COMPARATIVE ANALYSIS OF THE ACCURACY CODE OF MANUAL AND ARTIFICIAL INTELLIGENCE (AI) AT THE PUSKESMAS UMBULHARJO I IN 2025

Rezqa Amaliza[1], Nita Budiyanti, S.KM., M.H[2], Syarah Mazaya Fitriana, S.KM., M.P.H[3], Abdul Hadi Kadarusno, S.KM., M.P.H[4], Mutiara Pertiwi, A.Md[5], Rahmah Nindyakinanti, A.Md[6]
[1,2,3,4]Prodi DIII Rekam Medis dan Informasi Kesehatan Poltekkes Kemenkes Yogyakarta, [5,6]Puskesmas Umbulharjo I
Jl. Mangkuyudan MJ III/304, Kota Yogyakarta, Daerah Istimewa Yogyakarta
Email: rezqamaliza@gmail.com, nitabudiyanti.nita@gmail.com, syarah.fitriana@poltekkesjogja.ac.id, abdul.hadik@poltekkesjogja.ac.id, mutekmutiara@gmail.com, rnindyakinanti@gmail.com

ABSTRACT

Background: *Inaccurate ICD-10 diagnosis codes in primary health centers (puskesmas) negatively affect the quality of morbidity data and disease reporting, with errors commonly occurring at the code subcategory level (Munandziroh et al., 2024), while AI has been proven capable of improving coding accuracy and consistency as a supporting tool for health workers (Baigi et al., 2025.) A preliminary study at Puskesmas Umbulharjo I examining 60 outpatient medical records revealed inaccuracies in 35 manual-coded records (57.38%) and only 4 AI-coded records (6.56%), all falling within the subcategory group. This underscores the need to compare the accuracy of both methods and identify their error groups as a basis for recommending AI as a coding assistance tool.*

Objective: *This study aims to analyze the difference in ICD-10 diagnosis code accuracy between the manual method used at Puskesmas Umbulharjo I and the AI method using Claude, as well as to identify diagnosis code error groups based on expert judgment validation.*

Methods: *This study employed a descriptive quantitative approach with a cross-sectional design. The sample consisted of 398 outpatient medical records from 2025, obtained through simple random sampling from a population of 68,995 records. Manual codes were compared with AI-generated codes produced by Claude using a structured prompt, and both were validated against expert judgment codes in accordance with ICD-10 Volume 2 (2010). Reliability testing was conducted using the test-retest method across two separate testing sessions to assess Claude's consistency in generating codes.*

Results: *Based on the researcher's observations, diagnosis coding was not carried out in accordance with standard operating procedures (SOPs), there was no final validation by a Medical Records and Health Information Officer (PMIK), and frequent system loading issues occurred in the SIMPUS application. Manual codes were accurate in 93 medical records (23.37%), while AI codes were accurate in 345 medical records (86.68%). Manual coding errors were predominantly subcategory errors in 291 cases (95.41%) and category errors in 14 cases (4.59%). All AI coding errors were subcategory-level (100%). Claude's consistency reached 99.50%, with only 2 medical records producing different codes between sessions.*

Conclusion: *There is a significant accuracy gap between manual coding (23.37%) and AI coding (86.68%). Manual coding errors stem from non-compliance with SOPs, the absence of final PMIK validation, and limitations of the SIMPUS system. Manual codes experienced errors at both the category and subcategory levels, whereas all AI code errors were subcategory-level only. Claude's high consistency makes it a promising coding assistance tool, however final verification by a PMIK officer remains essential.*

Keywords: *Accuracy Code Diagnosis, Artificial Intelligence, Claude, Code Error Groups*

ANALISIS PERBANDINGAN AKURASI KODE DIAGNOSIS SECARA MANUAL DAN *ARTIFICIAL INTELLIGENCE (AI)* DI PUSKESMAS UMBULHARJO I TAHUN 2025

