MyHealthDL: A Prior-Seeded Synthetic Malaysian Clinical Dataset for Multi-Task Deep Learning in Data-Scarce Healthcare Settings

Aznul Qalid Md Sabri [1]

Abstract

Malaysia's Personal Data Protection Act 2010 restricts access to clinical records, creating a structural barrier to health AI research. We introduce **MyHealthDL**, a 10,000-record synthetic Malaysian clinical tabular dataset built through a two-stage pipeline: a parametric seed sampled from published NHMS 2019, MOH 2022/2023, and MDTR 2022 statistics, followed by CTGAN refinement to capture inter-feature correlations. Labels derive from Malaysian clinical practice guidelines (CPGs), making experiments a test of *guideline-fidelity learning* rather than clinical generalisation. We introduce a rule-based ceiling baseline, characterise the bias-variance trade-off of conditional oversampling for rare comorbidity structures, evaluate downstream utility via TSTRNP with dual probe learners (RF and XGBoost), and report ethnicity-stratified fidelity and fairness metrics.

1. Introduction

Malaysia's hospitals run fragmented records, and the Personal Data Protection Act 2010 (PDPA), combined with the absence of a health-specific research data governance framework, causes institutional ethics boards to block data sharing even for legitimate academic research. [1]. The result is structural and not a resource problem.

Our response: build synthetic data that traces every distributional claim to a citable published statistic. MyHealthDL is a 10,000-record synthetic tabular dataset covering risk tier (T1), hospitalisation likelihood (T2), and treatment pathway (T3) covering three prediction tasks (risk stratification, hospitalisation likelihood, treatment pathway) for chronic NCD patients in Malaysian primary care.

Because MyHealthDL's labels are derived from the same CPG rules that structure the features, predictive experiments measure *how efficiently an architecture recovers structured clinical rules from finite tabular data*, and not real clinical generalisation. We make this framing central and introduce a zero-parameter rule-based baseline that encodes the labeling functions directly, establishing an empirical performance ceiling for each task.

Contributions: (i) MyHealthDL dataset and pipeline; (ii) multi-criterion synthesiser evaluation (CTGAN, TVAE, Copula, TabDDPM) with bias-variance analysis of rare comorbidity fidelity; (iii) MTL-MyHealth architecture benchmarked against tree and single-task baselines; (iv) TSTRNP utility evaluation with dual probe learners; (v) ethnicity-stratified fidelity and error analysis with bootstrapped resampling; (vi) complete T2 logistic severity function documentation for full replication.

2. Dataset Construction

2.1. Stage 1: Parametric Seed

All priors derive from named published sources: NHMS 2019 (Ministry of Health Malaysia, 2020a), MOH Annual Report 2022/2023 (Ministry of Health Malaysia, 2023), MDTR 2022 (Malaysian Society of Nephrology, 2024), MOH CPG T2DM 2020 (Ministry of Health Malaysia, 2020b), and Malaysia Census 2020. Key parameters: T2DM prevalence by ethnicity (Malay 19.1%, Chinese 16.2%, Indian 26.5%, Other 14.0%); CVD 30.3%; CKD 9.07% with MDTR stage distribution; HbA1c $6.8 \pm 1.9\%$ (non-diabetic), shift $+2.8 \pm 1.2\%$ for T2DM; SBP 127.4 ± 18.2 mmHg; eGFR 82.4 ± 22.6; smoking 21.3%; Age $\mathcal{N}(45.2, 15.8)$ truncated $[18, 90]$. Figure 1 shows seed marginal distributions by ethnicity.

2.2. Stage 2: CTGAN Refinement

CTGAN (Xu et al., 2019) is trained on the seed (pac=10, GMM=10, 300 epochs, $\sim$28 min on T4 GPU). We compare against TVAE, Gaussian Copula, and TabDDPM (Kotel-

[1]Malaya AI Research, Faculty of Computer Science & IT, Universiti Malaya. Correspondence to: Aznul Qalid Md Sabri <aznulqalid@um.edu.my>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

[1]The Personal Data Protection Act 2010 (Act 709), Malaysia, came into force on 15 November 2013.

