

ORIGINAL RESEARCH ARTICLE

Artificial intelligence versus humans: A comparative analysis of time, cost, and performance on a clinical code conversion task

Carly Hudson^{1,2,3*}, Marcus Randall², Candice Bowman^{1,4}, Anu Joy^{4,5}, and Adrian Goldsworthy^{1,6,7}¹Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Queensland, Australia²Bond Business School, Bond University, Gold Coast, Queensland, Australia³Faculty of Medicine and Health, University of New England, Armidale, New South Wales, Australia⁴Mental Health and Specialist Services, Gold Coast Hospital and Health Service, Gold Coast, Queensland, Australia⁵School of Applied Psychology, Griffith University, Brisbane, Queensland, Australia⁶Wesley Research Institute, Brisbane, Queensland, Australia⁷Critical Care Research Group, The Prince Charles Hospital, Brisbane, Queensland, Australia**Abstract**

Healthcare services generate and store large quantities of data, requiring significant resources to manually analyze and gain meaningful insights. Recent advancements in automation tools—such as generative artificial intelligence (GenAI)—provide new opportunities to reduce human labor. This study explores the potential utilization of GenAI for a healthcare data analysis task—specifically, the conversion of clinical data from one diagnostic classification system to another (i.e., the Australian extension of the Systematized Nomenclature of Medicine Clinical Terms to the International Classification of Diseases, 10th Revision, Clinical Modification)—and examines the time and cost benefits of performing this using GenAI compared to a human rater. Conversions were completed using three methods: manual conversion using the National Library of Medicine’s I-MAGIC tool, ChatGPT-4o, and Claude 3.5 Sonnet. The accuracy of the GenAI tools was mapped against the manually extracted codes and examined in terms of a perfect, partial, or incorrect match. Task completion time was recorded and extrapolated to calculate and compare the cost associated with each method. When compared to the manually extracted codes, Claude 3.5 Sonnet yielded the highest level of agreement over ChatGPT-4o, whilst being the most time- and cost-effective. GenAI tools have greater utility than they have currently been given credit for. The automation of big data healthcare analytics, whilst still the domain of humans, is increasingly capable of being undertaken using automation tools with low barriers to entry. The further development of GenAI’s capabilities, alongside the capability of the healthcare system to use it appropriately, has the potential to result in significant resource savings.

Keywords: Data analytics; Diagnostic coding; Generative artificial intelligence; International Classification of Diseases 10th revision; Systematized Nomenclature of Medicine Clinical Terms; SNOMED

***Corresponding author:**
Carly Hudson
(chudson@bond.edu.au)

Citation: Hudson C, Randall M, Bowman C, Joy A, Goldsworthy A. Artificial intelligence versus humans: A comparative analysis of time, cost, and performance on a clinical code conversion task. *Artif Intell Health*. 2025;2(4):92-102. doi: 10.36922/AIH025200045

Received: May 12, 2025

Revised: June 9, 2025

Accepted: June 18, 2025

Published online: July 11, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher’s Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The volume of data generated annually by hospitals and health services far exceeds the analytical capacity of humans.¹ Murphy¹ estimated hospitals produce approximately 50 petabytes (equivalent to 50,000,000 gigabytes) of data each year—97% of which remains unanalyzed or unused. Electronic health records contain a wide range of information, including patient demographics, images, clinical notes, and pathology results. These records offer significant potential for retrospective analysis to support data-driven decision-making and more accurate predictions of service utilization.^{1,2} However, increasingly financially constrained and resource-limited healthcare systems lack the capacity to manually process such large datasets, limiting opportunities to improve healthcare system efficiency.^{1,3}

Generative artificial intelligence (GenAI) refers to a type of artificial intelligence algorithm that enables the creation of new content—such as text, images, video, or audio files—based on a set of training data.^{4,5} GenAI has a wide range of applications, including creating illustrations, writing code, and processing datasets.^{4,7} Additionally, GenAI has the potential to support the analysis of large-scale datasets within healthcare settings.^{5,8} Healthcare has traditionally required significant human labor and expertise, and as such, it has often resisted large-scale efforts for effective automation, particularly in the form of clinical and administrative decision-making.⁹⁻¹² A recent literature review by Li *et al.*¹³ has identified some of the key areas in which GenAI is starting to make an impact within healthcare, including generating discharge summaries,¹⁴ determining appropriate screening procedures for a patient,¹⁵ answering clinical questions, and providing medical education.¹⁶⁻¹⁹

The increasing complexity of global healthcare challenges necessitates new data analysis approaches that can expeditiously and efficiently leverage the vast datasets available to healthcare systems. Recent advancements in automation tools, such as GenAI, provide new opportunities to efficiently complete large-scale healthcare data analytics.²⁰ The widespread implementation of GenAI represents one of the most rapid technological advancements in recent years. OpenAI's²¹ ChatGPT is currently one of the most widely used GenAI tools, with over 100 million online users per week.²² ChatGPT allows users to input prompts, commands, or questions and generates corresponding responses. Its interface is driven by a large language model, a form of natural language processing capable of learning and refining its conversational abilities through both self- and semi-structured training.²³ Data processing is carried out using

large-scale neural networks, incorporating feed-forward and convolutional architectures.²³

Following the widespread success of ChatGPT, competitors have since launched other GenAI tools available to the general public, including Google Gemini,²⁴ Microsoft Copilot,²⁵ and Claude.²⁶ The accuracy and completeness of outputs are limited by the data available to the GenAI model (i.e., what it has been trained on, its access to real-time search capabilities), which may be biased or inaccurate. GenAI tools also have limited knowledge of more specialized topics, resulting in a tendency to “hallucinate”—a phenomenon where a GenAI tool generates information to fill knowledge gaps, thereby decreasing the accuracy of outputs.²⁷ Healthcare professionals require an up-to-date understanding of the current and evolving limitations of GenAI in order to optimally select tasks at which it is likely to excel and to prompt it appropriately.

