Supplementary Material

A Text-to-tabular Approach to Generate Synthetic Patient Data using LLMs

Fig. 4. Prompt used to generate a synthetic Alzheimer's disease population from ADNI.

Prior knowledge

Give an example table of 10 rows from ADNI data set. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is an observational study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD).
Only consider patients with Alzheimer's Disease diagnosis.

Instructions

The table must have one row by patient, no missing values and include all the following columns:

PTID: patient unique identifier, integer
CDRSB: CDR-SB | float | |
PTGENDER: Sex | integer | {0: male, 1: female} |
WholeBrain_bl: UCSF WholeBrain, volume of the whole brain measured by MRI, with the analysis conducted at the University of California, San Francisco (UCSF) | integer | |
ADAS11: Alzheimer's Disease Assessment Scale, 11-item version | float | |
ICV_bl: UCSF ICV, intracranial volume measured by MRI, with the analysis conducted at the University of California, San Francisco (UCSF) | integer | |
AGE: Age in years | float | |
PTEDUCAT: Years of education | integer | |
APOE4: APOE e4 carrier status | integer | |
Ventricles_bl: UCSF Ventricles, volume of the brain's ventricles measured by MRI, with the analysis conducted at the University of California, San Francisco (UCSF) | integer | |
MMSE: Mini-Mental State Examination | float | |

Return the table as a dictionary in JSON format with keys as index. JSON format strictly requires double quotes for strings. The column names need to be the same than those provided.
Only return the dictionary, do not repeat the question, introduce your answer or comment on it. Do not truncate the table, provide all the rows.

Context

Here is an example for one patient with the associated key of the dictionary to output:

{'0': {'PATNO': '022_S_0004', 'AGE': 74, 'PTGENDER': 1, 'PTEDUCAT': 15.5, 'APOE4': 1, 'CDRSB': 4.3, 'ADAS11': 18.6, 'MMSE': 23.3, 'Ventricles_bl': 39300, 'WholeBrain_bl': 1066000, 'ICV_bl': 1500000}}

