Distributed Learning from Multiple EHR Databases: Contextual Embedding Models for Medical Events

Supplementary materials

Ziyi Li[1], Kirk Roberts[2], Xiaoqian Jiang[2,*] and Qi Long[3,*]

[1]Emory University, Department of Biostatistics and Bioinformatics, Atlanta, GA 30332, USA
[2]University of Texas, Health Science Center at Houston, School of Biomedical Informatics, Houston, Texas 77030, USA
[3]University of Pennsylvania, Perelman School of Medicine, Department of Biostatistics and Epidemiology, Philadelphia, PA 19104, USA
[*]Xiaoqian Jiang (Xiaoqian.Jiang@uth.tmc.edu) and Qi Long (qlong@pennmedicine.upenn.edu) are joint corresponding authors.

February 11, 2019

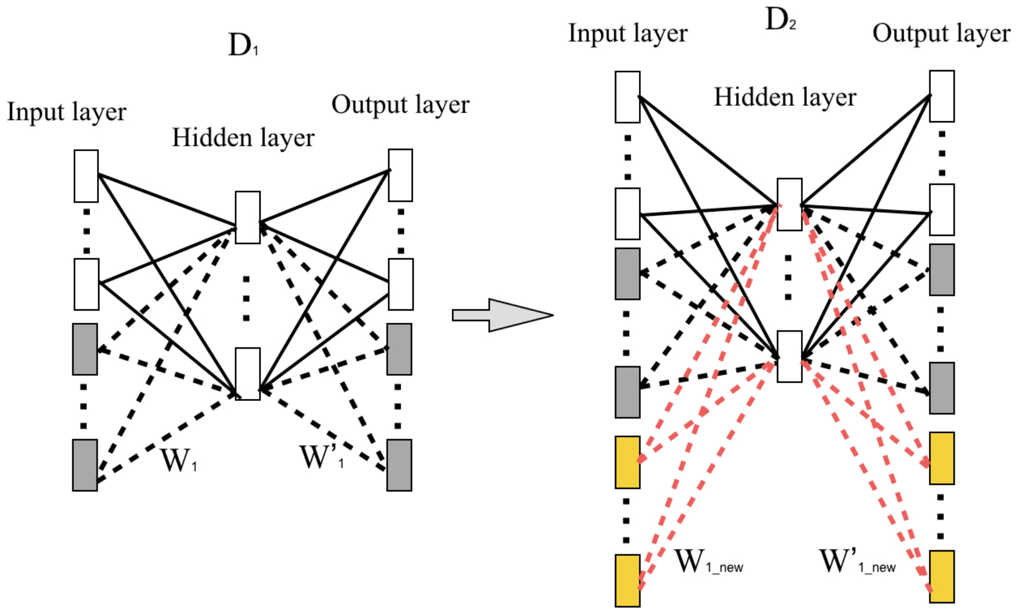


Figure S1: Naive updates. Left figure represents M_1 and right figure represents M_2. Empty squares represent the words exclusive to D_1. Gray squares are the words shared by both D_1 and D_2. Yellow squares are the words exclusive to D_2.