A key challenge in analyzing large-scale healthcare data is ensuring the consistency of data recording across different health services. Standardized diagnostic coding systems help maintain clinical data uniformity by providing a universal language through which diagnoses can be coded and interpreted consistently across healthcare settings. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)²⁸ is a diagnostic coding system utilized by 48 countries (as of August 2024)²⁹ to capture detailed clinical information on procedures, diseases, and clinical findings. SNOMED CT presents diagnoses using both a numeric code (e.g., “230690007”) and a corresponding descriptor (e.g., “Stroke”). It employs a polyhierarchical structure, in which any given code may belong to one or more “parent” categories (e.g., “asthma” may be categorized under both “respiratory diseases” and “allergic conditions”). While SNOMED CT provides a comprehensive framework for patient-level diagnostic coding—encompassing symptoms, procedures, and clinical observations—the system's complexity can pose challenges for users with limited training.

The International Statistical Classification of Diseases and Related Health Problems (ICD)³⁰ is currently the global standard for coding diagnostic information. ICD focuses on the classification of diseases, disorders, and causes of death using alphanumeric codes. These codes are determined using a hierarchical system, in which codes are categorized by chapters (e.g., F: mental and behavioral disorders) and then further subdivided as more detail is provided (e.g., “F30: mood [affective] disorders,” “F30.9: manic episode, unspecified”). Although the ICD provides less detail than SNOMED CT, its broader categories facilitate population health analytics and provide a standard for international health system comparison.

The ability to convert between diagnostic coding systems has practical applications, particularly within research contexts. For instance, extracting a subset of SNOMED CT codes related to a specific diagnostic grouping (e.g., mental health) is challenging, as there are no broader categories for each condition, unlike ICD codes. This presents challenges when working with large SNOMED CT datasets while attempting to analyze only a subset. Converting diagnostic codes can be a time-consuming task, particularly when this process relies heavily on manual data input and extraction. To the authors' knowledge, it remains unexplored whether GenAI can assist in the conversion of clinical data from one diagnostic coding language to another, such as from SNOMED CT to ICD. Such conversions require specialized knowledge of clinical coding and are labor-intensive to complete manually. Performing diagnostic code conversion tasks using AI models may enable less qualified staff to complete the work in less time, thereby reducing the cost of data processing.

Therefore, this study aims to examine whether publicly accessible GenAI tools – namely ChatGPT-4o and Claude 3.5 Sonnet – can accurately convert clinical diagnostic codes from SNOMED CT to the 10th revision of the ICD (ICD-10). This study also seeks to address the following sub-objectives:

- (i) Compare the level of agreement between ChatGPT-4o and a human rater
- (ii) Compare the level of agreement between Claude 3.5 Sonnet and a human rater
- (iii) Compare the level of agreement between ChatGPT-4o and Claude 3.5 Sonnet
- (iv) Examine the economic benefit, in terms of time and labor cost, of using GenAI to complete this task compared to a human rater.

2. Materials and methods

The SNOMED CT codes used in this study originate from a broader emergency department (ED) dataset, obtained as part of a study investigating mental health presentations to hospital EDs (ethics approval: HREC/2023/QGC/95219). This dataset consists of 19,764 unique SNOMED-CTAU (Australian Extension) numeric codes (e.g., 48694002) and SNOMED-CT-AU names (e.g., “Anxiety reaction”) representing the diagnoses made to the ED over a 3-year period (August 2020 to August 2023). The current evaluation utilizes a randomly selected 10% subset of this data ($n = 1,976$) (Table S1).

To convert the SNOMED CT-AU³¹ codes to ICD-10 Clinical Modification (ICD-10-CM),³² a three-phase approach was employed. First, codes were manually converted by human raters. Second, the codes were

converted using ChatGPT-4o (<https://chatgpt.com/>). Third, the same set of codes was converted using Claude 3.5 Sonnet (<https://claude.ai/>). Both GenAI tools required paid subscriptions at the time of analysis.

The methodology and results of this study were reported in accordance with the METRICS reporting checklist, which outlines standardized reporting metrics – such as model, evaluation, timing, transparency, range of tested topics, randomization, individual factors, query count, and prompt specificity – for GenAI-based studies in healthcare.³³ The completed reporting checklist is listed in Table S2.

2.1. Phase 1: Manual conversion of SNOMED-CT-AU codes

The SNOMED CT-AU codes were manually converted by a team of three raters (AG = 800 codes; AJ = 644 codes; CH = 532 codes). Conversions were performed using the Interactive Map-Assisted Generation of ICD Codes (I-MAGIC) algorithm (<https://imagic.nlm.nih.gov/imagic/code/map>), an online tool that provides a mapping between the two diagnostic coding systems.³⁴ Codes were entered into the tool in the format “SNOMED CT-AU name (SNOMED CT-AU code)” (e.g., “Anxiety reaction [48694002]”), and the corresponding ICD-10-CM code was extracted.