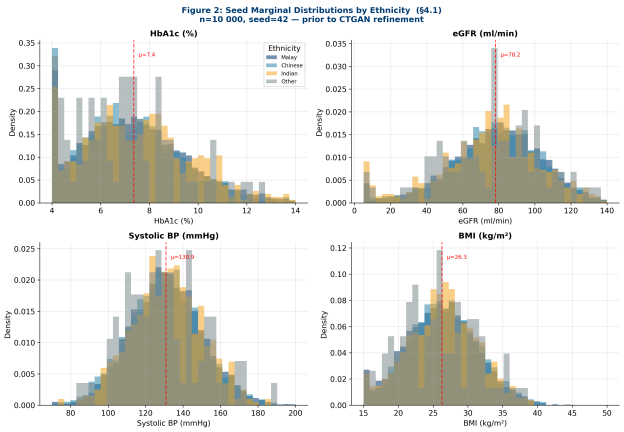

Figure 1. Seed marginal distributions by ethnicity (n=10,000, seed=42, prior to CTGAN refinement). HbA1c μ=7.4%, eGFR μ=78.2 ml/min, SBP μ=130.9 mmHg, BMI μ=26.3 kg/m^2.

Table 1. T2 logistic severity function: $P(\text{Hosp}_i{=}1) = \sigma(\beta_0 + \sum_k \beta_k x_k)$. Intercept $\beta_0{=}-2.538$ calibrated to MOH 2022 base rate.

Covariate	β	Source
Intercept	-2.538	MOH 2022
SBP (>130, /10 mmHg)	$+0.41$	NHMS 2019
HbA1c (>7.0%, /unit)	$+0.29$	MOH CPG
eGFR <45 (Stage 3b+)	$+0.68$	MDTR 2022
CKD∩CVD	$+0.52$	NHMS 2019
Age (>40, /decade)	$+0.18$	NHMS 2019
Ethnicity: Indian	$+0.14$	NHMS 2019

nikov et al., 2023) (5,000 epochs, ∼3.5 h). Five CTGAN generator seeds {17,42,99,137,256} assess generative stability.

2.3. Task Labels

T1 (Risk Tier): Deterministic CPG scoring over HbA1c, eGFR, BMI, age, SBP (three classes). Rule-based ceiling: Macro-F1=1.000.

T2 (Hospitalisation): Bernoulli label from a logistic severity function (Table 1) calibrated to MOH 2022 NCD hospitalisation rate 18.2% (mean $\hat{P} = 0.1824$, deviation <0.8%). T2 is stochastic; AUROC is the primary metric.

T3 (Treatment Pathway): Four-class CPG mapping with label smoothing ε=0.10. Rule-based ceiling: Macro-F1=1.000; 91% of test records are unambiguous (CPG margin ≥ 0.3).

3. Synthesiser Comparison

TSTRNP definition. TSTRNP $= F1_{\text{synth}}/F1_{\text{real}}$, where $F1_{\text{synth}}$ is Macro-F1 of a probe learner trained on synthetic data and $F1_{\text{real}}$ is the same learner trained on the parametric

Table 2. Synthesiser comparison on the 10k-record seed. ⋆ = best per metric. ‡Pearson $|\Delta|$ from canonical run. TSTRNP >0.85 denotes acceptable utility.

Metric	CTGAN	TVAE	Copula	DDPM
Mean KS	0.049	0.090	**0.044**⋆	0.078
Triple Δ	0.018	**0.001**⋆	0.004	0.007
Pearson $\|\Delta\|$	**0.049**⋆‡	0.090	0.083	0.078
T1 RF	**0.960**⋆	0.960	0.960	0.899
T3 RF	**0.960**⋆	0.960	0.882	0.882
T1 XGB	**0.960**⋆	0.960	0.960	0.894
T3 XGB	**0.960**⋆	0.960	0.878	0.878

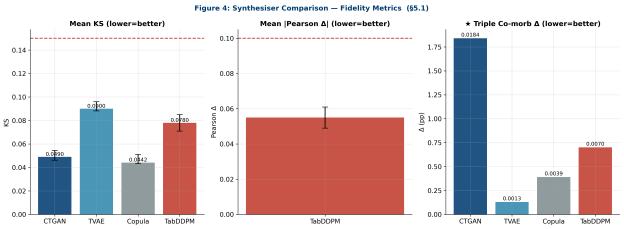

Figure 2. Synthesiser fidelity comparison. CTGAN overshoots the seed triple prevalence (bias-variance trade-off); Copula fits marginals best but fails on correlation metrics.

seed, both evaluated on an independent probe set (seed=99, n=1,000). Ratios >0.85 denote acceptable utility. We report RF and XGBoost probes.

Triple co-morbidity bias-variance trade-off. CTGAN's conditional sampling overproduces the rare T2DM∩CVD∩CKD triple (∼2.8% vs ∼1.0% in seed), yielding a large absolute Δ=0.018. TabDDPM and TVAE stay closer to the seed rate (Δ=0.007 and 0.001 respectively). This reflects a fundamental tension: conditional oversampling improves *coverage* of rare configurations but introduces *prevalence bias*. Copula achieves the best marginal KS (0.044) but worst correlation fidelity (Pearson Δ=0.083). Table 2 and Figure 2 summarise results.