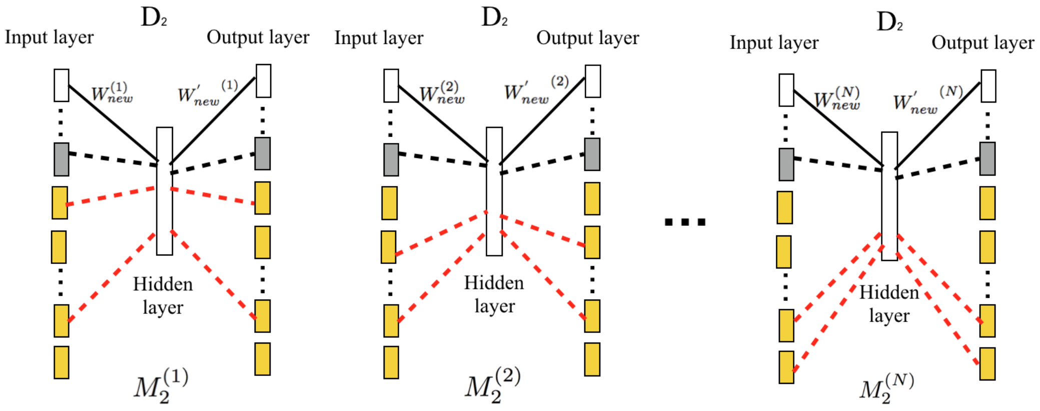
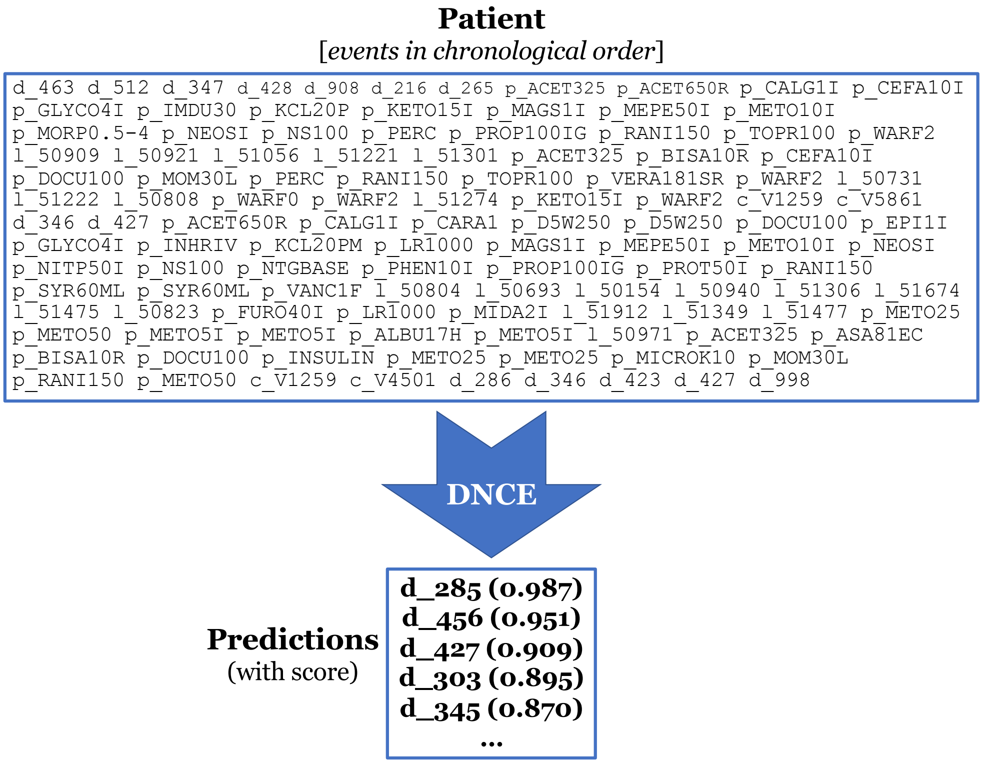


Figure S2: Dropout updates. This figure only demonstrate the second step in dropout updates, which is to update existing model using new dataset D_2. Empty squares represent the words exclusive to D_1. Gray squares are the words shared by both D_1 and D_2. Yellow squares are the words exclusive to D_2. When a node is not connected to hidden layer, it means this node is 'dropped' for the current training cycle.



```
d_463 d_512 d_347 d_428 d_908 d_216 d_265 p_ACET325 p_ACET650R p_CALG1I p_CEFA10I
p_GLYCO4I p_IMDU30 p_KCL20P p_KETO15I p_MAGS1I p_MEPE50I p_METO10I
p_MORP0.5-4 p_NEOSI p_NS100 p_PERC p_PROP100IG p_RANI150 p_TOPR100 p_WARF2
l_50909 l_50921 l_51056 l_51221 l_51301 p_ACET325 p_BISA10R p_CEFA10I
p_DOCU100 p_MOM30L p_PERC p_RANI150 p_TOPR100 p_VERA181SR p_WARF2 l_50731
l_51222 l_50808 p_WARF0 p_WARF2 l_51274 p_KETO15I p_WARF2 c_V1259 c_V5861
d_346 d_427 p_ACET650R p_CALG1I p_CARA1 p_D5W250 p_D5W250 p_DOCU100 p_EPI1I
p_GLYCO4I p_INHRIV p_KCL20PM p_LR1000 p_MAGS1I p_MEPE50I p_METO10I p_NEOSI
p_NITP50I p_NS100 p_NTGBASE p_PHEN10I p_PROP100IG p_PROT50I p_RANI150
p_SYR60ML p_SYR60ML p_VANC1F l_50804 l_50693 l_50154 l_50940 l_51306 l_51674
l_51475 l_50823 p_FURO40I p_LR1000 p_MIDA2I l_51912 l_51349 l_51477 p_METO25
p_METO50 p_METO5I p_METO5I p_ALBU17H p_METO5I l_50971 p_ACET325 p_ASA81EC
p_BISA10R p_DOCU100 p_INSULIN p_METO25 p_METO25 p_MICROK10 p_MOM30L
p_RANI150 p_METO50 c_V1259 c_V4501 d_286 d_346 d_423 d_427 d_998
```

d_285 (0.987)
d_456 (0.951)
d_427 (0.909)
d_303 (0.895)
d_345 (0.870)
...

Figure S3: A schematic plot to demonstrate the input and output of the proposed model. In this plot we demonstrate a sequence from one patient, d_* represent real diagnoses that were on record. l_* are events for lab tests. p_* are prescription event. Other possible events (not present in this patient) include s_* for symptoms and c_* for conditions.