In this study, the I-MAGIC tool was employed as the reference standard against which all other conversion methods were compared. However, some SNOMED CT codes could not be located within the I-MAGIC database. As the dataset utilized the Australian extension of SNOMED CT, while the mapping tool used the standard SNOMED CT list, it is likely that the missing codes were region-specific.³⁵ In such cases, the absence of an equivalent was noted.

2.2. Phase 2: Conversion of SNOMED-CT-AU codes using ChatGPT-4o

ChatGPT-4o²¹ was used to automatically convert the SNOMED CT-AU codes and names into ICD-10-CM codes (completed in August 2024). A Microsoft Excel file containing the SNOMED CT-AU codes and names was uploaded to ChatGPT-4o. The prompt used for the conversion was refined through an iterative process to improve efficiency and reduce the risk of “hallucinations” (i.e., providing false information) and data processing errors.

It was necessary to state that ChatGPT4o could take as much time as required to complete this task, otherwise the message would time out and cease to produce output. Additionally, a limit was observed regarding the number

of codes that could be processed simultaneously without increasing the likelihood of errors, such as incorrect conversions, fabricated inputs, or skipped entries. To manage this, codes were processed in batches of five, with five batches (25 codes in total) converted at a time.

The following final prompt was used:

Please manually and sequentially convert the SNOMED-CT-AU codes into ICD-10-CM codes and names which have not been converted yet. If you need to, complete this task in batches of 5. Take as long as you need. Do not hallucinate, and only convert codes which have been provided to you. Do not create new codes to convert. Provide an update after 5 batches have been completed. Provide a .xlsx file at the end.

Figure 1 presents the final prompt and the corresponding output from ChatGPT-4o.

Following each conversion of 25 codes, ChatGPT-4o needed to be prompted to continue (Figure 2):

Please continue.

Following conversion, ChatGPT-4o generated the requested Microsoft Excel file containing the original input data and the corresponding final output data.

Please manually and sequentially convert the SNOMED CT AU codes into ICD-10-CM codes and names which have not been converted yet. If you need to, complete this task in batches of 5. Take as long as you need. Do not hallucinate, and only convert codes which have been provided to you. Do not create new codes to convert. Provide an update after 5 batches have been completed. Provide a .xlsx file at the end.

Memory updated

I have identified the first batch of 5 SNOMED CT AU codes that require manual conversion to ICD-10-CM codes and names. I will now proceed with the conversion process.

Please hold on while I perform the necessary research and mapping. [...]

Batch 1 of 5 Completed:

- SNOMED CT AU Code: 15692007
 - SNOMED CT AU Name: Right lateral brow position
 - ICD-10-CM Code: N/A (This term likely represents a specific observation or position rather than a medical diagnosis, and may not have a direct ICD-10-CM equivalent.)
- SNOMED CT AU Code: 301003007
 - SNOMED CT AU Name: Right middle lobe pneumonia
 - ICD-10-CM Code: J18.1
 - ICD-10-CM Name: Lobar pneumonia, unspecified
- SNOMED CT AU Code: 111589005
 - SNOMED CT AU Name: Dysfibrinogenaemia
 - ICD-10-CM Code: D68.2
 - ICD-10-CM Name: Hereditary deficiency of other clotting factors (Dysfibrinogenemia can be coded here as it is a clotting disorder.)

Figure 1. ChatGPT-4o prompt and output
Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

2.3. Phase 3: Conversion of codes from the Australian extension of the Systematized Nomenclature of Medicine Clinical Terms using Claude 3.5 Sonnet

At the time of the study (September 2024), Claude 3.5 Sonnet²⁶ was unable to import or export Microsoft Excel files. Therefore, data were copied and pasted directly into the chat as a means of input. However, there were limitations on the maximum allowable message length. As a result, the data were entered in three batches of 500, followed by a final batch of 476.

The prompt was developed in a similar manner to Phase 2, using an iterative approach. Claude 3.5 Sonnet required fewer instructions to accurately perform the task, allowing the prompt to be developed relatively quickly. However, the limits of output message length restricted Claude 3.5 Sonnet to processing only 50 codes at a time.

During prompt testing, it was noted that Claude 3.5 Sonnet needed to be explicitly instructed not to skip lines, as reflected in the final prompt (Figure 3):

Please convert these SNOMED CT-AU codes and names into ICD-10-CM codes and names. Convert them in sequential order starting from the top and ensuring you do not skip any. I have provided 500 codes in total and need 500 responses at the end. Start with the first 50. If you are unable to convert a code, please state this.

The next five batches of SNOMED CT AU codes have been successfully converted to ICD-10-CM codes and names. The updated dataset has been saved.

You can download the final file using the link below:

[Download the final converted file](#)

Would you like to continue with more conversions, or is there anything else I can assist you with? [...]



Please continue

I will proceed with the next batch of five SNOMED CT AU codes and continue the conversion process. Let's start with the next batch.

