KANTAR

Addressing Sample Size Challenges in Linked Data Through Data Fusion using Artificial Neural Networks

Srikesh Arunajadai*, Lulu Lee, Tom Haskell

Kantar Inc. Three World Trade Center, 175 Greenwich St 35th Floor, New York, NY 10007

* srikesh.arunajadai@kantar.com

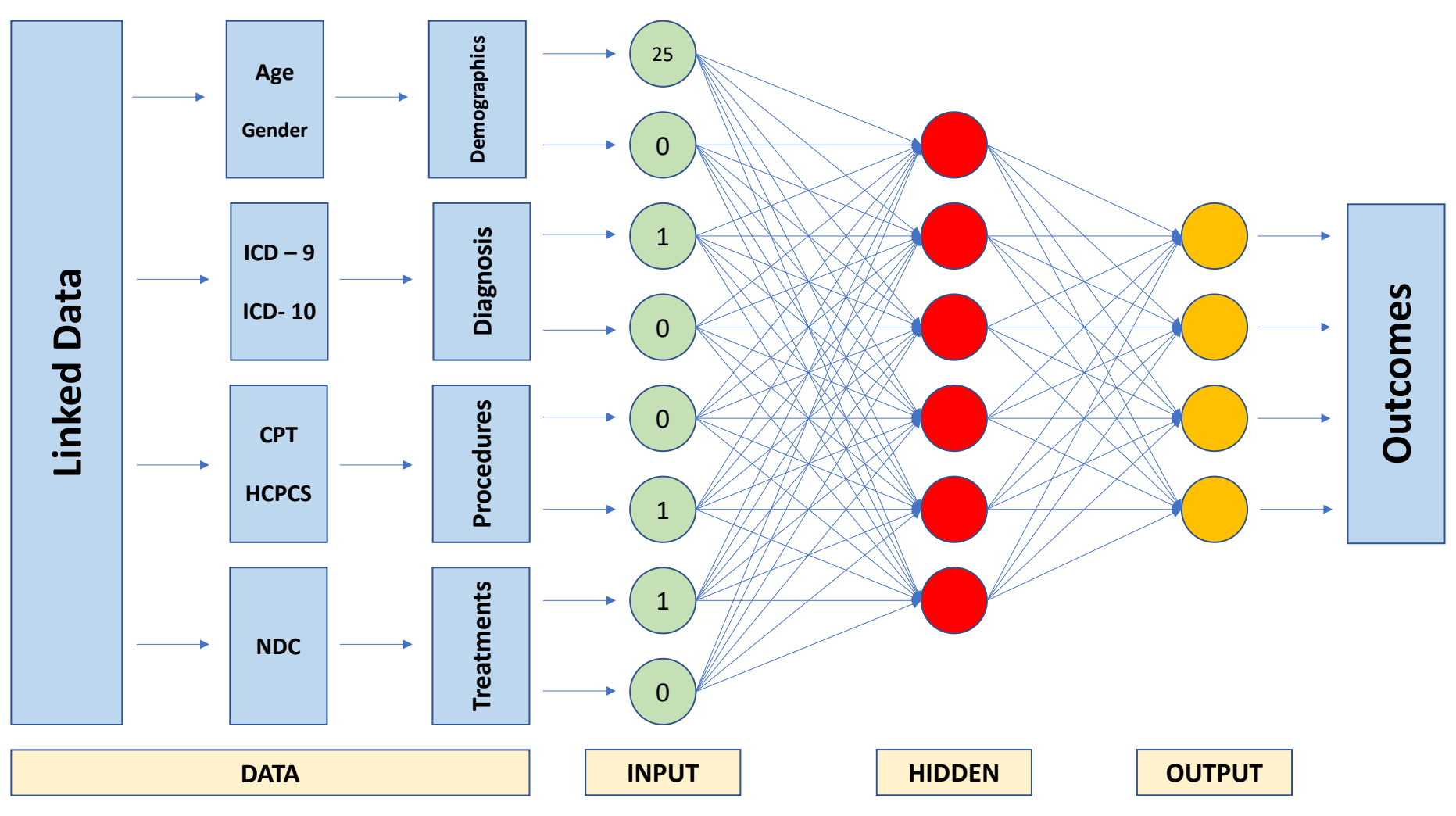
Introduction

- Linking secondary data with patient-reported data at the patient-level brings together a comprehensive view of the patient but sample sizes can be a challenge.
- Data fusion is a special case of data integration to generate a synthetic data set by combining two data sets that have disjoint records and some distinct variables.
- This study demonstrates the fusion of Patient Reported Outcomes (PROs) in surveys with clinical data in claims enabling the study of associations between quality of life and disease-treatment interactions at scale especially for rare diseases.
- The objective is to fuse/impute the PROs from the National Health and Wellness (NHWS) survey (donor) to the Komodo claims data (recipient).