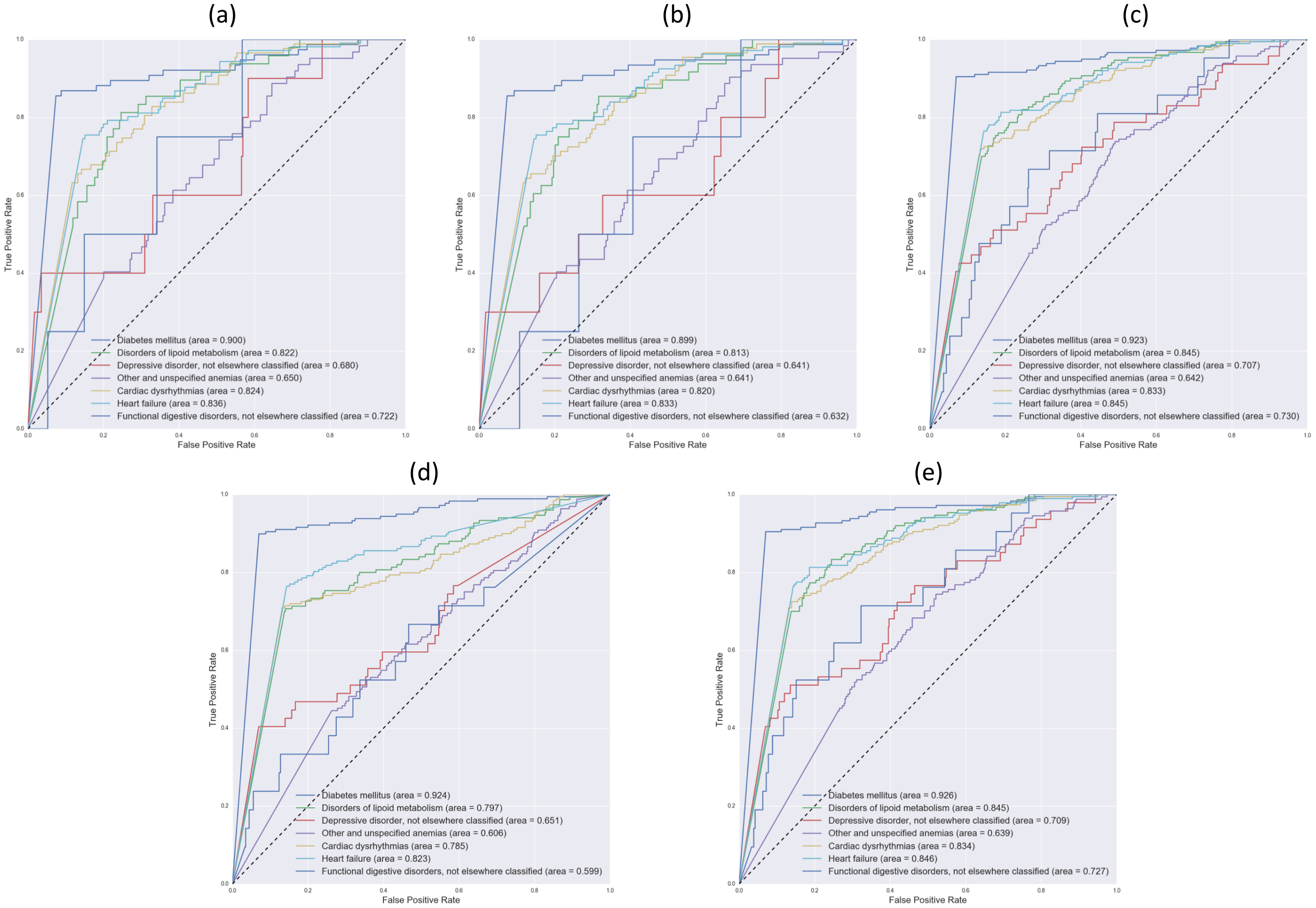