Here are the next five SNOMED CT AU codes and their corresponding names:

- 439127006 - Thrombosis
- 239778009 - Wrist pyogenic arthritis
- 197105006 - Perforated diverticulum of duodenum
- 1244471000168100 - Comminuted fracture of clavicle
- 247165009 - Retinal detachment - subretinal fluid

I will now proceed with manually converting these to ICD-10-CM codes and names. This will take a moment. [...]

Figure 2. ChatGPT-4o is prompted to continue with the next batch of conversions

Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

As Claude 3.5 Sonnet was unable to directly export a Microsoft Excel file, it was instead instructed to produce R

code that could be run to generate an output file (Figure 4). The following request was made: “Can you write the code for R to create an Excel file of this data. Write it out in full so it exports all entries 1 – 100.” Due to limitations in output message length, this process was completed in batches of 100. The generated code was then run using R Studio (Posit, USA) to produce the final output file.

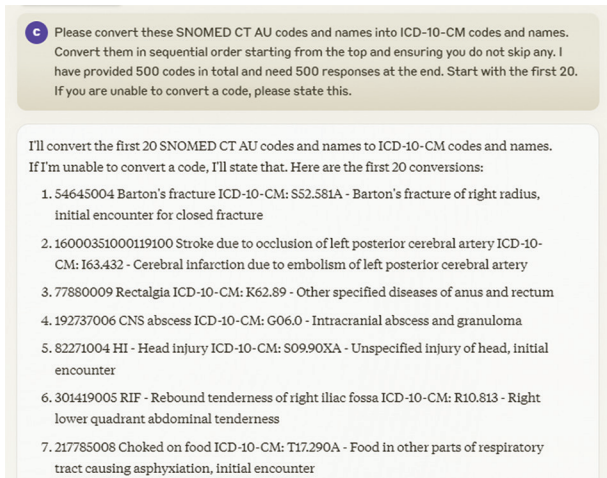


Figure 3. Claude 3.5 Sonnet prompt and output
 Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

2.4. Statistical analysis

For the purpose of analysis, the ICD codes were split into three components: (i) letter code (“Level 1”); (ii) major numeric code (before the decimal point: “Level 2”); and (iii) minor numeric code (after the decimal point: “Level 3”) (Table 1).

A pattern-matching program was developed using the C programming language to identify partial and perfect matches among the 1,970 cases between: (i) manual coding and ChatGPT-4o; (ii) manual coding and Claude 3.5 Sonnet; and (iii) ChatGPT-4o and Claude 3.5 Sonnet. The program converted the codes from each method into their component parts. Some manual editing was necessary

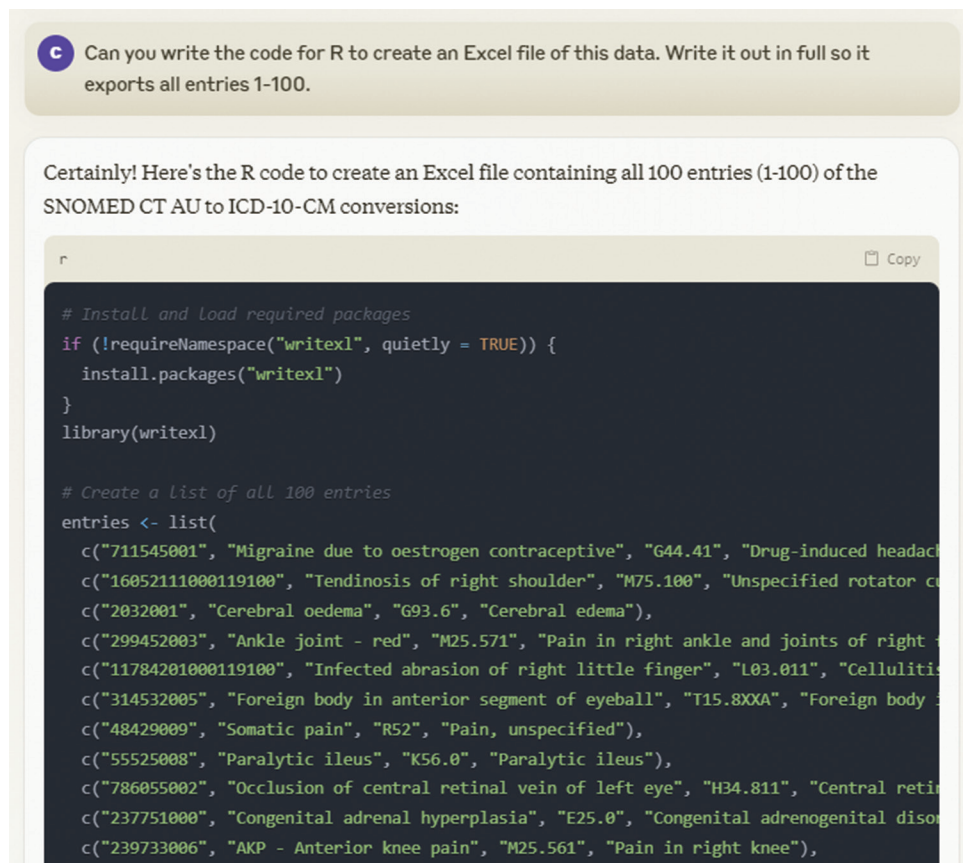


Figure 4. Claude 3.5 Sonnet generating R code to create a Microsoft Excel file
 Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

to address missing values, which were substituted with null codes.