Methods

1. The NHWS survey data are collected annually from nearly 75,000 - 95,000 respondents (adults aged 18 or older) in the US through a self-administered, internet-based survey which provides a unique look into the healthcare market from the viewpoint of the consumer.
2. The Komodo healthcare claims data is an expansive data set of medical and pharmacy claims (>65 billion clinical/prescription encounters) that come from a variety of sources within the United States (US) including hospital networks, physician networks, claims clearinghouses, pharmacies, and health insurers.
3. Variables to fuse: PROs - SF-36v2 (MCS, PCS), SF-6D health utilities index, and EQ-5D-5L
4. Independent variables: Age, gender, Diagnosis (ICD), Procedures(CPT/HCPCS), and Treatment (NDC)
5. All chronic conditions are considered. Acute conditions, procedures, and treatments within a year prior to the survey date are considered.
6. An Artificial Neural Network (ANN) model is fitted as schematically described on the right.
7. The predicted PROs from the recipient data set is matched with the PROs in the donor data set using random distance hot deck matching. The final matched value is the fused PRO for the recipient data
8. Multiple fused data sets are generated by a bootstrap based multiple imputation procedure.
9. The multiply-imputed fused data sets are analysed using procedures for such data sets.
10. There were a total of 104,132 patients in the linked data sample. The patients were divided into training set (N = 78,099, 80%), validation set (N = 20,826, 20%) and a test set (N=5,207, 5%).

Model



Analysis

1. We compare the performance of the fused data on the test data of $N = 5207$ across univariate, bivariate, and correlation analysis.
2. For each PRO, we provide the minimum sample size required N_{min} to make valid inferences calculated a priori based on the matching noise in the training data.
3. For the univariate analysis, we compare means across non-disease specific, type-2 diabetes, and Myasthenia Gravis (a rare disease) cohorts.
4. For each of the comparisons we provide the P-value associated with hypothesis test of no difference between the observed and fused data estimates.
5. For the correlation analysis we provide the 95% lower (LL) and upper (UL) confidence limits.
6. We compare the difference between the observed and fused estimates with Minimal Clinically Important Difference (MCID). The differences between PROs are meaningful only if they are greater than the MCID. Ideally we would want the differences between observed and fused estimates to be less than MCID.

Non-disease Specific

N = 5207			Observed		Fused		Difference		
PRO	MCID	N_{min}	Mean	SE	Mean	SE	Mean	SE	P-value
MCS	3.000	232	48.11	0.158	48.31	0.422	-0.202	0.442	0.656
PCS	2.000	386	50.18	0.132	50.17	0.251	0.008	0.278	0.977
EQ5D	0.180	13	0.82	0.002	0.82	0.004	0.003	0.005	0.448
SF6D	0.033	292	0.73	0.002	0.73	0.004	-0.003	0.004	0.494

Type-2 Diabetes

N = 883			Observed		Fused		Difference		
PRO	MCID	N_{min}	Mean	SE	Mean	SE	Mean	SE	P-value
MCS	3.000	225	49.73	0.369	50.00	0.928	-0.270	0.976	0.786
PCS	2.000	487	45.78	0.351	46.13	0.666	-0.352	0.741	0.636
EQ5D	0.180	14	0.79	0.005	0.79	0.010	0.002	0.011	0.860
SF6D	0.033	299	0.71	0.005	0.72	0.009	-0.009	0.010	0.367

Myasthenia Gravis

N = 100			Observed		Fused		Difference		
PRO	MCID	N_{min}	Mean	SE	Mean	SE	Mean	SE	P-value
MCS	3.000	150	48.09	1.172	49.64	1.612	-1.543	1.923	0.432
PCS	2.000	370	42.81	1.132	42.76	1.675	0.042	1.948	0.983
EQ5D	0.180	13	0.76	0.017	0.74	0.024	0.021	0.029	0.471
SF6D	0.033	194	0.68	0.015	0.69	0.016	-0.007	0.020	0.739

Type-2 Diabetes: By Age

			Observed		Fused		Difference		
PRO	MCID	N_{min}	Mean	SE	Mean	SE	Mean	SE	P-value
18 - 44 (N = 83)									
MCS	3.000	225	43.69	1.205	42.32	3.148	1.366	3.331	0.690
PCS	2.000	487	47.60	1.128	48.51	2.066	-0.911	2.308	0.695
EQ5D	0.180	14	0.80	0.019	0.76	0.041	0.035	0.044	0.438
SF6D	0.033	299	0.67	0.016	0.67	0.030	0.000	0.033	1.000
45 - 64 (N = 366)									
MCS	3.000	225	47.67	0.610	47.73	1.538	-0.058	1.634	0.972
PCS	2.000	487	45.89	0.562	45.93	1.061	-0.041	1.176	0.972
EQ5D	0.180	14	0.77	0.009	0.77	0.018	0.000	0.019	0.987
SF6D	0.033	299	0.70	0.007	0.70	0.015	0.001	0.016	0.955
65 - 79 (N = 381)									
MCS	3.000	225	52.63	0.489	53.44	1.248	-0.810	1.330	0.550
PCS	2.000	487	45.20	0.522	45.93	1.152	-0.721	1.256	0.570
EQ5D	0.180	14	0.81	0.007	0.81	0.019	-0.002	0.021	0.909
SF6D	0.033	299	0.72	0.006	0.74	0.015	-0.022	0.016	0.193
80 + (N = 53)									
MCS	3.000	225	52.58	1.161	52.99	2.406	-0.413	2.679	0.878
PCS	2.000	487	46.35	1.359	45.32	3.470	1.030	3.670	0.786
EQ5D	0.180	14	0.81	0.018	0.82	0.038	-0.008	0.042	0.851
SF6D	0.033	299	0.73	0.018	0.74	0.035	-0.008	0.038	0.835