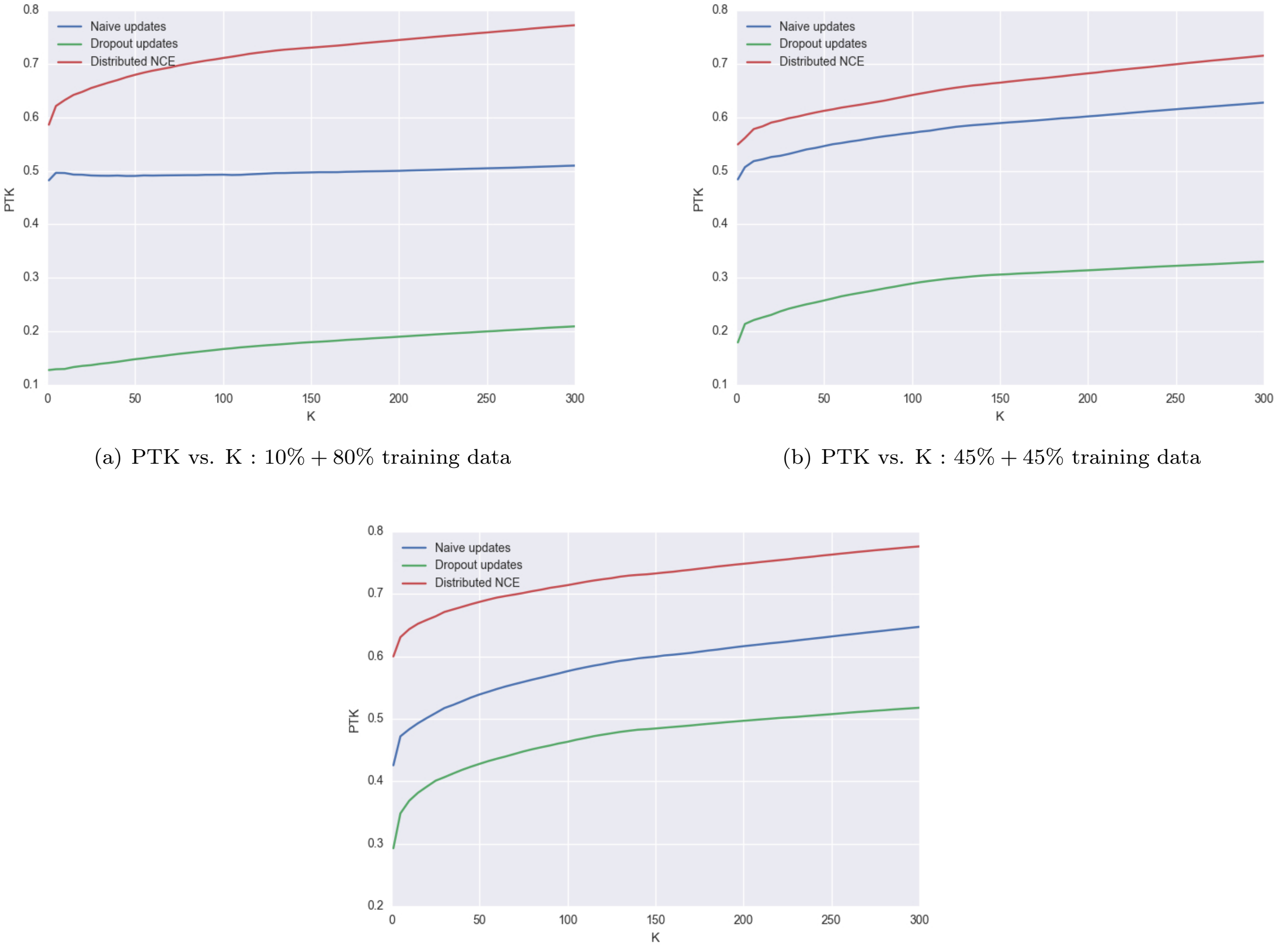


