Predicting Type 2 Diabetes in Individuals with Severe Mental Illness (SMI) Longitudinal analysis of Electronic Health Records

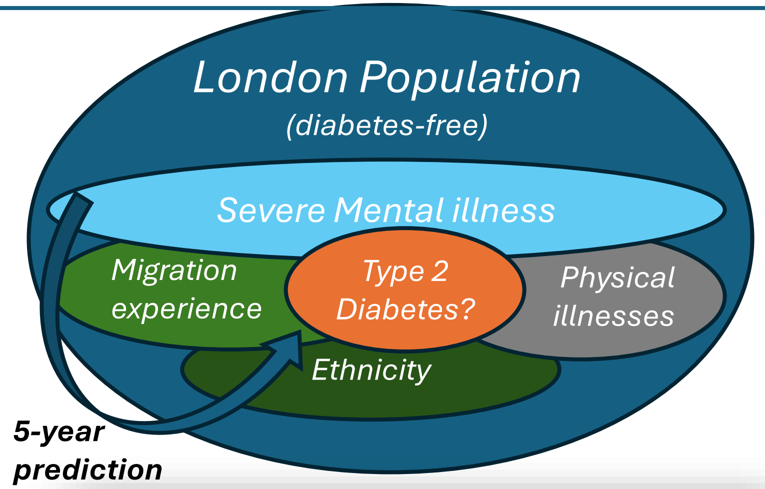
Diana Shamsutdinova, PhD; Daniel Stahl, PhD[1]; Jay Das-Munshi, MD, PhD[2]

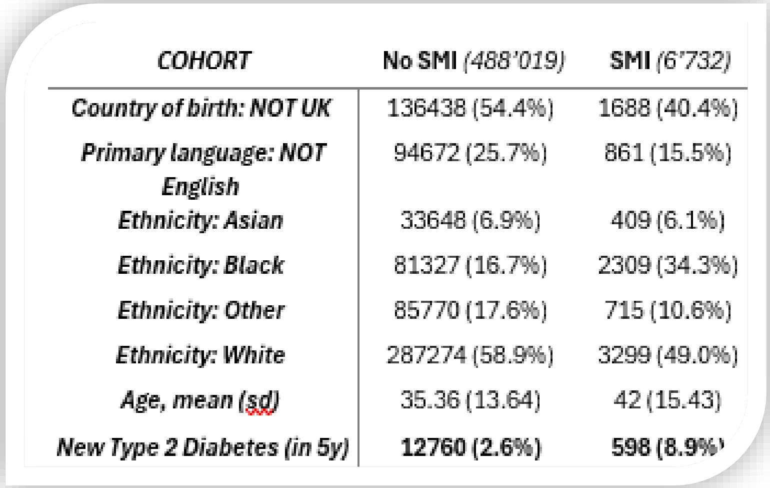
[1] Biostatistics and Health Informatics, IoPPN, King's College London, [2]Department of Public Health, King's College London

NIHR | Maudsley Biomedical Research Centre

Prediction Modelling Group

https://github.com/dianashams/survcompare





COHORT	No SMI *(488'019)*	SMI *(6'732)*
Country of birth: NOT UK	136438 (54.4%)	1688 (40.4%)
Primary language: NOT English	94672 (25.7%)	861 (15.5%)
Ethnicity: Asian	33648 (6.9%)	409 (6.1%)
Ethnicity: Black	81327 (16.7%)	2309 (34.3%)
Ethnicity: Other	85770 (17.6%)	715 (10.6%)
Ethnicity: White	287274 (58.9%)	3299 (49.0%)
Age, mean (sd)	35.36 (13.64)	42 (15.43)
New Type 2 Diabetes (in 5y)	**12760 (2.6%)**	**598 (8.9%)**

Granulating Ethnic Groups:

Primary Language (English/ Other)

Country of birth (UK/Other)

Background: The prevalence of type 2 diabetes mellitus (T2DM) in individuals with severe mental illness (SMI) is 2–3 times higher than in the general population, independent of traditional risk factors. **Over 30% of individuals with SMI over 50 in some ethnic groups may develop T2DM**. Predictive modelling is crucial for early detection in this vulnerable population.

Aims:

1) Develop and validate prediction model for T2DM in SMI;
2) Assess predictive value of social and clinical exposures;
3) Test if simple model is enough or machine learning is needed?

Methods:

Data: Electronical Health Records, 450,000 London residents, 6,700 with SMI

Timeframe: Nov 2012 – Nov 2019 (Latest Pre-COVID).

Predictors: Age, sex, deprivation index, previous hypertension, microvascular comorbidities, depression, ethnicity, primarily language spoken, country of birth.

Algorithms: 1) Cox-PH, 2) Machine Learning Ensemble (Prediction = (1-λ) Cox + λ RandomSurvivalForest)

Results and Methodological insights

- Using EHR data, one can predict T2DM cases reasonably well for people with SMI (AUC-ROC 0.73)
- Machine Learning Ensemble performed only marginally better (AUCROC 0.7280 vs 0.7255)
- Optimal ensemble's share in predictions was 34% (λ = 0.34)

Clinical Insights

- Being born outside of the UK (1st generation migrants) or not reporting English as primary language increases T2DM chances by about 40%, for both SMI and non-SMI groups. For SMI, country of birth was an important prediction at par with ethnicity. This emphasises the critical role of accessible interpretation services within the NHS.

Methodological Challenges and Solutions

Longitudinal Missing Data

- **Multilevel multivariate** imputation: use all time points to inform missing
- **Multiple** imputation: don't analyse as if imputed is same as known

Time-to-event prediction

- Survival methods CoxPH, Survival Random Forests

Non-linearity

- Compare predictions by CoxPH and Random Forest, **'survcompare'** package

Model assessment

- Repeated Nested Cross-Validation
- Discrimination and Calibration

Limitations

- No antipsychotic medication
- Future research to include

Shamsutdinova, D., Das-Munshi, J., Ashworth, M., Roberts, A., & Stahl, D. (2023). Predicting type 2 diabetes prevalence for people with severe mental illness in a multi-ethnic East London population. *International Journal of Medical Informatics*, *172*, 105019.

Shamsutdinova, D., Stamate, D., & Stahl, D. (2025). Balancing accuracy and Interpretability: An R package assessing complex relationships beyond the Cox model and applications to clinical prediction. *International Journal of Medical Informatics*, *194*, 105700.