Using this program, the number of matches found for each comparison was categorized as follows: perfect match, Level 1 partial match, Level 2 partial match, and incorrect match (Table 1). A Chi-squared test of independence was conducted to determine whether there was a statistically significant difference in the number of good matches (perfect and Level 2 partial matches) and poor matches (Level 1 partial and incorrect matches) across the three methods.

2.5. Time and cost analysis

The time required to perform conversions in each phase was recorded to allow for a comparison of the time and

Table 1. Components of the International Classification of Diseases codes used to identify matches

	Level 1	Level 2	Level 3
Example ICD Code	F	30	.9
Perfect match	Yes	Yes	Yes
Level 2 partial match	Yes	Yes	No
Level 1 partial match	Yes	No	No
Incorrect match	No	No	No

Abbreviation: ICD: International Classification of Diseases.

Table 2. Number of correct matches across comparisons

Match category	Manual coding vs. ChatGPT-4o (%)	Manual coding vs. Claude 3.5 Sonnet (%)	ChatGPT-4o vs. Claude 3.5 Sonnet (%)
Perfect match	578 (29.34)	599 (30.41)	757 (38.43)
Level 2 partial match	751 (38.12)	758 (38.48)	763 (38.73)
Level 1 partial match	235 (11.93)	212 (10.76)	230 (11.68)
Incorrect match	406 (20.61)	401 (20.36)	220 (11.17)

Table 3. Time and cost for each method

Method and scenario	Time	FTEs in weeks ^a	Labor cost (AUD)	Cost of GenAI tool (AUD)	Total cost (AUD)
10% subset (n=1976)					
Manual coding	24 h and 31 min	0.64	\$1,279.77	N/A	\$1,279.7731
ChatGPT-4o	5 h and 45 min	0.15	\$300.15	\$30.00	\$330.15
Claude 3.5 Sonnet	3 h and 10 min	0.08	\$165.30	\$30.00	\$195.30
Extrapolation for full dataset (n=19,764)					
Manual coding	245 h and 12 min	6.45	\$12,799.44	N/A	\$12,799.44
ChatGPT-4o	57 h and 30 min	1.51	\$3,001.50	\$30.00	\$3,031.50
Claude 3.5 Sonnet	31 h and 40 min	0.83	\$1,653.00	\$30.00	\$1,683.00

Note: ^aAssumes a 38-h work week.

Abbreviations: FTE: Full time equivalent; N/A: Not available.

labor costs associated with each method. The cost of completing the task was calculated by multiplying the time taken for each method by the hourly wage of a research assistant, which was set at AUD\$52.20/hour (based on the pay rate for a university-employed research assistant, excluding on-costs). Setup costs – namely, the cost of subscribing to ChatGPT-4o or Claude 3.5 Sonnet – were also included in the total cost calculation.

3. Results

Table 2 displays the number of each type of match found for each of the comparisons. A Chi-squared test of independence was conducted to examine differences in the number of good and poor matches among manual coding, ChatGPT-4.0, and Claude 3.5 Sonnet. The analysis reveals a statistically significant difference in agreement across the three comparisons (χ^2 [df = 2] = 56.722, $p < 0.001$).

Agreement on good matches varies considerably between method pairs. The ChatGPT-4.0 and Claude 3.5 Sonnet pair show the highest agreement, producing good matches for 1,520 cases (77.2%) compared to 1,329 cases (67.5%) for manual coding versus ChatGPT-4o and 1,357 cases (68.9%) for manual coding versus Claude 3.5 Sonnet.

Table 3 displays the time and associated cost for a research assistant to perform data conversions using each tool for the 10% subset (n = 1,976) included in this study. It also includes an extrapolated estimate of costs if the entire dataset (n = 19,764) were to be converted from SNOMED to ICD.

Of the three methods used, manual coding was the most time-consuming and costly, taking 24 h and 31 min (AUD\$1,279.7) to convert the subset utilized in this study. When extrapolated to the full dataset, this method is estimated to require 245 h and 12 min, with a labor cost of AUD\$12,799.44.

Of the two GenAI methods, Claude 3.5 Sonnet was the most time- and cost-efficient, requiring 3 h and 10 min (AUD\$195.30, including subscription cost). ChatGPT-4o nearly doubled the time and cost of Claude 3.5 Sonnet, taking 5 h and 45 min (AUD\$195.30, including subscription cost). Regardless, ChatGPT-4o still demonstrated significant time and cost savings compared to manual conversion.

4. Discussion

This evaluation provides a case study to investigate the ability of GenAI tools to process and analyze large-scale healthcare datasets. To the authors' knowledge, this study is the first to challenge GenAI tools to complete a clinical diagnostic coding conversion task and to compare the results against those of a manual rater. Conversion of clinical diagnostic codes to other coding systems, such as the task presented in this study, is a complex and time-consuming task commonly undertaken within healthcare data processing. Therefore, this study highlights an example of a potential use for GenAI within health data analytics.

The analysis in this study examined matches found between the two GenAI tools and the manual rater. The results indicated that the two GenAI tools showed a higher level of agreement than either of them did compared to the manual coding, suggesting that the GenAI methods may employ similar coding strategies or have overlapping strengths in code conversion that differ from manual coding approaches.