Figure S4: ROC curves for the top 7 most common diagnoses. Figure (a) is the results by PDPS using only the data from CareVue. Figure (b) is by PDPS using only the data from MetaVision. Figure (c), (d), and (e) are by Naive, Dropout, and DNCE algorithm respectively using both CareVue and MetaVision.



(a) PTK vs. K : 10% + 80% training data

(b) PTK vs. K : 45% + 45% training data

(c) PTK vs. K : 80% + 10% training data

Figure S5: Precision-top-K versus K for all methods

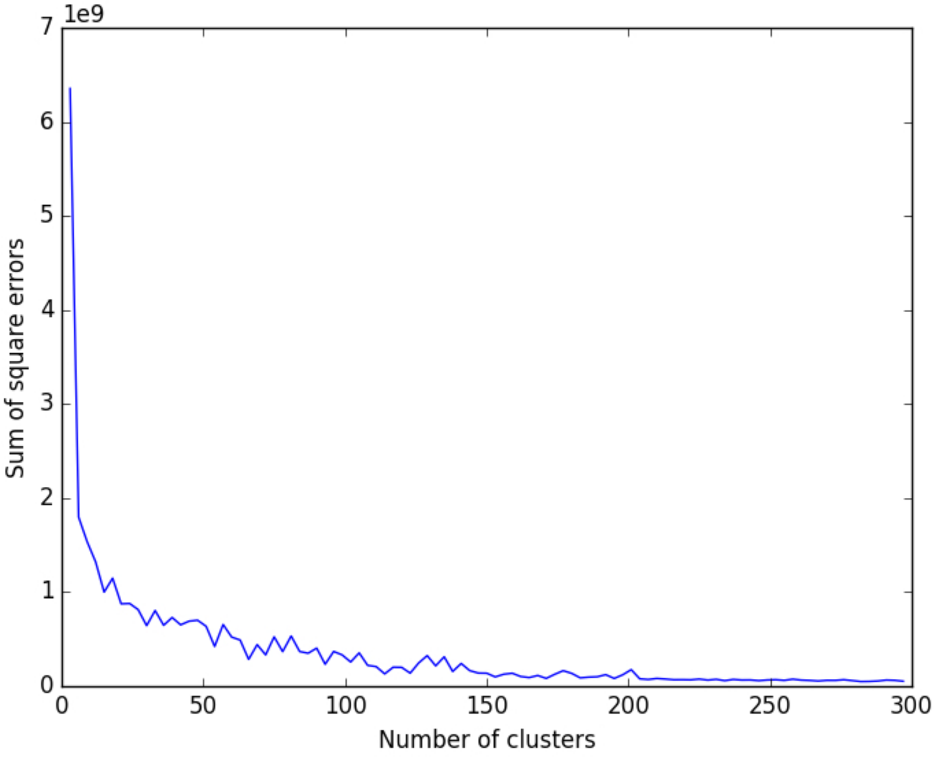


Figure S6: Sum of Squared Errors (with noise-added centroids) by Number of Clusters.

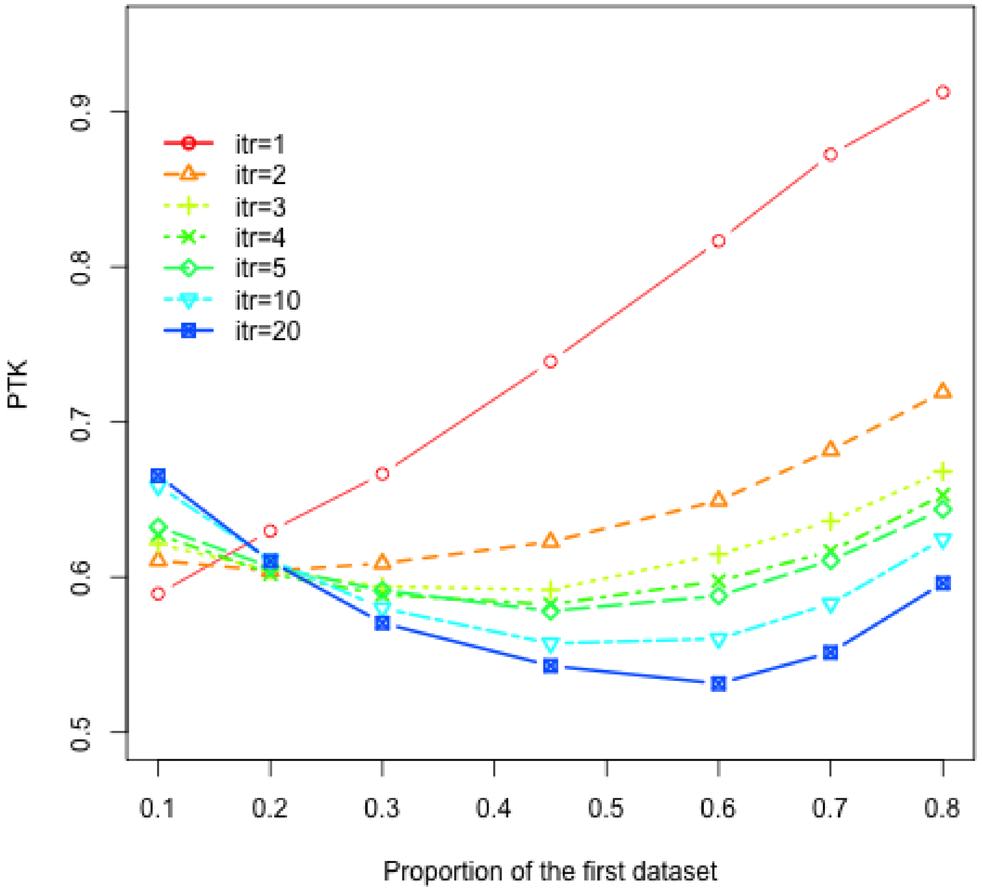


Figure S7: Precision-Top-K versus K by Distributed NCE using different training-set-partitions for different number of iterations. X-axis is the proportion of first training set. The first and second training sets add up to 90% of total data. All Distributed NCE models are compared with gold standard model, which is defined as global model trained will all the training data (90% of data).

