**Data Absenteeism in LLM-Generated Synthetic Health Surveys: A Comparative Study across ASEAN Countries**

Huanyu Bao, Dandan Wang, **Edmund W.J. Lee**

**Background**: This study evaluates "data absenteeism"—the absence or misrepresentation of population characteristics and health behaviours in synthetic data—by assessing large language models' (LLMs) ability to generate accurate synthetic health surveys across six ASEAN countries (Singapore, Indonesia, Thailand, Vietnam, Malaysia, Philippines).

**Methods:** We collected large-scale ASEAN survey data (N = 6,174) on the relationship between health technology use (apps and wearables) and self-reported physical and mental well-being outcomes. Synthetic datasets were generated using ChatGPT, Claude, and Gemini. Demographic distributions were compared using two-sample t-tests, and health outcome predictions were assessed with regression analyses.

**Results**: LLM-generated data closely matched real-world data in terms of gender, ethnicity, and education levels. However, LLMs underestimated the influence of social determinants on health outcomes. Survey data showed stronger age effects (e.g., Indonesia BMI: $\beta = .18$, $p < .001$), gender disparities (e.g., Vietnam BMI: $\beta = -.22$, $p < .001$), and ethnic disparities (e.g., Singapore emotional well-being: $\beta = .14$, $p < .001$), while LLM data produced weaker effects on physical and mental health. Health technology impacts were overestimated in LLM data (e.g., Vietnam psychological well-being: $\beta = .43$, $p < .001$ vs. $\beta = .01$ in survey data). Additionally, alignment between datasets was moderate in relatively developed countries (Singapore, Malaysia) but weaker in lower-middle-income (LMIC) countries (Indonesia, Philippines, Vietnam).

**Conclusion:** While LLMs accurately capture basic demographic distributions, synthetic data often underestimate the relationships between social determinants, technology use, and health outcomes, indicating the presence of data absenteeism for LMIC countries.