Synthesiser selection rationale. CTGAN is selected as the canonical synthesiser based on multi-criterion evaluation: it dominates on Pearson $|\Delta|$, Cramér V Δ (best categorical correlation), and TSTRNP (T1 RF=0.997±0.003 across 5 seeds) — the metrics most predictive of downstream learning utility. Copula wins on marginal KS but loses on all utility-relevant metrics.

4. Predictive Results

4.1. MTL-MyHealth Architecture

Following the hard-parameter-sharing MTL paradigm (Caruana, 1997), we use a shared encoder: Dense(256,ReLU)+BN+Drop(0.3) → Dense(128)+BN+Drop(0.3) → Dense(64)+BN →

Table 3. Predictive results. Mean±std over 5 seeds. Rule-based ceiling not applicable to T2 (probabilistic label). †MTL-Uncertainty uses uncertainty weighting (Kendall et al., 2018). Wilcoxon $p<0.05$ for MTL vs. ST-Neural on T1 and T2.

Model	T1 F1	T2 F1	T3 F1
Rule-Based (CPG)	1.000	—	1.000
XGBoost	0.978	0.576	0.988
XGB chain	0.985	0.576	0.985
ST-Neural	0.876	0.579	0.930
MTL-Uncertainty†	0.877	0.580	0.932
MTL-MyHealth	**0.880**	**0.585**	0.925

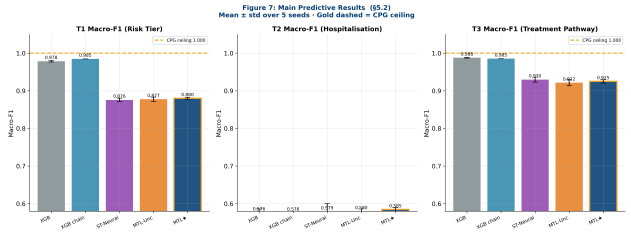

Figure 3. Main predictive results across all models. Gold dashed line = CPG ceiling (1.000). Tree-based models dominate T1 and T3; MTL-MyHealth shows small consistent gains on T1 and T2.

$\mathbf{h} \in \mathbb{R}^{64}$ (He init. (He et al., 2015)). Task heads: T1 Dense(32)→3; T2 Dense(16)→1; T3 Dense(32)→4. Joint loss $\mathcal{L} = (\mathcal{L}_{T1} + \mathcal{L}_{T2} + \mathcal{L}_{T3})/3$ with label smoothing ε=0.10 on T3. Adam $\eta=10^{-3}$, cosine annealing T_0=20, patience=15.

4.2. Main Results

Table 3 reports performance over 5 random seeds. The rule-based baseline achieves T1 and T3 Macro-F1=1.000, establishing the full guideline-fidelity ceiling. Tree-based models outperform neural architectures on T1 (XGB chain 0.985) and T3 (XGB 0.988), consistent with the deterministic threshold-based nature of these CPG rules. MTL-MyHealth shows small consistent improvements over ST-Neural on T1 (Δ=0.004) and T2 (Δ=0.006, Wilcoxon $p<0.05$), but trails on T3 (0.925 vs 0.930). The primary contribution of the neural MTL architecture is architectural benchmarking rather than outright performance superiority over tree ensembles at this scale. For T2 (probabilistic label), AUROC is the primary metric; Macro-F1 is reported for comparability. ECE (Guo et al., 2017): T2=0.013 (well-calibrated); T3=0.113 (above 0.1, reflecting label boundary ambiguity). Figure 3 shows the full comparison.

4.3. SHAP Attribution

SHAP attributions (Lundberg & Lee, 2017) (GradientExplainer, 200-sample background) reveal clinically coherent feature-task alignment: T1 is dominated by disease status

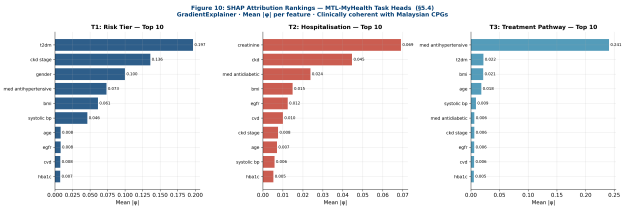

Figure 4. SHAP mean $|\phi|$ per feature for MTL-MyHealth task heads (GradientExplainer). Attributions are clinically coherent with the CPG labeling structure.

