of the episode. The sample division structure into groups enables an analysis of complexity based on the specificity of the information in episodes containing only diagnoses, procedures, or both.

Table 1 summarizes the episode count based on their structure of diagnosis and procedure numbers. Of the 150 episodes, 95 (63.3%) belong to Group 1 (exclusively 163 diagnosis codes); Group 2 has 29 (19.3%) episodes (exclusively 31 procedure codes); and, finally, Group 3 comprises 26 (17.3%) episodes (distribution among 68 diagnosis codes and 55 procedure codes).

Table 1. Count of episodes distributed by diagnosis and procedure composition.

		Number of Procedures				
		0	**1**	**2**	**≥3**	**Count**
Number of Diagnoses	**0**	-	27	2	0	29
	1	51	3	2	1	57
	2	31	6	3	3	43
	≥3	13	2	3	3	21
	Count	95	38	10	7	150

Considering the representativeness of diagnosis and procedure codes, only codes from one chapter of the ICD-10-CM are not present (XIII Diseases of the musculoskeletal system and connective tissue), and in ICD-10-PCS the sample focused especially on the Medical and Surgical Section, including codes from 19 of the 31 Body Systems in this section. The different chapters present on the ICD-10-CM are identified in Supplementary S1. This research work is comprehensive concerning the representativeness relative to the ICD-10-CM/PCS. While it is not exhaustive, it is significant for the conclusions drawn.

For analyzing ChatGPT's performance in assigning ICD10-CM/PCS codes to each episode, the sample classification criteria were assessed in two scopes: individualized and independent evaluation with a focus on the code, and global evaluation with a focus on the episode.

Individualized and independent evaluation with a focus on the code includes the assessment of the code's identity, focusing exclusively on the returned code. The performance of ChatGPT was measured by the following indicators: correct codes, the count of completely correct codes; partially correct codes, the count of codes with at least three correct initial characters (applicable to codes longer than three characters); incorrect codes, all others that do not meet the above requirements; missing codes, the count of missing condition not coding mentioned in the episode; and excess codes: count of returned codes not mentioned in the episode.

A global evaluation with a focus on the episode involves the overall integrity of the episode, evaluating the specificity of the code in characterizing the disease/procedure that best represents the clinical description. Considering ChatGPT's accuracy rate, which is the ratio of the number of correctly returned codes by the number of expected codes, four performance levels were considered.

The segmentation was based on quartiles (see Table 2). Another criterion could have been used; however, a division into 25% slots was considered to structure reasoning in quantitative levels that can be easily associated with qualitative levels. The evaluation was structured in two levels:

- First Level—Code Evaluation: Each code is evaluated individually, with correct codes (1 point), partially correct codes (0.5 points), and other evaluations (0 points).
- Second Level—Episode Evaluation: The episode evaluation is calculated by taking the average score from the first level, considering the total number of codes and their individual evaluations.

This approach ensures that the episode evaluation is preceded by an assessment of code accuracy, maintaining alignment with this study's core objective.

Table 2. ChatGPT accuracy rate considering the returned codes.

Error	Classification (Episode)
0–25%	Good
26–50%	Satisfactory
51–75%	Weak
76–100%	Inadequate

4. Data Analysis

The coding responses from ChatGPT for the presented clinical cases were collected in an Excel file and organized into diagnostic and procedural codes. The codes returned were compared with the coding performed by the medical coder and auditor.

After this, the results were reorganized according to the quality of the code returned and the episode coded by ChatGPT (Figure 1).

The data were analyzed using RStudio. In summarizing the data obtained from the study, descriptive statistics pertaining to central moments were considered for continuous variables, while frequencies (percentages) were used for categorical variables. The chi-square test was employed to compare independent proportions between two groups and a *p*-value of less than 0.05 was considered statistically significant.