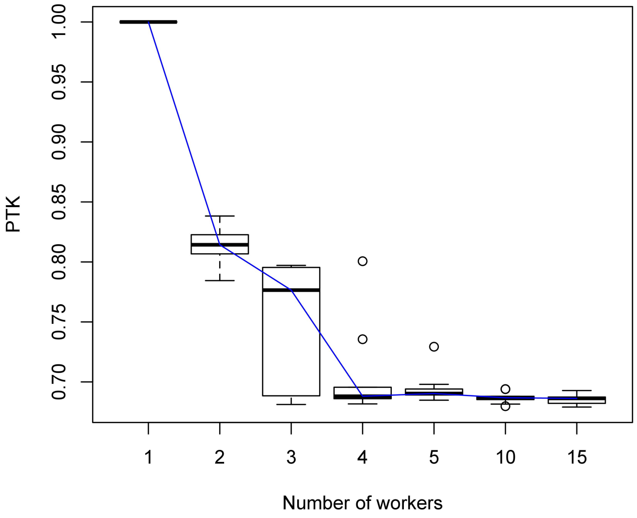


Figure S8: Boxplot of PTK between two global models versus Number of workers. Each box include the PTK values of two repetitive global models using ten-fold cross validation training data. Blue line connects group median.

	Diagnostic	Freq	Prescription	Freq (Person)	Lab tests	Freq (Person)	Symptom	Freq (Person)	Condition	Freq (Person)
Total Number	712		3553		1138		174		293	
Top 10 common	Cardiac dysrhythmias	2086	INSULIN	26409 (4157)	Hemoglobin	13350 (5544)	Septic shock	890 (757)	Long-term use anticoagul	1355 (918)
	Essential hypertension	2021	FURO40I	26080 (3380)	Red Blood Cells	13315 (5504)	Bacteremia	567 (505)	Aortocoronary bypass	1221 (732)
	Heart failure	2003	NACLFLUSH	22650 (4576)	Hematocrit	13298 (5536)	Convulsions	556 (409)	Long-term use of insulin	1038 (635)
	Disorders of fluid, electrolyte, and acid-base balance	1974	VANC1F	17958 (3645)	Glucose	13035 (5463)	Hypoxemia	448 (402)	Status-post ptca	850 (550)
	Diabetes mellitus	1787	VANCOBASE	17943 (3646)	Urea Nitrogen	10858 (4973)	Ascites NEC	381 (281)	History of tobacco use	822 (617)
	Other diseases of lung	1707	NS1000	17761 (3528)	Phosphate	10540 (5120)	Diarrhea	380 (328)	Hx-ven thrombosis/embols	779 (488)
	Acute renal failure	1583	NS500	17622 (3656)	White Blood Cells	10383 (5106)	Abnrml coagultion prfile	360 (319)	Tracheostomy status	461 (302)
	Other and unspecified anemias	1580	D5W250	17454 (3670)	PT	10195 (4871)	Cardiogenic shock	288 (266)	Status cardiac pacemaker	439 (265)
	Disorders of lipoid metabolism	1447	HEPA5I	16454 (4536)	Neutrophils	10154 (5011)	Retention urine NOS	266 (235)	Do not resusctate status	387 (301)
	Other forms of chronic ischemic heart disease	1356	MAG2PM	15891 (2773)	Calcium, Total	9986 (4941)	Sleep apnea NOS	192 (141)	Long-term use steroids	372 (266)

Table S1: Total number and top 10 diagnostics, prescriptions, lab tests, symptoms and conditions of all 5642 patients.

Database	N	Mean(SD)	n(%)				
		Age	Gender	Hispanic	Black	White	Asian
CareVue	2922	70.61 (53.2)	1615 (55.3)	81 (2.8)	333 (11.4)	2150 (73.6)	81 (2.8)
MetaVision	2693	76.12 (55.0)	1513 (56.2)	137 (5.1)	369 (13.7)	1993 (74.0)	66 (2.5)
Both or NA	15	-					

Table S2: Patient characteristics of database systems CareVue and MetaVision.

	Diagnostic	Freq	Prescription	Freq (Person)	Lab tests	Freq (Person)	Symptom	Freq (Person)	Condition	Freq (Person)
Total Number	635		2572		632		160		253	
Top 10 common	Disorders of fluid, electrolyte, and acid-base balance	1107	NACLFLUSH	14763 (2680)	Red Blood Cells	6783 (2667)	Septic shock	563 (471)	Long-term use anticoagul	980 (621)
	Essential hypertension	1047	MAG2PM	13576 (2253)	Hemoglobin	6781 (2659)	Hypoxemia	350 (309)	Long-term use of insulin	761 (422)
	Cardiac dysrhythmias	1038	INSULIN	12588 (2067)	Hematocrit	6728 (2658)	Ascites NEC	344 (251)	History of tobacco use	656 (472)
	Other and unspecified anemias	916	NS1000	11801 (2175)	Glucose	6723 (2679)	Bacteremia	277 (244)	Aortocoronary bypass	632 (347)
	Disorders of lipoid metabolism	912	FURO40I	10813 (1610)	Urea Nitrogen	5500 (2409)	Diarrhea	261 (223)	Hx-ven thrombosis/embols	629 (377)
	Heart failure	901	NS500	10188 (2062)	Phosphate	5410 (2515)	Abnrml coagultion prfile	249 (213)	Status-post ptca	527 (321)
	Diabetes mellitus	893	VANC1F	9978 (1941)	Neutrophils	5336 (2462)	Convulsions	184 (156)	Do not resusctate status	387 (301)
	Other diseases of lung	847	VANCOBASE	9970 (1942)	White Blood Cells	5141 (2460)	Dysphagia NOS	160 (135)	Renal dialysis status	354 (181)
	Acute renal failure	843	HEPA5I	9890 (2436)	PT	5111 (2360)	Retention urine NOS	152 (134)	Hx TIA/stroke w/o resid	315 (225)
	Chronic kidney disease	752	NS250	7556 (1782)	INR(PT)	5105 (2362)	Tachycardia NOS	146 (131)	Long-term use steroids	254 (170)