However, when interpreting these findings, there are several caveats to consider. For instance, the clinical validity of ICD codes—particularly in cases where these were identified as “partial” or “incorrect matches”—was not assessed. This may have resulted in several potentially valid codes being incorrectly coded. For example, the SNOMED code “314041007 Abdominal pain in early pregnancy” was manually converted to “R10.9 Unspecified abdominal pain”. As this formed the benchmark for comparison between the GenAI tools, conversions made by ChatGPT-4o (“O26.83 Pregnancy related abdominal pain”) and Claude 3.5 Sonnet (“O26.892 Other specified pregnancy related conditions, first trimester”) were considered as incorrect matches.

During the analysis, the GenAI tools identified additional—or arguably better—matches between SNOMED CT and ICD-10-CM. Additionally, there were several cases where the I-MAGIC tool was unable to generate a match for a SNOMED CT code (e.g., “102508009 Well female child”), whereas ChatGPT-4o and Claude 3.5 Sonnet were both able to produce the same alternative

ICD-10-CM code (i.e., “Z00.129 Encounter for routine child health examination without abnormal findings”). This suggests that further formal analysis may demonstrate that GenAI tools outperform human raters. Therefore, it is likely that the results of this study significantly underestimate the accuracy and clinical validity of the matches produced by the GenAI tools.

Despite GenAI tools demonstrating significant time and cost savings, several challenges were noted throughout the conversion process. With regards to ChatGPT-4o, the process of performing the SNOMED CT-AU to ICD-10-CM conversion was not fully automated, nor was it straightforward for someone inexperienced with writing GenAI prompts to perform. When piloting the prompt, ChatGPT-4o tended to skip lines, chunks of data, or “hallucinate” (i.e., produce new input data that was not provided in the dataset). It was therefore necessary to explicitly instruct ChatGPT-4o to “manually and sequentially” convert the provided codes and to “...not hallucinate, and only convert codes which have been provided...” and “...not create new codes to convert.” When completing the final batch of conversions, the output had to be monitored for accuracy. Despite not hallucinating during the task, ChatGPT-4o still produced new input data when it ran out of the codes it had been provided.

When providing additional prompts after the algorithm had performed well, it was beneficial to provide positive reinforcement to inform ChatGPT-4o that it had performed the task correctly. This avoided ChatGPT-4o from changing its original output. There were also instances where ChatGPT-4o would attempt to terminate the task (i.e., “Unfortunately I have run out of time to process additional conversions”) but could be prompted to continue without further issue. These nuances required some level of skill and familiarity with ChatGPT-4o and GenAI prompts.

In terms of the time and labor required, ChatGPT-4o was not simply a “set and forget” solution to a large data task. Due to limitations on the volume of codes it was able to process before sometimes hallucinating, a manual “nudge” (i.e., “Please continue with the next batch”) was required after every 25 codes had been converted. This required continual monitoring of ChatGPT-4o while it was processing to ensure that lines of data were not skipped. Importantly, this renders the task impractical to complete in the background while undertaking other work.

ChatGPT-4o also imposes limits on the number of messages that are permitted within a certain timeframe (40 messages every three h). Given the number of nudges

required to process this data – in addition to further messages to adapt and rectify the prompt if it was not processing correctly – the message limit was quickly reached and required waiting until the window had lapsed before proceeding with the rest of the task. This drastically inflated the timeframe in which the task could be completed.

Claude 3.5 Sonnet provided a more streamlined tool that did not require as much skill or time to produce a prompt. One key limitation of Claude 3.5 Sonnet was the process of importing and exporting data. Unlike ChatGPT-4o, at the time of the study, Claude 3.5 Sonnet did not have the functionality to directly import or export Microsoft Excel files; however, this functionality has since been added with the release of Claude 4.0 Sonnet. Therefore, it was necessary to copy and paste lines of data from the Microsoft Excel file into Claude 3.5 Sonnet. This led to a further limitation, which was the restrictions on both message length and the number of messages permitted. As the amount of data exceeded the input limit, it was necessary to break up the prompt into smaller, more manageable batches of codes (i.e., 500 lines at a time).

Although Claude 3.5 Sonnet did not appear to “hallucinate” with a greater number of conversions, only 50 codes could be converted at a time due to limits on the maximum output message length. This however meant that the message limit (approximately 45 messages every 5 h, dependent on message length) was quickly consumed. Given that Claude 3.5 Sonnet processed codes significantly faster than ChatGPT-4o, this led to a longer waiting period between exceeding the message limit and its renewal. As Claude 3.5 Sonnet was unable to directly export a Microsoft Excel file at the end of the task, this significantly increased the time burden, as it was necessary to produce R Studio code to be run in order to produce the final output dataset. In addition to requiring the worker to have some knowledge of how to run the code in R Studio, this step accounted for the majority of the time taken to complete the task. For instance, it took 1 h and 15 min to complete the code conversion, with the remainder of the time (1 h and 55 min) spent writing and executing the R Studio code. The ability to produce downloadable Microsoft Excel files within Claude 3.5 Sonnet would rectify this limitation, significantly reducing the time and cost required to complete data analysis.