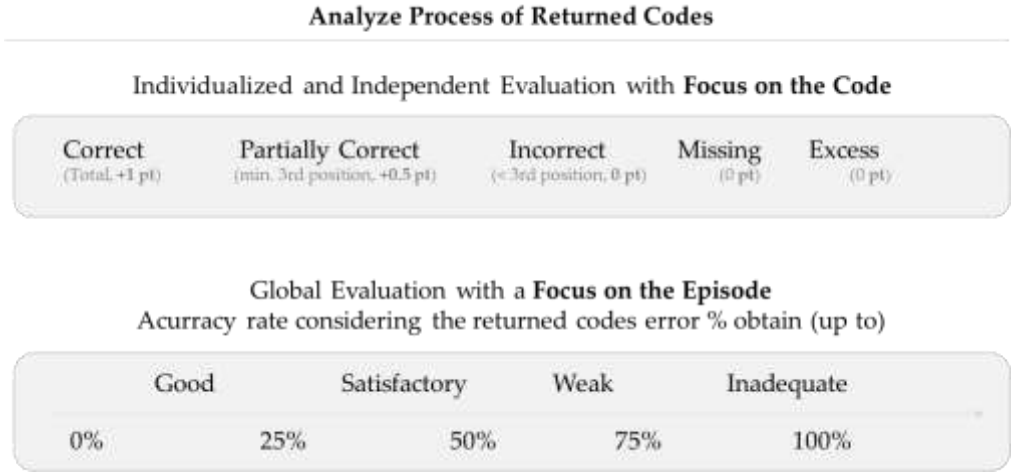

Figure 1. Summary of the coding performance analysis criteria used to evaluate ChatGPT's performance.

5. Results

The following results reflect ChatGPT's performance in coding episodes using the ICD-10-CM/PCS, based on the GPT-4 architecture developed by OpenAI, providing a broad knowledge base up to September 2021.

In accordance with the methodology described, a statistical analysis was carried out on the characteristics of the returned code and the quality of the coded episode for each of the three study groups. The alignment of the text considered in the prompt with the discharge summary does not diminish the legitimacy of this research. It does not limit itself to coding but rather uses text with codifiable and unambiguously identifiable characteristics for the purposes of quantitative evaluation.

5.1. Individualized and Independent Assessment with a Focus on Code

Based on an individualized and independent evaluation focusing on the code according to the stratified model of correct, partially correct, missing, and exceeded codes, the results show the differences in ChatGPT's coding of diagnoses and procedures for each episode.

This analysis enables us to assess, as a first approximation, the overall performance of ChatGPT in supporting the medical coder in their clinical coding practice. The results are presented in Figure 2.

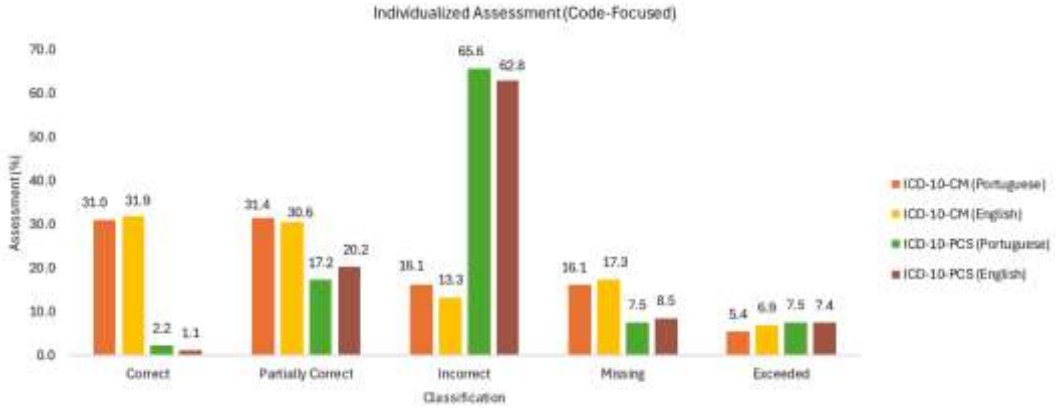

Figure 2. Individualized assessment (code-focused).

Regarding the correct codes, ChatGPT's performance was higher by approximately 29 percentage points for diagnoses compared to procedures, demonstrating greater proficiency in diagnostic coding. The accuracy rates of 31.0% and 31.9% indicate similar performance across languages, with slightly better results in the native English language of ICD-10. As for procedures, no useful conclusions can be drawn due to the poor performance in returning correct procedure codes.

In the return of partially correct codes (ICD-10-PCS), the performance for procedures considerably increased compared to the previous classifier, accounting for 17.2% and 20.2%, with better results in the native ICD-10 language.

The error rate in procedure codes is substantially higher than that in diagnostic codes, at 65.6% vs. 16.1%, and 62.8% vs. 13.3%, almost four times higher. Incorrect diagnostic codes are higher in Portuguese than in English, differing by almost three percentage points, consistent with the previous analysis. As for procedures, the error rate is significant, exceeding 60% in both cases, 65.6% and 62.8% for Portuguese and English, respectively, knowing that it is higher in Portuguese.

