Synthetic Medical Coding Challenge

About Swisscoding Technologies

We're building AI solutions to transform medical coding—the process of converting written medical records into standardized codes used for billing, statistics, and patient care. Currently, medical coding is mostly done manually, which is slow, expensive, and error-prone. Our AI aims to make this process faster, more accurate, and scalable.

The Problem We're Solving

Creating good AI for medical coding requires lots of training data. However, real patient data is heavily protected by privacy laws and regulations. This is why we need realistic synthetic (artificially created) medical data that mimics real patient records without using actual patient information.

Your Challenge: Creating Synthetic Medical Data

Imagine you're working with our AI team on a crucial project. Matteo, our AI Engineer, has identified a major bottleneck in our development pipeline: we need high-quality training data for our osteoarthritis coding models, but privacy regulations make real patient data difficult to use.

"This is where you come in," Matteo explains during your onboarding. "We need someone who can generate realistic synthetic medical data that captures the nuances of knee osteoarthritis cases."

Your challenge is to create a synthetic dataset based on the sample data we've provided that will serve as the foundation for our next-generation medical coding models.

What We Provide You With

- A JSON file (synthetic_medical_cases.json) with 10 example cases of knee osteoarthritis, including:
    - Discharge summary text (what a doctor writes when a patient leaves the hospital)
    - Primary diagnosis code (the main condition - a type of knee osteoarthritis)
    - List of secondary diagnosis codes (other related conditions)
- A CSV file (codes_icd_diagnosis.csv) with descriptions of medical diagnosis codes

What You Need to Create

A synthetic dataset with:

- At least 100 sample cases
- Three columns:
    - Synthetic discharge summary texts (make sure that the notes you generate are diverse e.g. by adjusting text length, structure, patient backgrounds (age, gender, medical history and clinical details and treatment approaches)
    - Associated primary ICD code. Should be one of the following options:
        - M1710: Unilateral primary osteoarthritis, unspecified knee
        - M1711: Unilateral primary osteoarthritis, right knee
        - M1712: Unilateral primary osteoarthritis, left knee
    - Bonus (optional): Include relevant secondary diagnosis codes

As Matteo says: "The strength of our AI models depends entirely on the quality of data they learn from. Your synthetic dataset will be the cornerstone of our medical coding platform."

Tools You Can Use

You may use Google Gemini's models (free tier) to help generate the synthetic dataset.

How We'll Evaluate Your Work

- Your reasoning and approach: Why did you choose to do what? Experiments or evidence in literature (e.g. papers, technical blog posts)
- The quality of your dataset in terms of how diverse and realistic it is
- Code quality: Readability, organization, and maintainability

What to Expect in the Interview

- No formal presentation needed
- Casual but detailed discussion about your methods and thought process
- Focus on understanding how you approached the problem
- Be prepared to show your dataset and code during the interview

Finally: Don’t worry!!! We are a young team that values creativity, collaboration, and a willingness to learn — so have fun and show us what you can do! 😊