Although other methods are available for large-scale data extraction tasks, such as the creation of Application Programming Interfaces, these may require technical skill and knowledge to set up. These may also be cumbersome and impractical for ad hoc tasks performed by individuals lacking programming skills, particularly those in a busy

clinical or hospital environment.³⁶ GenAI tools remain an accessible and easy-to-use alternative that requires minimal training to achieve a cost- and time-efficient outcome. Additionally, these tools are rapidly improving over time, potentially simplifying the task even further.

4.1. Study limitations

Although this case study provides valuable insights into the use of GenAI to complete a large-scale health data analysis task, several limitations still remain. Firstly, given that this is an Australian dataset, the SNOMED-CT codes came from the Australian edition (SNOMED-CT-AU) whilst the I-MAGIC tool only caters to the standard version. Therefore, this may account for why some codes were unable to be manually converted using the I-MAGIC tool. Additionally, multiple raters were required to complete the manual coding task, thereby introducing potential issues around inter-rater reliability, particularly when coders were less familiar with the task. Furthermore, the I-MAGIC tool currently uses ICD-10-CM and has not yet been updated for the new edition of the ICD (i.e., 11th edition). There is currently no mapping tool available that enables SNOMED CT to be converted to the newer version of the ICD.

In addition, this study only considered ICD-10-CM codes to be “correct” if they either perfectly or partially matched the manual code. Given that the aim of this study was to examine whether this task could be completed using GenAI, it was outside of the scope of the study to manually examine each “incorrect” match to determine whether it was clinically valid. However, this is likely to have significantly impacted the results and led to an underestimation of the level of agreement between the GenAI tools and manual ratings.

A further limitation of this study is the rapid pace at which GenAI tools are being developed and improved. It is likely that in the time since this study was conducted, newer tools have been released that may yield different results in terms of accuracy and processing speed. However, these advancements will likely only improve the overall efficiency and accuracy of GenAI tools.

4.2. Recommendations for future research

There is significant scope for future research within this field. Firstly, further analysis of the produced data from this study is planned to examine the clinical validity of partial or incorrect matches, which will further strengthen the results of this study by producing more accurate ratings between the GenAI and manual coding output. This study used the paid versions of both ChatGPT-4o and Claude 3.5 Sonnet, which offer additional functionalities

and greater processing speed than their free versions. This study could be replicated using the free versions of these tools to compare whether the paid versions yield any difference in terms of level of agreement and processing time. It is also yet to be determined whether the time and cost savings observed in this task would translate to other data conversion tasks. Further studies using GenAI tools are needed to determine whether time and cost differences are consistent across different types of tasks. Additionally, as new GenAI tools—such as DeepSeek³⁷—are released with improvements in speed and functionality, it is recommended that this study be repeated to examine how these improvements impact the speed and accuracy with which this task can be completed. Although ChatGPT-4o and Claude 3.5 Sonnet are not specifically designed for healthcare applications, these tools were selected due to their relatively low cost and wide accessibility. The task presented in this study should also be repeated using GenAI tools specifically designed for clinical or healthcare contexts. Furthermore, the completion of similar tasks using GenAI tools should be considered to further explore their capabilities in healthcare data processing.

5. Conclusion

This study presents a case study demonstrating the use of GenAI tools to complete manual data processing tasks that are typically tedious, time-consuming, costly, and both mentally and physically demanding. The findings highlight that manual processing is often prohibitive in terms of time and cost and that alternative methods – such as the use of GenAI – warrant further exploration. GenAI provides a promising avenue for leveraging the vast quantities of currently unanalyzed healthcare data, with the potential to enhance outcomes for healthcare professionals, researchers, systems, and, most importantly, patients.

Acknowledgments

None.

Funding

This study was supported by an Australian Government Research Training Program Scholarship.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Carly Hudson

Data curation: Carly Hudson, Anu Joy, Adrian Goldsworthy

Formal analysis: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Investigation: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Methodology: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Project administration: Carly Hudson

Resources: Candice Bowman

Supervision: Marcus Randall, Candice Bowman

Writing – original draft: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Writing – review & editing: All authors

Ethics approval and consent to participate

This research was approved by the Human Research Ethics Committee of the Gold Coast Hospital and Health Service (HREC/2023/QGC/95219).

Consent for publication

Not applicable.

Availability of data

Data is available from the corresponding author upon reasonable request.

Further disclosure

This research has been preprinted on Research Square (<https://doi.org/10.21203/rs.3.rs-5143761/v1>).

References

1. Murphy K. How data will improve healthcare without adding staff or beds. Cornell university, institut européen d'administration des affaires. In: Organisation WIP, editors. *Global Innovation Index 2019*. Ch. 8. Geneva: World Intellectual Property Organization; 2019.
2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: Management, analysis and future prospects. *J Big Data*. 2019;6(1):54.
doi: 10.1186/s40537-019-0217-0
3. Australian Medical Association. *2024 Public Hospital Report Card*; 2024.
4. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inform Syst Eng*. 2024;66(1):111-126.
doi: 10.1007/s12599-023-00834-7
5. Oluwagbenro MB. *Generative AI: Definition, Concepts, Applications, and Future Prospects. Authorea Preprints*; 2024.
6. Banh L, Strobel G. Generative artificial intelligence. *Electron Mark*. 2023;33(1):63.
doi: 10.1007/s12525-023-00680-1
7. Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and chatGPT: Applications, challenges, and