From the missing information, a higher incidence is observed in diagnoses compared to procedures of slightly more than double the comparative rates. Note the rate of missing diagnostic codes being higher in English (17.3%) than in Portuguese (16.1%).

Regarding procedures, given the number of incorrect codes, the result is statistically less significant.

The occurrence of excess codes, which are presented without clinical information, is higher in procedures and almost similar in both languages under study. In the codification of diagnoses, cases in English have a higher incidence of excess codes than in Portuguese, differing by 1.5 percentage points.

Next, the results are analyzed by thematic groups, based on the information contained in each episode: Group 1 (95 episodes), exclusively diagnoses; Group 2 (29 episodes), exclusively procedures, and Group 3 (26 episodes), diagnoses and procedures. In this context, only the returned codes are considered.

5.1.1. Group 1: Exclusive Diagnostics Evaluation

Figure 3 presents the results of the sample from Group 1 (163 codes). Note that the sub-samples based on complexity have different dimensions as indicated in Table 1.

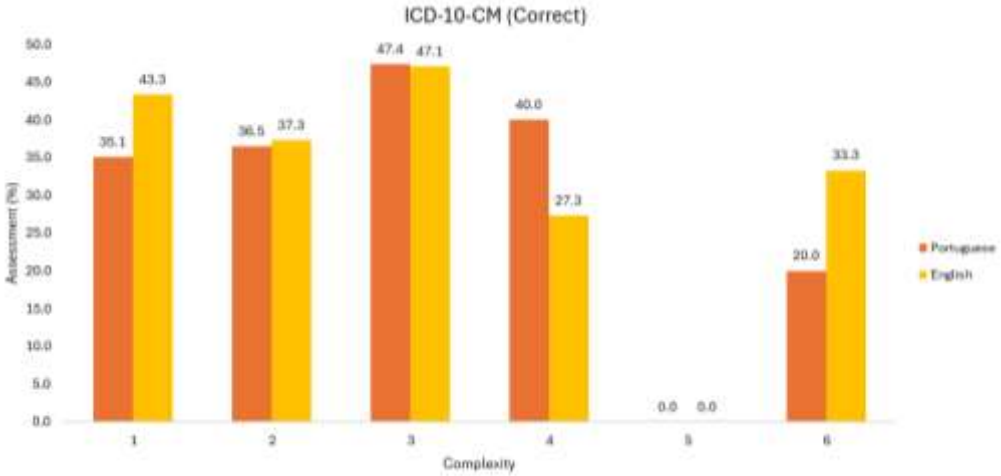

Figure 3. Evaluation of correct ICD-10-CM codes (Group 1).

Focusing on the Portuguese language, there is an increase in correct codes from 35.1% to 47.4% among the first three levels of complexity. However, the sample size in these first three levels decreases with 51, 31, and 6 observations, making it impractical to compare these values.

Nonetheless, in first three levels there is a tendency for a higher accuracy rate in English, without an understanding of measuring behavior at higher levels.

Partially correct ICD-10-CM code results are shown in Figure 4. Concerning the English language, it is observed that the representativeness grows with the increase of episode complexity; nevertheless, this finding is not present at the last level.

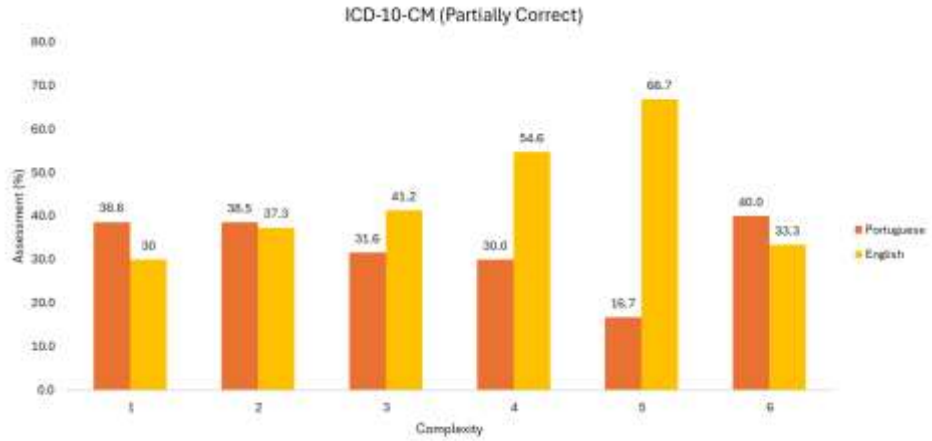

Figure 4. Evaluation of partially correct ICD-10-CM codes (Group 1).

