

## МЕТОДИ ВАЛІДАЦІЇ СИНТЕТИЧНИХ ТАБЛИЧНИХ МЕДИЧНИХ ДАНИХ

Калінін Д. В., Северин В. П.

*Національний технічний університет  
«Харківський політехнічний інститут», м. Харків*

Синтетичні табличні медичні дані відіграють важливу роль у забезпеченні приватності пацієнтів, дозволяючи проводити дослідження без розкриття персональної інформації. Однак, ефективне використання таких даних потребує надійних методів валідації, які гарантують їхню відповідність реальним медичним показникам. Основною задачею валідації є оцінка того, наскільки синтетичні дані зберігають статистичні, структурні та кореляційні характеристики оригінальних наборів.

Для перевірки якості синтетичних табличних даних використовуються різні методи аналізу, які тісно пов'язані з такими аспектами, як статистична, структурна та прикладна відповідності. Так, статистична відповідність забезпечує порівняння розподілу значень, середніх показників та кореляцій між атрибутами, що дозволяє оцінити схожість синтетичних даних із реальними [1]. Структурна відповідність фокусується на збереженні логічних зв'язків між медичними параметрами, зокрема між діагнозами, результатами аналізів та призначеними терапіями, що особливо важливо для клінічних прогнозувальних моделей [2]. Прикладна відповідність визначає, чи можуть синтетичні дані ефективно використовуватися для тренування моделей машинного навчання та бути достатньо корисними у реальних аналітичних задачах [3].

Одним із ключових викликів при генерації синтетичних табличних даних є проблема перенавчання моделі, коли нейронні мережі починають відтворювати фрагменти оригінальних даних, що може призвести до ризику деанонізації. Щоб запобігти цьому, використовуються моделі, які зберігають загальні статистичні властивості без прямого копіювання вихідних даних. Зокрема, варіаційні автокодувальники (VAEs) і генеративні змагальні мережі (GANs) демонструють ефективні результати у синтезі медичних табличних наборів, зберігаючи основні закономірності даних [2].

### **Література:**

1. Apellániz P. A., Jimenez A., Galende B. A. et al. Synthetic Tabular Data Validation: A Divergence-Based Approach // IEEE Access. 2024. Vol. 12.
2. Chen J., Chun D., Patel M. et al. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures // BMC Medical Informatics and Decision Making. 2019. Vol. 19, no. 1.
3. Vallevik V. B., Babic A., Marshall S. E. et al. Can I trust my fake data –A comprehensive quality assessment framework for synthetic tabular data in healthcare // International Journal of Medical Informatics. 2024. Vol. 185, 105413.