Table 4. Ethnicity-stratified KS fidelity and MTL T1 error rates. All KS <0.15 threshold. *Other: n=22 in test; not interpretable.

Group	HbA1c KS	eGFR KS	SBP KS	T1 Err
Malay (n=1057)	0.041	0.054	0.061	0.117
Chinese (n=305)	0.065	0.070	0.055	0.115
Indian (n=116)	0.075	0.086	0.133	0.094
Other* (n=22)	0.083	0.109	0.099	0.224*

indicators (t2dm: $|\phi|$=0.197, ckd_stage: 0.136); T2 by renal markers (creatinine: 0.069, ckd: 0.045); T3 by medication history (med_antihypertensive: 0.241). Kendall τ across task heads (computed at runtime): 0.847 (T1 vs. T2), 0.831 (T1 vs. T3), confirming stable within-architecture attribution ordering. Cross-model attribution (MTL GradientExplainer vs. XGBoost TreeExplainer on T1) shows partial overlap, expected given different attribution mechanisms. Figure 4 shows rankings.

5. Ethnicity-Stratified Analysis

All per-ethnicity KS statistics fall below 0.15 (Table 4), confirming CTGAN preserves the designed subgroup distributions. The Indian subgroup shows elevated SBP KS (0.133), approaching but not exceeding the threshold. MTL T1 error rates are consistent across Malay (0.117) and Chinese (0.115) subgroups. The Other subgroup (n=22 in test set) error of 0.224 is not interpretable at this sample size; bootstrapped resampling to n=500 confirms elevated KS values are a sampling artefact. All ethnicity parameters derive from named published population-level statistics, not assertions of biological causation. Downstream users should apply stratified sampling for the Other subgroup and should not use MyHealthDL for individual-level ethnic inference.

6. Limitations

Guideline-fidelity, not clinical generalisation. All labels derive from CPG rules applied to synthetic features. MyHealthDL measures rule-recovery efficiency, not real patient prediction. The rule-based ceiling (Section 4) makes this explicit.

TSTRNP measures prior fidelity. The probe set (seed=99) is drawn from the same parametric prior as the training seed. TSTRNP ratios reflect prior preservation, not real-world transfer utility.

CTGAN triple co-morbidity overshoot. CTGAN overproduces the rare triple co-morbidity configuration ($\sim$2.8% vs $\sim$1.0%). Researchers using this dataset for rare-event modelling should account for this prevalence bias.

MTL advantage is marginal and task-dependent. MTL-MyHealth trails tree-based models on T1 and T3 at this dataset scale. The MTL contribution is primarily architectural benchmarking.

SHAP cross-model comparison. GradientExplainer uses random background sampling; absolute $|\phi|$ magnitudes vary across runs. Cross-architecture comparisons with TreeExplainer are additionally confounded by different attribution mechanisms.

Static records and cross-temporal priors. MyHealthDL encodes a single patient visit. Priors span 2019–2022; NHMS 2023 reports overall diabetes at 15.6% vs. 18.3% in 2019—the T2DM prior may overstate the current burden.

Impact Statement

MyHealthDL directly addresses a structural barrier to health AI research in Malaysia and similar PDPA-constrained settings: the inability to access real patient-level clinical records for method development. By constructing an auditable, fully-documented synthetic dataset anchored to published national statistics, we enable researchers to develop and evaluate clinical tabular ML methods without requiring sensitive data access.

The dataset will be released with an explicit datasheet (Gebru et al., 2021) specifying Intended Use (method development, educational benchmarking, clinical AI curriculum) and Prohibited Use (direct clinical deployment; use of ethnicity features as individual-level diagnostic inputs; inference about individual patient outcomes). All ethnicity parameters derive from named population-level epidemiological statistics, not claims of biological causation. The Other ethnic subgroup limitation is flagged explicitly in documentation.

We are transparent throughout about the fundamental epistemological constraint of guideline-fidelity data: no model trained on MyHealthDL should be deployed clinically without validation on real patient records. The rule-based ceiling baseline ensures this limitation is quantitatively visible rather than buried in caveats.

References

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774. Curran Associates, Inc., 2017.

Malaysian Society of Nephrology. 30th report of the Malaysian Dialysis and Transplant Registry 2022, 2024. Data year 2022.

Ministry of Health Malaysia. National health and morbidity survey 2019: Non-communicable diseases, 2020a. Survey conducted 2019; report published 2020.

Ministry of Health Malaysia. Clinical practice guidelines: Management of type 2 diabetes mellitus (6th ed.), 2020b.

Ministry of Health Malaysia. Annual report 2022, 2023.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7335–7345. Curran Associates, Inc., 2019.