Regarding the Portuguese language, the behavior is the opposite. However, it would not be fair to generalize this conclusion, since the sample size decreases substantially. Regarding the incorrect codes returned (Figure 5), it would be necessary to analyze the behavior of ChatGPT, whether there was a disregard, invention, or interpretative degradation of the information from the discharge report.

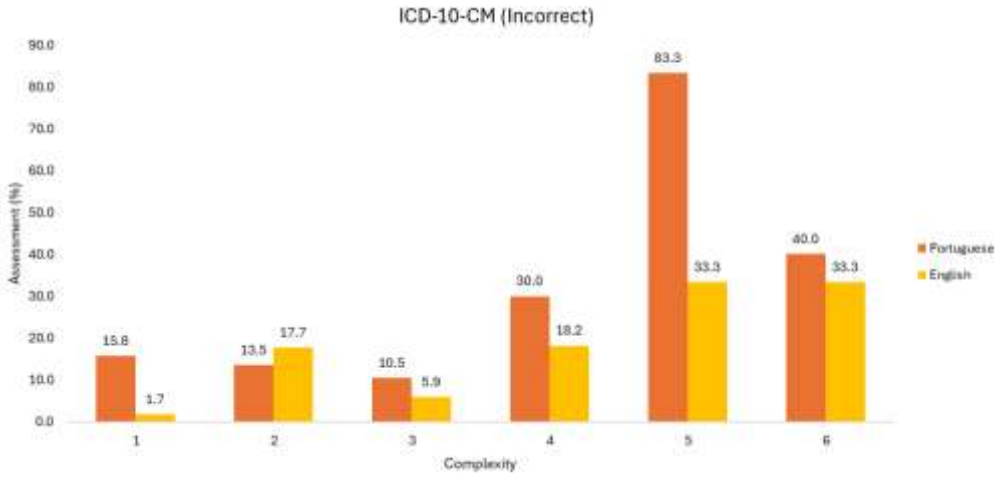

Figure 5. Evaluation of incorrect ICD-10-CM codes (Group 1).

We found incorrect codes, like inadequate structure of the code, incorrect identification of diagnoses and/or procedures, but also excess and missing codes. The reason for this behavior is not clear. The incorrect codes are more pronounced in the Portuguese language, except at level two.

The excess codes are seen only in low-complexity medical discharge reports, corresponding to episodes with one, two, or three diagnoses. Conversely, reports with four, five, and six diagnoses show no instances of excess codes. In scenarios with one diagnosis, more excess diagnoses are returned in English, differing by 5% between the studied languages. At the next level, the difference is negligible, while at level three, they differ again by 5%, with an inverted pattern.

It is imperative to understand if the codes returned by ChatGPT were based in a correct interpretation of the medical discharge report, namely about the diagnosis.

For this purpose, the following matrix (Table 3) is based on the difference between the expected and obtained code counts, emphasizing the variation in scenarios: codes missing (expected > obtained) or excess codes (expected < obtained). A comprehensive view of this matrix enables the identification of relevant scenarios regarding the predictive capability of clinical information present in the sections of the discharge notes.

Table 3. Matrix of performance in identifying information returned by ChatGPT.

		Difference in English							
		−2	**−1**	**0**	**1**	**2**	**3**	**4**	**Count**
Difference in Portuguese	**−2**	1	-	1	-	-	-	-	2
	−1	-	3	6	1	-	-	-	10
	0	-	4	51	5	1	-	-	61
	1	-	3	5	11	-	-	-	19
	2	-	-	-	-	-	-	-	0
	3	-	-	1	-	-	1	-	2
	4	-	-	-	-	-	-	1	1
	Count	1	10	64	17	1	1	1	95

In this context, the matrix can be interpreted in five regions: four quadrants and the shaded main diagonal. The shaded main diagonal indicates predictive similarity, meaning when the predictive behavior is identical, regardless of the language, represented by 68 (71.6%) of the discharge notes.

On the other hand, it is possible to distinguish in four quadrants separated by the vertical and horizontal lines of zeros. Thus, the upper-left quadrant indicates false

negatives, corresponding to 66 (69.5%) of the cases, with a notable higher tendency in Portuguese, about 63.6% of those showing a difference between languages. In the lower-right quadrant, missing codes are presented, corresponding to 25 (26.3%) cases. The upper-right and lower-left quadrants represent scenarios of opposite behavior between languages, with low expression, containing four (4.2%) of the cases. The predictive capability of clinical passages with information to consider in the clinical coding exercise is similar.