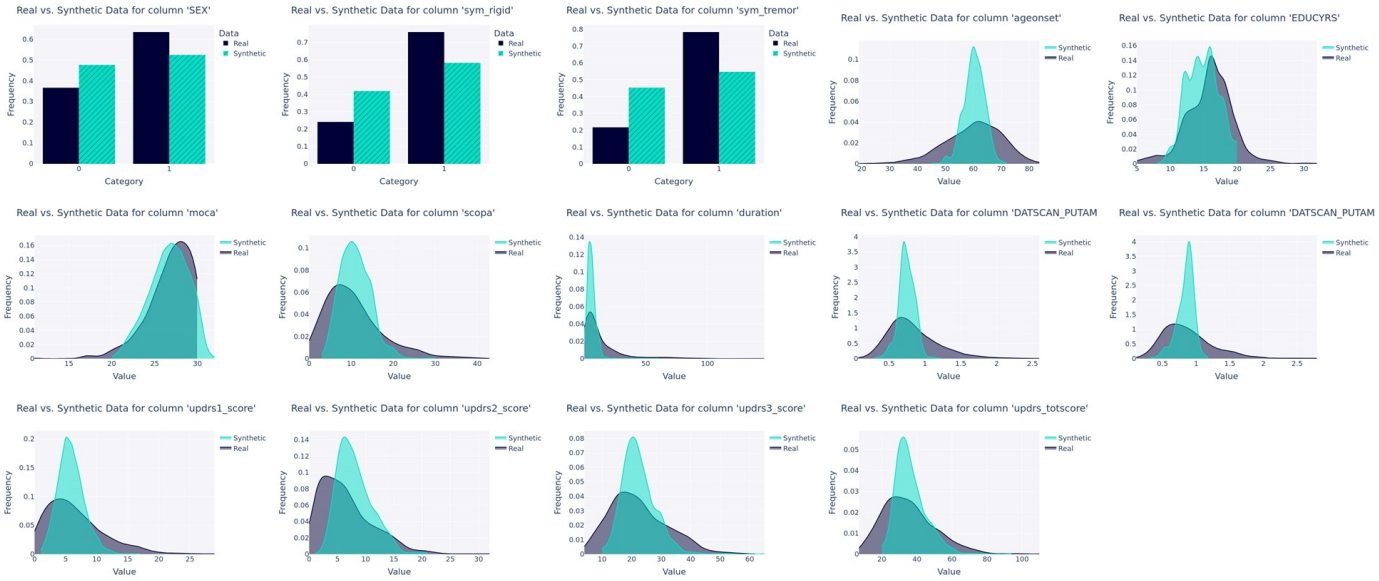

Fig. 5. Univariate distribution plots comparing text-to-tabular generated synthetic data with real data from PPMI. The synthetic data consists of 1,000 patients generated from one of our evaluation framework experiments. The real data encompasses the entire #PPMI2024 dataset to ensure a comprehensive comparison. Of note, here the plots are kernel density estimate plots.

Fig. 6. Univariate distribution plots comparing text-to-tabular generated synthetic data with real data from ADNI. The synthetic data consists of 1,000 patients generated from one of our evaluation framework experiments. The real data encompasses the entire #ADNI dataset to ensure a comprehensive comparison. Of note, here the plots are kernel density estimate plots.

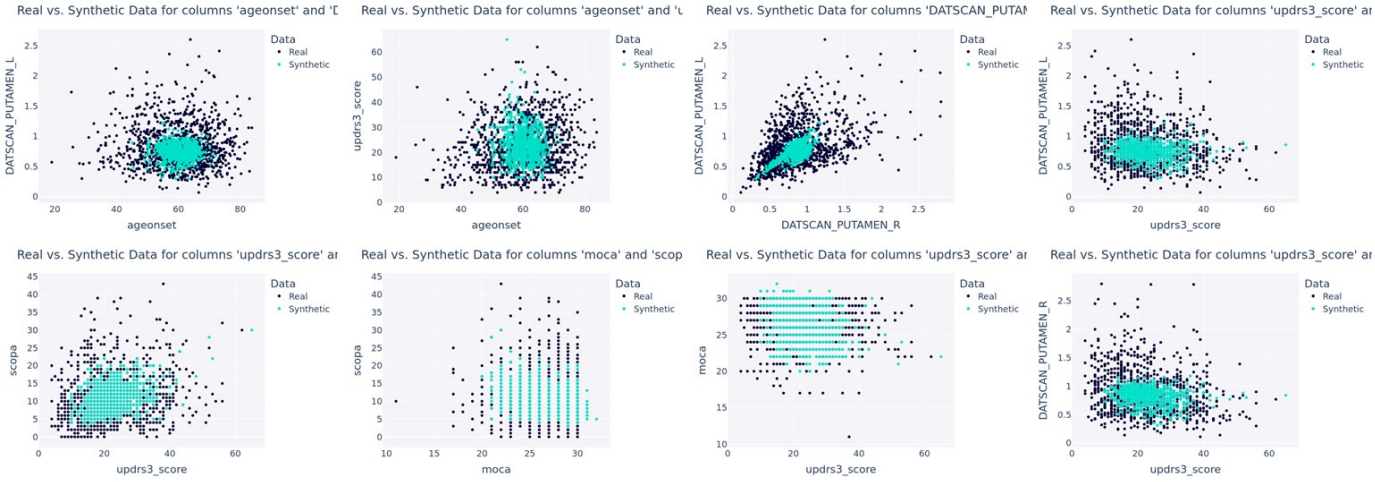

Fig. 7. Joint bivariate distribution plots comparing text-to-tabular generated synthetic data with real data from PPMI. The synthetic data consists of 1,000 patients generated from one of our evaluation framework experiments. The real data includes the entire #PPMI2024 dataset to ensure a comprehensive comparison.