T2D: Correlation with Age

		Observed			Fused		
PRO	Variable	ρ	LL	UL	ρ	LL	UL
EQ5D	age	0.07	0.01	0.14	0.07	-0.03	0.16
MCS	age	0.30	0.23	0.36	0.19	0.12	0.26
PCS	age	-0.07	-0.14	0.00	-0.03	-0.13	0.06
SF6D	age	0.12	0.05	0.19	0.11	0.03	0.19

Type-2 Diabetes: By Gender

			Observed		Fused		Difference		
PRO	MCID	N_{min}	Mean	SE	Mean	SE	Mean	SE	P-value
Male (N = 409)									
MCS	3.000	225	50.45	0.517	50.66	1.359	-0.210	1.430	0.886
PCS	2.000	487	47.46	0.466	47.66	1.348	-0.202	1.421	0.889
EQ5D	0.180	14	0.81	0.007	0.80	0.017	0.010	0.018	0.603
SF6D	0.033	299	0.73	0.006	0.73	0.018	-0.006	0.019	0.760
Female (N = 474)									
MCS	3.000	225	49.11	0.522	49.44	1.027	-0.321	1.108	0.773
PCS	2.000	487	44.33	0.506	44.81	0.887	-0.481	1.003	0.632
EQ5D	0.180	14	0.77	0.008	0.78	0.015	-0.005	0.016	0.773
SF6D	0.033	299	0.69	0.006	0.70	0.012	-0.012	0.013	0.350

T2D: Correlation between PROs

		Observed			Fused		
PRO-1	PRO-2	ρ	LL	UL	ρ	LL	UL
EQ5D	SF6D	0.75	0.70	0.79	0.71	0.67	0.75
MCS	EQ5D	0.55	0.50	0.61	0.49	0.43	0.55
MCS	PCS	0.20	0.14	0.27	0.20	0.11	0.27
MCS	SF6D	0.71	0.66	0.76	0.71	0.66	0.75
PCS	EQ5D	0.68	0.63	0.73	0.68	0.64	0.72
PCS	SF6D	0.71	0.67	0.76	0.71	0.64	0.76

Results: Summary

Univariate Analysis:

1. **Non-disease Specific**: The differences in means are below 0.2 in absolute value and we fail to reject the hypothesis of no difference across all the PROs. As the sample sizes $N = 5207$ are much larger than N_{min}, the differences are well below MCID.
2. **Type-2 Diabetes (T2D)**: The differences in means are below 0.5 in absolute value and we fail to reject the hypothesis of no difference across all the PROs. As the sample sizes $N = 883$ are well above N_{min}, the differences are well below MCID.
3. **Myasthenia Gravis**: We fail to reject the null hypothesis of no difference across all PROs. Although the differences are lower than MCID, the probability of the difference being greater than MCID are higher except in EQ5D where the sample size $N = 100$ is greater than N_{min}.

Bivariate Analysis (Type-2 Diabetes):

1. **By Age**: The difference is less than 1 in absolute value when the sample size N is greater than N_{min}.
2. **By Gender**: The differences are less the 0.5 in absolute value across all cases when the sample N is greater than N_{min}.

Correlation Analysis (Type-2 Diabetes):

1. **Between PROs and Age**: The difference between observed and fusion based estimates is less the 0.05 and the 95% confidence intervals from the fused data includes the observed estimate. Except in the case of MCS where the difference is slightly larger and less than 0.1 and the confidence intervals barely miss the observed point estimate.
2. **Between PROs**: The correlation from the fused data are either identical to the correlation from the observed data or at least within the range of the 95% confidence intervals from the observed data estimates.

Conclusion

- In this work, we show the ability to implement data fusion in a disease agnostic way thereby enabling the use of more advanced machine learning algorithms on larger data sets, while still being able to use the resulting fused data to perform disease specific analysis.
- The advantages of using the linked data are twofold - (1) we do not have to impose the untestable and often unrealistic assumption of conditional independence and (2) the input variables for the data fusion model come from the same data source as the recipient data, thereby avoiding any concerns regarding consistency of definitions or time-frame of data collection amongst others.
- We demonstrate how to maximize the use of distinct non-overlapping healthcare data sets to gain insights using machine learning methods.