5.1.2. Group 2: Exclusive Evaluation of Procedures

For a better interpretation of Group 2, the exclusive procedure sample (ICD-10-PCS) results are in table format (Table 4), because they have only one or two procedure codes.

Table 4. Evaluation of ICD-10-PCS procedure codes (Group 2).

		Complexity	
		1 Code	**2 Codes**
Representativity		27	2
Correct	**Portuguese**	7.4%	0%
	English	3.7%	0%
Partially correct	**Portuguese**	0%	100%
	English	11.1%	100%
Incorrect	**Portuguese**	92.6%	0%
	English	85.2%	0%
In excess	**Portuguese**	0%	0%
	English	0%	0%

By reading Table 4, it is evident that the accuracy rate is excessively low, namely 7.4% and 3.7%, respectively. For episodes with two codes, there are no observations; they all fall into the partially correct codes category, considering the first three characters as the minimum criterion. From the codes provided by ChatGPT, it can be concluded that they are 92.6% and 85.2% incorrect in Portuguese and English, respectively. There are no observations of excess procedure codes. Similar to the analysis matrix for ICD-10-CM codes, the corresponding matrix for ICD-10-PCS is presented to examine the appropriateness of information interpretation and the returned codes.

According to the matrix in Table 5, there is adequacy of information in 93.1% of the cases studied. On the other hand, the arrangement on the main diagonal of the table indicates that the behavior is indifferent to the language of composition of the medical discharge report, considering that statistically, the predominance of differences by language is correlatable in all cases.

Table 5. Matrix of performance in identifying information returned by ChatGPT (Group 2).

		Difference in English		
		0	**1**	**Count**
Difference in Portuguese	0	27	0	27
	1	0	2	2
	Count	27	2	29

5.1.3. Group 3: Evaluation of Diagnoses and Procedures

Because of the dimension of the samples and the complexity of the clinical records, the results presented are the most relevant.

The investigative rationale from Groups One and Two is maintained for interpreting the results of Group 3. The analysis related to diagnoses exclusively follows. It is noteworthy that diagnoses in episodes with both diagnosis and procedure codes are being studied.

On the *x*-axis, the count value of codes present in the episode is indicated, including diagnoses and procedures, with the evaluation of diagnoses using the ICD-10-CM being considered in the systematized information.

Episodes with two, seven, and 17 codes exhibit a high accuracy rate for diagnoses in both languages, corresponding to 66.7% and 66.7%, 75.0% and 75.0%, and, for the last case, 75.0% and 51.1%, respectively. It is noteworthy that, excluding the last case, no variability differences are observed. On the other hand, in episodes ranging from the 2nd, 3rd, 4th, and 5th levels, corresponding to episodes with three, four, five, and six codes, there is instability in the results, with no apparent monotony based on complexity. The data presented for episodes with three codes are higher in English, with a 10-percentage-point difference between languages, a difference tripled in medical discharge reports with six codes. Intermediate cases, corresponding to four and five, range between 14.3% and 23.1%. As for higher values, there are no correct codes in episodes with eight codes, and with nine codes, there is an 8-percentage-point difference between languages, being higher in English.

Regarding incorrect codes, analyses indicate a potential misinterpretation of clinical reports by ChatGPT.

Medical discharge reports with fewer codes have error rates ranging from 7.7% to 33.3%. Reports with five, eight, and 17 codes exhibit higher error rates.

Through the performance matrix (Table 6), for the preceding groups, ICD-10-CM Group 3, it is observed that 50.0% of cases have an excess of codes in at least one of the languages.

Table 6. Matrix of performance in identifying diagnoses codes returned by ChatGPT (Group 3).

		Difference in English					
		−1	**0**	**1**	**3**	**4**	**Count**
Difference in Portuguese	**−1**	1	-	-	-	-	1
	0	1	10	4	-	-	15
	1	1	4	3	-	-	8
	2	-	-	-	1	-	1
	3	-	-	-	-	1	1
	Count	3	14	7	1	1	26

Approximately 61.5% statistically follows the same interpretation behavior in both languages. Of the procedure codes, only two were correct. It is worth noting that ChatGPT's interpretation of the information reported in the episode was not acceptable (<39.0%).

Considering that only a minimal number of correct codes were obtained, the analysis of ICD-10-PCS results focuses especially on the incorrect codes returned by ChatGPT.