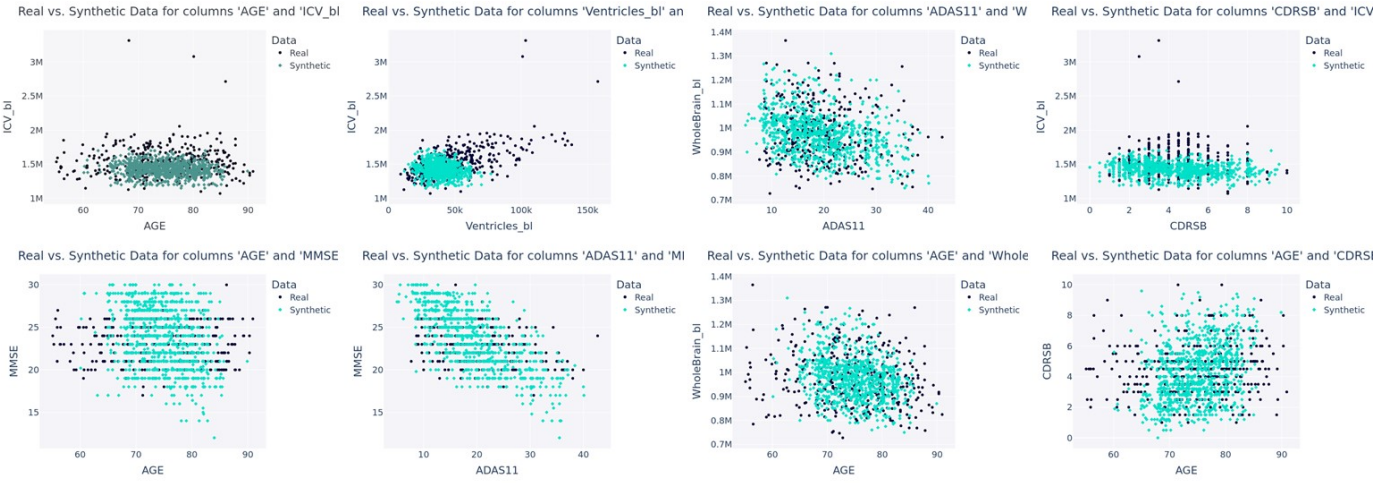

Fig. 8. Joint bivariate distribution plots comparing text-to-tabular generated synthetic data with real data from ADNI. The synthetic data consists of 1,000 patients generated from one of our evaluation framework experiments. The real data includes the entire #ADNI dataset to ensure a comprehensive comparison.

TABLE VII
DESCRIPTION OF A SUBSET OF VARIABLES FROM PPMI CLINICAL DATABASE (CURATED DATACUT V.20240129)

Name	Description	Group	Category	Type	Mapping
moca	MOCA Score (adjusted for education)	Clinical assessment	continuous	int	
DATSCAN_PUTAMEN_L	Striatal Binding Ratio of the Left Putamen Small Volume	Imaging	continuous	float	
DATSCAN_PUTAMEN_R	Striatal Binding Ratio of the Right Putamen Small Volume	Imaging	continuous	float	
updrs_totscore	MDS-UPDRS Total Score (OFF)	Clinical assessment	continuous	int	
updrs1_score	MDS-UPDRS Part I: Non-Motor Aspects of Experiences of Daily Living (OFF)	Clinical assessment	continuous	int	
updrs2_score	MDS-UPDRS Part II: Motor Aspects of Experiences of Daily Living (OFF)	Clinical assessment	continuous	int	
updrs3_score	MDS-UPDRS Part III: Motor Examination (OFF)	Clinical assessment	continuous	int	
duration	Duration of Parkinson disease from Diagnosis to Visit Date	Clinical assessment	continuous	float	
EDUCYRS	Years of Education	Demographic	continuous	int	
SEX	Gender	Demographic	binary	int	{1: male, 0: female}
ageonset	Age at onset of Parkinson disease diagnosis (Years)	Demographic	continuous	float	
sym_tremor	Initial symptom (at diagnosis) - Resting Tremor	Baseline symptom	binary	int	{0: absence, 1: presence}
scopa	SCOPA-AUT Total Score	Clinical assessment	continuous	int	
sym_rigid	Initial symptom (at diagnosis) - Rigidity	Baseline symptom	binary	int	{0: absence, 1: presence}