Table S3: Total number and top 10 diagnostics, prescriptions, lab tests, symptoms and conditions of 2693 patients from MetaVision.

	Diagnostic	Freq	Prescription	Freq (Person)	Lab tests	Freq (Person)	Symptom	Freq (Person)	Condition	Freq (Person)
Total Number	579		2374		814		138		207	
Top 10 common	Heart failure	1093	FURO40I	15134 (1753)	Hematocrit	6510 (2852)	Convulsions	370 (251)	Aortocoronary bypass	581 (383)
	Cardiac dysrhythmias	1042	INSULIN	13644 (2065)	Hemoglobin	6509 (2859)	Septic shock	323 (282)	Long-term use anticoagul	372 (294)
	Essential hypertension	970	D5W250	10228 (1896)	Red Blood Cells	6472 (2811)	Bacteremia	288 (259)	Status-post ptca	314 (224)
	Diabetes mellitus	884	MAGS1I	9698 (1734)	Glucose	6256 (2758)	Prev matern surg aff NB	176 (142)	Long-term use of insulin	273 (210)
	Disorders of fluid, electrolyte, and acid-base balance	859	MICROK10	8889 (1809)	Urea Nitrogen	5310 (2541)	Cardiogenic shock	147 (135)	Inf mcrg rstn pncllins	262 (222)
	Other diseases of lung	851	NS250	7931 (1684)	White Blood Cells	5199 (2623)	Sleep apnea NOS	146 (102)	Tracheostomy status	228 (171)
	Acute renal failure	731	VANC1F	7913 (1683)	Phosphate	5078 (2580)	Diarrhea	117 (103)	Status cardiac pacemaker	183 (119)
	Other and unspecified anemias	658	VANCOBASE	7905 (1683)	PT	5040 (2487)	Retention urine NOS	113 (100)	Hx of breast malignancy	162 (91)
	Other forms of chronic ischemic heart disease	654	NACLFLUSH	7747 (1869)	Calcium, Total	4897 (2510)	Abnrml coagultion prfile	109 (104)	History of tobacco use	162 (142)
	Septicemia	535	KCL20PM	7410 (1595)	Neutrophils	4772 (2525)	Fever NOS	107 (96)	Hx-ven thrombosis/embols	148 (110)

Table S4: Total number and top 10 diagnostics, prescriptions, lab tests, symptoms and conditions of 2922 patients from CareVue.

	Proportion of First Training Data						
	10%	20%	30%	45%	60%	70%	80%
Distributed NCE	0.637(2e-3)	0.612(4e-3)	0.593(2e-3)	0.581(2e-3)	0.588(2e-3)	0.610(2e-3)	0.648(3e-3)
Distributed NCE with DP (10 clusters)	0.615(7e-3)	0.586(8e-3)	0.569(5e-3)	0.557(5e-3)	0.562(3e-3)	0.577(6e-3)	0.609(5e-3)
Distributed NCE with DP (30 clusters)	0.627(8e-3)	0.606(7e-3)	0.588(5e-3)	0.570(5e-3)	0.572(6e-3)	0.582(6e-3)	0.616(1e-2)
Distributed NCE with DP (50 clusters)	0.627(7e-3)	0.600(6e-3)	0.583(3e-3)	0.567(6e-3)	0.570(4e-3)	0.582(5e-3)	0.615(8e-3)
Distributed NCE with DP (100 clusters)	0.621(7e-3)	0.601(8e-3)	0.586(5e-3)	0.566(5e-3)	0.568(5e-3)	0.583(7e-3)	0.606(1e-2)
Distributed NCE with DP (150 clusters)	0.620(7e-3)	0.599(4e-3)	0.578(5e-3)	0.563(7e-3)	0.565(8e-3)	0.578(7e-3)	0.604(6e-3)

Table S5: Precision-Top-K of Distributed NCE without or with DP. Results are summarized over 10-folds cross validation. $Skip - gram$ model is used. Distributed NCE is Distributed Noise Contrastive Estimation. DP is Differential Privacy. The total proportion of training data is 90%. For example, if the proportion of first training data is 10%, the proportion of second training data should be 80%.