It is noteworthy that the majority of ICD-10-PCS codes are incorrect.

For the ICD-10-PCS procedure codes, the difference presenting in the data is shown in Table 7.

Table 7. Matrix of performance in identifying procedure codes returned by ChatGPT (Group 3).

		Difference in English				
		−3	**−1**	**0**	**1**	**Count**
Difference in Portuguese	**−1**	-	-	1	-	1
	0	1	-	13	1	15
	1	-	1	5	2	8
	2	-	-	-	1	1
	3	-	-	-	1	1
	Count	1	1	19	5	26

The analysis of Table 7 allows us to conclude that the behavior is distinct in Portuguese and English regarding the interpretation and return of information. In English, there is a higher rate of false negatives, while in Portuguese there is a higher incidence of missing codes. Both scenarios pose risks, revealing the immaturity of the responses presented, demonstrating a lack of confidence in the interpretation of procedures.

5.2. Overall Evaluation with Focus on the Episode

Overall evaluation by group with focus on the episode is seen in Table 8. For Group 1 and 2, it is evident that ChatGPT performs better in English than in Portuguese. In Group 1, at the preferred acceptance level of up to 25%, it is observed that the performance in English was approximately five percentage points higher. However, in Group 2, based on the results presented, there is no evidence of utility in this assertion, as more than 90% of cases in both languages have inadequate performance, ranging from 75% to 100% error.

Group 3 has a performance similar in Portuguese and English, differing by one observation. Nevertheless, no episode was classified as Good (<25%) and more than 60% of episodes were classified as Weak or Inadequate in both languages. Thus, the accuracy in diagnostic coding was significantly higher than in procedures.

Table 8. Overall evaluation by group.

Error Classification	Group 1 (Only Diagnoses)		Group 2 (Only Procedures)		Group 3 (Diagnoses & Procedures)	
	Portuguese	**English**	**Portuguese**	**English**	**Portuguese**	**English**
Good	24 (25.3%)	29 (30.5%)	2 (6.9%)	1 (3.5%)	0 (0.0%)	0 (0.0%)
Satisfactory	15 (15.8%)	17 (17.9%)	0 (0.0%)	0 (0.0%)	4 (15.4%)	4 (15.4%)
Weak	6 (6.3%)	5 (5.3%)	0 (0.0%)	0 (0.0%)	5 (19.2%)	2 (7.7%)
Inadequate	50 (52.6%)	44 (46.3%)	27 (93.1%)	28 (96.5%)	17 (65.4%)	20 (76.9%)

6. Discussion

Given this context, the exploration of tools or solutions to assist the coding physician, aiming to enhance efficiency and potentially alleviate the coding workload, is commendable. The performance of ChatGPT as a potential aiding instrument for the coding physician was evaluated considering two key indicators: the individualized and independent assessment of each returned code and the overall integrity of the entire episode in allowing for a comprehensive characterization. Each indicator was independently verified for each of the three categorized episode groups, Group 1 with only diagnoses, Group 2 with only procedures, and Group 3 with both.

The evaluation spectrum considered the behavior of ChatGPT when exposed to queries in the form of medical discharge reports. It was observed that the returned response depended on the clarity of the submitted text. Surprisingly, the word/character count did not seem to influence the response. The episode's structure, categorized into three groups, was identified as one of the determining factors in ChatGPT's performance, as it directly correlates with the complexity of ICD-10-CM/PCS codes.

The 150 episodes were submitted to ChatGPT in both Portuguese and English. The following noteworthy results were observed:

Better performance (higher accuracy, lower rate of incorrect codes, whether excess or missing) in diagnosing with a similar performance in both languages. The composition of ICD-10-CM diagnostic codes, apart from Chapter 19, Injury, Poisoning, and Certain Other Consequences of External Causes, does not exceed four alphanumeric characters. This simplicity explains the results.

Poorer performance in procedural coding. This can be justified by the more complex structure of ICD-10-PCS codes, consisting of seven alphanumeric characters, such as 0SRB02Z Replacement of Left Hip Joint with Metal on Polyethylene Synthetic Substitute, Open Approach.

In partially correct codes, the performance was superior in procedural coding. Generally, ChatGPT correctly returns the first three characters and, in some cases, the first four characters of the Medical and Surgical Section, identifying the Section, Body System, Root Operation, and sometimes the Body Part.