Rezqa Amaliza[1], Nita Budiyanti, S.KM., M.H[2], Syarah Mazaya Fitriana, S.KM., M.P.H[3], Abdul Hadi Kadarusno, S.KM., M.P.H[4], Mutiara Pertiwi, A.Md[5], Rahmah Nindyakinanti, A.Md[6]
[1,2,3,4]Prodi DIII Rekam Medis dan Informasi Kesehatan Poltekkes Kemenkes Yogyakarta, [5,6]Puskesmas Umbulharjo I
Jl. Mangkuyudan MJ III/304, Kota Yogyakarta, Daerah Istimewa Yogyakarta
Email: rezqamaliza@gmail.com, nitabudiyanti.nita@gmail.com, syarah.fitriana@poltekkesjogja.ac.id, abdul.hadik@poltekkesjogja.ac.id, mutekmutiara@gmail.com, rnindyakinanti@gmail.com

ABSTRAK

Latar Belakang: Ketidakakuratan kode diagnosis ICD-10 di puskesmas berdampak pada kualitas data morbiditas dan pelaporan penyakit, dengan kesalahan yang umumnya terjadi pada subkategori kode (Munandziroh *et al.*, 2024), sementara *AI* terbukti mampu meningkatkan akurasi dan konsistensi kodefikasi sebagai alat bantu petugas (Baigi *et al.*, 2025). Studi pendahuluan di Puskesmas Umbulharjo I terhadap 60 rekam medis rawat jalan menunjukkan ketidakakuratan kode manual sebesar 35 rekam medis (57,38%) dan kode *AI* hanya 4 rekam medis (6,56%), seluruhnya pada kelompok subkategori, sehingga diperlukan perbandingan akurasi dan identifikasi kelompok kesalahan keduanya sebagai dasar rekomendasi pemanfaatan AI sebagai alat bantu kodefikasi.

Tujuan: Penelitian ini bertujuan menganalisis perbedaan akurasi kode diagnosis ICD-10 antara metode manual oleh Puskesmas Umbulharjo I dan metode *AI* oleh *Claude*, serta mengidentifikasi kelompok kesalahan kode diagnosis berdasarkan validasi *expert judgment*.

Metode: Penelitian menggunakan pendekatan kuantitatif deskriptif dengan desain *cross sectional*. Sampel berjumlah 398 rekam medis rawat jalan tahun 2025 yang diperoleh melalui *simple random sampling* dari populasi 68.995 rekam medis. Kode manual dibandingkan dengan kode *AI* yang dihasilkan *Claude* menggunakan *prompt*, kemudian keduanya divalidasi terhadap kode *expert judgment* yang sesuai dengan ICD-10 *Volume* 2 Tahun 2010. Uji reliabilitas dilakukan dengan metode *test-retest* pada dua sesi pengujian yang berbeda untuk melihat konsistensi dari *Claude* dalam menghasilkan kode.

Hasil: Berdasarkan pengamatan peneliti, pemberian kode diagnosis tidak sesuai SOP, tidak ada validasi akhir PMIK, dan juga sering *loading* pada SIMPUS. Kode

manual akurat sebanyak 93 rekam medis (23,37%), sementara kode *AI* akurat sebanyak 345 rekam medis (86,68%). Kelompok kesalahan kode manual didominasi oleh kesalahan subkategori sebanyak 291 rekam medis (95,41%) dan kesalahan kategori sebanyak 14 rekam medis (4,59%). Seluruh kesalahan kode *AI* adalah subkategori sebanyak 53 rekam medis (100%), dan konsistensi *Claude* mencapai 356 rekam medis (99,50%), dengan hanya 2 rekam medis yang menghasilkan kode berbeda antar sesi.

Kesimpulan: Terdapat kesenjangan akurasi yang signifikan antara kode manual (23,37%) dan kode *AI* (86,68%). Kesalahan kode manual bersumber dari tidak terlaksananya SOP, absennya validasi akhir PMIK, dan keterbatasan SIMPUS. Kode manual mengalami kesalahan pada kelompok kategori dan sub kategori, sedangkan kode *AI* seluruhnya bersifat subkategori. Konsistensi *Claude* yang tinggi menjadikannya potensial sebagai alat bantu pemberian kode, namun verifikasi akhir oleh PMIK tetap diperlukan.

Kata Kunci: Akurasi Kode Diagnosis, *Artificial Intelligence*, *Claude*, Kelompok Kesalahan Kode