	Naive updates		Dropout udpates		Distributed	
	PTK	Avg_AUC	PTK	Avg_AUC	PTK	Avg_AUC
10:80:10	0.496 (2e-3)	0.770 (8e-3)	0.131 (9e-4)	0.720 (5e-3)	0.637 (2e-3)	0.774 (8e-3)
20:70:10	0.532 (2e-3)	0.773 (7e-3)	0.162 (3e-3)	0.721 (7e-3)	0.612 (4e-3)	0.774 (7e-3)
30:60:10	0.532 (3e-3)	0.772 (8e-3)	0.184 (9e-4)	0.722 (7e-3)	0.593 (2e-3)	0.774 (7e-3)
45:45:10	0.516 (2e-3)	0.772 (8e-3)	0.219 (3e-3)	0.723 (7e-3)	0.581 (2e-3)	0.773 (8e-3)
60:30:10	0.492 (3e-3)	0.774 (8e-3)	0.260 (3e-3)	0.724 (5e-3)	0.588 (2e-3)	0.773 (8e-3)
70:20:10	0.483 (2e-3)	0.774 (8e-3)	0.307 (5e-3)	0.727 (6e-3)	0.610 (2e-3)	0.773 (7e-3)
80:10:10	0.485 (3e-3)	0.775 (8e-3)	0.370 (4e-3)	0.736 (7e-3)	0.648 (3e-3)	0.773 (7e-3)

Table S6: Simulation results of all methods using $Skip-Gram$ model. Results are summarized over 10-folds cross validation. Distributed NCE is Distributed Noise Contrastive Estimation. PTK is Precision-Top-K. Avg-AUC is averaged Area-Under-Curve. $n_{t1} : n_{t2} : n_{test}$ means that the two training datasets are n_{t1}% and n_{t2}% of total data. Testing dataset is n_{test}% of total data.

	Age: Mean(std)		Naive updates		Dropout updates		Distributed NCE		Distributed NCE with DP (30 clusters)	
	Group_1	Group_2	PTK	$Avg-$AUC	PTK	$Avg-$AUC	PTK	$Avg-$AUC	PTK	$Avg-$AUC
$b_1 = -0.002$	77.7(1.0)	71.3(0.5)	0.5348 (4e-3)	0.7742 (7e-3)	0.1696 (2e-3)	0.7202 (8e-3)	0.5961 (3e-3)	0.7757 (8e-3)	0.5933 (3e-3)	0.7757 (8e-3)
$b_1 = -0.02$	84.0(0.5)	58.4(0.5)	0.5036 (3e-3)	0.7756 (8e-3)	0.2272 (3e-3)	0.7230 (6e-3)	0.5694 (2e-3)	0.7761 (8e-3)	0.5575 (4e-3)	0.7762 (7e-3)
$b_1 = -0.04$	79.0(0.2)	50.1(0.5)	0.4715 (5e-3)	0.7763 (7e-3)	0.2947 (5e-3)	0.7255 (6e-3)	0.5909 (3e-3)	0.7753 (7e-3)	0.5690 (6e-3)	0.7758 (7e-3)

Table S7: Simulation results of all methods using $Skip - Gram$ model for age-divided populations. Distributed NCE is Distributed Noise Contrastive Estimation. PTK is Precision-Top-K. Avg-AUC is averaged Area-Under-Curve. Patients are assigned to Group_1 and Group_2 by $Logit(Pr(S = 1)) = 1 + b_1 \times Age$. Here $S = 1$ means subject is assigned to Group_2. Testing dataset is 10% of total data and is randomly chosen without age-correlation. Results are summarized over 10 simulation datasets.

	Age $\leq$ 53 : Age $>$ 53		Age $\leq$ 66 : Age $>$ 66		Age $\leq$ 77 : Age $>$ 77	
	PTK	*Avg_AUC*	PTK	*Avg_AUC*	PTK	*Avg_AUC*
Naive udpates	0.5070	0.7726	0.4897	0.7700	0.4749	0.7701
Dropout updates	0.1860	0.7228	0.2700	0.7285	0.3438	0.7325
Distributed NCE	0.5730	0.7748	0.5543	0.7718	0.5907	0.7704
Distributed NCE with DP (30 clusters)	0.5722	0.7750	0.5451	0.7722	0.5890	0.7692

Table S8: Simulation results of all methods using $Skip - Gram$ model for age-divided populations. Distributed NCE is Distributed Noise Contrastive Estimation. PTK is Precision-Top-K. Avg-AUC is averaged Area-Under-Curve. "Age ≤ 53 : Age > 53" means that the two training datasets are divided by the age of subjects. One training dataset includes all subjects with age smaller than 53 and the other training dataset includes all subjects with age greater than 53. Testing dataset is 10% of total data and is randomly chosen without age-correlation.