In the case of incorrect codes, the performance is significantly negative in procedures, as explained earlier. A peculiar behavior of ChatGPT was noted, consistently incurring errors in assigning the last three or four characters of the code.

Regarding missing codes, the data do not allow for a conclusion because the number of cases submitted for diagnostic coding was higher than cases with procedures.

Lastly, the return of excess codes, where no clinical information justifying their coding was present, was higher for procedures and identical in both languages. Some authors describe this as "hallucinations" of ChatGPT, returning unsolicited and contextually irrelevant information.

While analyzing the 150 episodes, critical error situations were identified in ChatGPT's performance as a tool to assist the coding physician in their clinical practice. In summary, the categorized interpretation of ChatGPT's behavior during interactions systematically reveals the following behaviors:

Instability in returned responses: Each time questioned about the returned code, particularly in situations of incorrect, missing, or excessive codes, ChatGPT changes its response. This behavior also occurs with correct codes when questioned, occasionally switching to incorrect codes or providing incorrect considerations about them. It should be noted that researchers have the answer a priori. In concluding this study, we find that the inclination of the suggestion was incorrect, indicating a systematic error. Search results do not consistently yield the same results, demonstrating this instability, which generates a lack of confidence in the information provided. However, the information is extremely convincing and even politely educational.

Return of codes with incorrect descriptions: Both existing and non-existent codes are returned with inaccurate descriptions.

Persistent error in clinical coding of laterality: Substitution of right for left and vice-versa.

Shifts responsibility for clinical coding: ChatGPT emphasizes the need for compliance with ICD-10-CM/PCS conventions and guidelines, but never applies the rules in the codification process, redirecting the responsibility to the coding specialist.

The detailed behavior of ChatGPT, along with examples categorized into analysis groups that illustrate the main errors observed during the study of its adequacy as a clinical coding support tool, is described in Supplementary S2.

7. Conclusions

This study allowed for the evaluation of the performance of the 2023 free version of ChatGPT in clinical coding within a significant sample, revealing insufficient results to consider it an automated clinical coding tool, unlike some developments already validated in this field. This process is essential as it does not provide sufficient confidence to replace the coding physician in this exercise. The main results of this study conclude that ChatGPT was not considered an exclusive tool for clinical coding. We found that performance is better for diagnoses than for procedures, decreases with the complexity of the information, and we did not measure significant differences between Portuguese and English. It could be considered a search engine or an aid to the coding physician or auditor. The information returned by ChatGPT in this study ranged from perfect to disastrous, allowing us to assert that the result is of low reliability. There are multiple relationships from data version to other limitations, considered below. On the other hand, sensitivity in the domain of information is reduced, being effective in code association but also concerning laterality.

The limitations of this study include the lack of representativeness of categories to analyze if there is any bias in performance. The free version of ChatGPT was used; it could

be that using the improved version would reveal different results. However, physicians and auditors would likely prefer the free version. The 2023 ChatGPT free version used has data up to 2019 and, knowing that the knowledge management of ICD-10-CM/PCS codes is dynamic and continually updated, this could be a limiting factor to performance, as it counts for about four years.

For future work, other chats with artificial intelligence and even paid versions could be considered to assess performance. Regarding the sample structure, other aspects could be considered, such as language permitting a greater variety of translations, but also considering the representativeness of diagnostic categories and the distribution up to the third character of the procedures, at least. This study would take on another scope and reach, not intended in this present work. There are interesting and disastrous results. Addressing the research question, based on the described context, it is concluded that the use of the free version of ChatGPT is not a tool that can be used unsupervised and that replaces the coding physician in their exercise. Considering the evolution of technology and the constant updating of these applications, we hope to provide a positive answer to the base question in this work soon. Given the results, we cannot affirm that the use is reliable now, urging responsible use by coding physicians, ensuring the quality and rigor of the product of their exercise. Regardless of the quality of clinical coding obtained by AI tools, validation by a certified Clinical Coding Physician is essential.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/informatics11040084/s1, Supplementary S1: Representation of Chapters and Categories; Supplementary S2: Defects in Interpreting Presented Information.

Author Contributions: Conceptualization, B.N.T. and A.L.; methodology, B.N.T. and A.L.; validation, B.N.T. and A.L.; formal analysis, B.N.T. and A.L.; investigation, B.N.T., A.L., G.N., A.C.-F. and F.C.; data curation, B.N.T. and A.L.; writing—original draft preparation, B.N.T. and A.L.; writing—review and editing, B.N.T., A.L., G.N., A.C.-F. and F.C.; visualization, B.N.T., A.L., G.N., A.C.-F. and F.C.; supervision, B.N.T., A.L., G.N., A.C.-F. and F.C.; project administration, B.N.T., A.L., G.N., A.C.-F. and F.C. 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 conflict of interest.