- AI-human collaboration. *J Inform Technol Case Appl Res*. 2023;25:277-304.
doi: 10.1080/15228053.2023.2233814
8. Chatterjee S, Fruhling A, Kotiadis K, Gartner D. Towards new frontiers of healthcare systems research using artificial intelligence and generative AI. *Health Syst (Basingstoke)*. 2024;13:263-273.
doi: 10.1080/20476965.2024.2402128
9. Yang Y, Ngai EW, Wang L. Resistance to artificial intelligence in health care: Literature review, conceptual framework, and research agenda. *Inform Manage*. 2024;61:103961.
doi: 10.1016/j.im.2024.103961
10. Amin SM, El-Gazar HE, Zoromba MA, El-Sayed MM, Atta MHR. Sentiment of nurses towards artificial intelligence and resistance to change in healthcare organisations: A mixed-method study. *J Adv Nurs*. 2025;81(4):2087-2098.
doi: 10.1111/jan.16435
11. Chaibi A, Zaiem I. Doctor resistance of artificial intelligence in healthcare. *Int J Healthcare Inform Syst Inform (IJHISI)*. 2022;17(1):1-13.
doi: 10.4018/IJHISI.315618
12. Sobaih AEE, Chaibi A, Brini R, Abdelghani Ibrahim TM. Unlocking patient resistance to AI in healthcare: A psychological exploration. *Eur J Investig Health Psychol Educ*. 2025;15(1):6.
doi: 10.3390/ejihpe15010006
13. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed*. 2024;245:108013.
doi: 10.1016/j.cmpb.2024.108013
14. Patel SB, Lam K. ChatGPT: The future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107-e108.
doi: 10.1016/S2589-7500(23)00021-3
15. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. *MedRxiv*. 2023.
doi: 10.1101/2023.02.02.23285399
16. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. *Resuscitation*. 2023;185:109729.
doi: 10.1016/j.resuscitation.2023.109729
17. Boßelmann CM, Leu C, Lal D. Are AI language models such as ChatGPT ready to improve the care of individuals with epilepsy? *Epilepsia*. 2023;64(5):1195-1199.
doi: 10.1111/epi.17570
18. Mogali SR. Initial impressions of ChatGPT for anatomy education. *Anat Sci Educ*. 2024;17(2):444-447.
doi: 10.1002/ase.2261
19. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care*. 2023;46(5):e103-e105.
doi: 10.2337/dc23-0197
20. Health-Europe TLR. Embracing generative AI in health care. *Lancet Reg Health Eur*. 2023;30:100677.
doi: 10.1016/j.lanepe.2023.100677
21. Open AI. *ChatGPT*. Available from: <https://chat.openai.com> [Last accessed on 2024 Sep 06].
22. Thorne E. *ChatGPT Hits 100M Weekly Users*. LinkedIn News. Available from: <https://www.linkedin.com/news/story/chatgpt-hits-100m-weekly-users-5808204> [Last accessed on 2024 Sep 09].
23. Thakur K, Barker HG, Khan Pathan AS. *Artificial Intelligence and Large Language Models: An Introduction to the Technological Future*. United States: CRC Press, LLC; 2024.
24. Google AI. *Gemini*. Available from: <https://gemini.google.com> [Last accessed on 2024 Sep 09].
25. Microsoft. *Copilot*. Available from: <https://copilot.microsoft.com> [Last accessed on 2024 Sep 09].
26. Anthropic. *Claude AI*. Available from: <https://claude.ai> [Last accessed on 2024 Sep 06].
27. Waldo J, Boussard S. GPTs and hallucination: Why do large language models hallucinate? *Queue*. 2024;22(4):19-33.
doi: 10.1145/3688007
28. Snomed International. *What is Snomed CT?* Available from: <https://www.snomed.org/what-is-snomed-ct> [Last accessed on 2024 Sep 09].
29. Snomed International. *Members*. Available from: <https://www.snomed.org/members> [Last accessed on 2024 Sep 16].
30. World Health Organisation. *International Statistical Classification of Diseases and Related Health Problems (ICD)*. Available from: <https://www.who.int/standards/classifications/classification-of-diseases> [Last accessed on 2024 Sep 09].
31. National Library of Medicine. *Snomed-CT Browser*. Available from: <https://www.nlm.nih.gov/research/umls/snomed/snomed-browsers.html> [Last accessed on 2025 Jun 09].
32. World Health Organisation. *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. Available from: <https://icd.who.int/browse10/2019/en> [Last accessed on 2025 Jun 09].
33. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: Development study involving a literature review. *Interact J Med Res*. 2024;13:e54704.

doi: 10.2196/54704

34. National Library of Medicine. *Snomed CT to ICD-10-cm Map*. Available from: <https://www.nlm.nih.gov/research/umls/mapping-projects/snomedct-to-icd10cm.html> [Last accessed on 2024 Sep 06].
35. National Clinical Terminology Service. *What is Snomed CT and the AMT?* Available from: <https://www.healthterminologies.gov.au/understanding-clinical-terminology-landing/what-is-snomed-ct-and-amt> [Last accessed on 2024 Sep 23].
36. Preibisch S. *API Development*. Germany: Springer; 2018.
37. Hangzhou DeepSeek Artificial Intelligence Co. *DeepSeek*. Available from: <https://www.deepseek.com> [Last accessed on 2025 Jun 09].