TABLE VIII
DESCRIPTION OF A SUBSET OF VARIABLES FROM THE ADNI DATASET

Name	Description	Group	Category	Type	Mapping
AGE	Age in years	Baseline characteristic	Categorical	float	
PTGENDER	Sex	Demographic	Binary	int	{0: male, 1: female}
PTEDUCAT	Years of education	Demographic	Continuous	int	
APOE4	APOE e4 carrier status	Clinical assessment	Continuous	int	
CDRSB	CDR-SB	Clinical assessment	Continuous	float	
ADAS11	Alzheimer's Disease Assessment Scale, 11-item version	Clinical assessment	Continuous	float	
MMSE	Mini-Mental State Examination	Clinical assessment	Continuous	float	
Ventricles_bl	UCSF Ventricles, volume of the brain's ventricles measured by MRI, with the analysis conducted at the University of California, San Francisco (UCSF)	Imaging	Continuous	int	
WholeBrain_bl	UCSF WholeBrain, volume of the whole brain measured by MRI, with the analysis conducted at the University of California, San Francisco (UCSF)	Imaging	Continuous	int	
ICV_bl	UCSF ICV, intracranial volume measured by MRI, with the analysis conducted at the University of California, San Francisco (UCSF)	Imaging	Continuous	int	

TABLE IX

BENCHMARK RESULTS OF SDG MODELS ON #ADNI IN TERMS OF EVALUATION OF FIDELITY, PRIVACY PRESERVATION, AND UTILITY (MEAN AND STANDARD DEVIATION). FOR EACH METRIC, THE BEST RESULTS ARE REPORTED IN BOLD.

	Ours		**Baselines**					
Model	**GPT-4**		**Copula**		**CTGAN**		**TVAE**	
Metric	D_{Train}	D_{Test}	D_{Train}	D_{Test}	D_{Train}	D_{Test}	D_{Train}	D_{Test}
Fidelity								
Column Shapes	0.824±0.004	0.812±0.006	**0.923±0.005**	0.904±0.015	0.784±0.033	0.773±0.035	0.863±0.016	0.845±0.021
Column Pair Trends	0.785±0.011	0.768±0.014	**0.903±0.007**	0.862±0.013	0.819±0.014	0.806±0.012	0.847±0.013	0.818±0.017
KSComplement	0.795±0.007	0.785±0.010	**0.915±0.004**	0.897±0.016	0.742±0.041	0.735±0.044	0.866±0.016	0.847±0.015
TVComplement	0.937±0.013	0.919±0.016	**0.954±0.010**	0.932±0.023	0.951±0.016	0.925±0.029	0.849±0.047	0.834±0.062
CorrelationSimilarity	0.877±0.011	0.869±0.013	**0.983±0.001**	0.954±0.011	0.909±0.009	0.904±0.010	0.957±0.004	0.942±0.006
ContingencySimilarity	0.912±0.016	0.880±0.023	**0.913±0.015**	0.881±0.043	0.897±0.033	0.845±0.046	0.743±0.079	0.705±0.101
WD (↓)	16,037±805	18,282±1,192	**6,091±264**	6,147±472	21,836±6,788	22,094±7,445	5,701±1,676	7,471±2,145
JSD (↓)	0.029±0.001	0.031±0.001	0.032±0.002	0.033±0.001	0.050±0.003	0.051±0.004	**0.032±0.001**	0.033±0.001
LogisticDetection	0.500±0.029	0.543±0.058	**0.966±0.026**	0.972±0.034	0.305±0.132	0.315±0.146	0.662±0.074	0.669±0.105
Privacy								
NewRowSynthesis	1±0	1±0	1±0	1±0	1±0	1±0	1±0	1±0
DCR (↓)	0.967±0.030	0.995±0.040	**0.973±0.014**	0.952±0.046	1.064±0.048	1.079±0.053	0.970±0.047	1.035±0.070
NNDR (↓)	0.714±0.027	0.727±0.030	**0.674±0.023**	0.672±0.027	0.690±0.023	0.691±0.041	0.718±0.029	0.736±0.034
CategoricalCAP (↑)	0.860±0.008	0.850±0.014	**0.898±0.005**	0.894±0.004	0.907±0.012	0.905±0.018	0.883±0.012	0.891±0.023
Utility								
TSTR (F1 score)	-	0.904±0.082	-	**0.947±0.066**	-	0.827±0.124	-	0.924±0.087
TATR (F1 score)	-	0.904±0.082	-	**0.947±0.066**	-	0.834±0.125	-	0.928±0.083