References

1. Tatham, A. The increasing importance of clinical coding. *Br. J. Hosp. Med*. **2008**, *69*, 372–373.
2. Atasoy, H.; Greenwood, B.N.; McCullough, J.S. The digitization of patient care: A review of the effects of electronic health records on health care quality and utilization. *Annu. Rev. Public Health* **2019**, *40*, 487–500.
3. Baumann, L.A.; Baker, J.; Elshaug, A.G. The impact of electronic health record systems on clinical documentation times: A systematic review. *Health Policy* **2018**, *122*, 827–836.
4. Brynjolfsson, E.; McAfee, A. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*; WW Norton & Company: New York, NY, USA, 2014.
5. Menachemi, N.; Collum, T.H. Benefits and drawbacks of electronic health record systems. *Risk Manag. Healthc. Policy* **2011**, *4*, 47–55.
6. INE 2023. Instituto Nacional de Estatística—Estatísticas da Saúde: Lisbon, Portugal, 2021. ISSN 2183-1637. ISBN 978-989-25-0599-2. Available online: https://www.ine.pt/xurl/pub/11677508 (accessed on 11 October 2023).
7. Pordata. Estatísticas Sobre Portugal e Europa. População Residente: Censos, 2023. Available online: https://www.pordata.pt/subtema/portugal/populacao+residente-558 (accessed on 12 October 2023).
8. Dong, H.; Falis, M.; Whiteley, W.; Alex, B.; Matterson, J.; Ji, S.; Wu, H. Automated clinical coding: What, why, and where we are? *NPJ Digit. Med*. **2022**, *5*, 159.
9. Venkatesh, K.P.; Raza, M.M.; Kvedar, J.C. Automating the overburdened clinical coding system: Challenges and next steps. *NPJ Digit. Med*. **2023**, *6*, 16.

10. Schlagwein DWillcocks, L. 'ChatGPT et al.': The ethics of using (generative) artificial intelligence in research and science. *J. Inf. Technol.* **2023**, *38*, 232–238.
11. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.L.; Tang, Y. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA *J. Autom. Sin*. **2023**, *10*, 1122–1136.
12. Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv* **2023**, arXiv:2302.02083.
13. Vaishya, R.; Misra, A.; Vaish, A. ChatGPT: Is this version good for healthcare and research? *Diabetes Metab. Syndr. Clin. Res. Rev*. **2023**, *17*, 102744.
14. Johnson, D.; Goodman, R.; Patrinely, J.; Stone, C.; Zimmerman, E.; Donald, R.; Wheless, L. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. *Res. Square* **2023**, *preprint*.
15. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ*. **2023**, *9*, e45312. https://doi.org/10.2196/45312.
16. Mintz, Y.; Brodie, R. Introduction to artificial intelligence in medicine. Minimally *Invasive Ther. Allied Technol*. **2019**, *28*, 73–81.
17. Purdy, M.; Daugherty, P. Why artificial intelligence is the future of growth. Remarks at AI now: The social and economic implications of artificial intelligence technologies in the near term. *Accenture* **2016**, 1–72.
18. Schmidt, A.; Willard, P.; Krawzik, K.; Kenny, A. ICD-10-CM *Professional for Hospitals. The Complete Official Code set. Optum 360 Coding. n/a. USA*. Optum 360°: Eden Prairie, MN, USA, 2017.
19. Sanmillán, M.; Cebrián, R.N.; Pato-Alonso, S.; Asensio-Villahoz, P.; Salido-Campos, C.; Anso-Borda, I.; Rodríguez-Martinez, G.; Roces Fernández, A.; Gutiérrez Miras, A.; Echevarría Echarri, L.; et al. Manual de Codificación. CIE-10-ES Diagnósticos. Unidad Técnica de Codificación CIE-10-ES. Ministerio de Sanidad, Servicios Sociales e Igualdad, 2016. Available online: https://www.sanidad.gob.es/estadEstudios/estadisticas/normalizacion/CIE10/UT_MANUAL_DIAG_2016_prov1.pdf (accessed on 13 July 2023).
20. Patil, R.; Heston, T.F.; Bhuse, V. Prompt Engineering in Healthcare. *Electronics* **2024**, *13*, 2961.