Appendices for

Title: Building Bridges Across Electronic Health Record Systems Through Inferred Phenotypic Topics

Authors: You Chen[1], Joydeep Ghosh[2], Cosmin Adrian Bejan[1], Carl A. Gunter[3], Siddharth Gupta[3], Abel Kho[4], David Liebovitz[4], Jimeng Sun[5], Joshua Denny[1,6], and Bradley Malin[1,7]

Author Affiliations:

[1]Dept. of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN

[2]Dept. of Electrical & Computer Engineering, University of Texas, Austin, TX

[3]Dept. of Computer Science, University of Illinois at Urbana-Champagne, Champaign, IL

[4]School of Medicine, Northwestern University, Chicago, IL

[5]School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, GA

[6]Department of Medicine, Vanderbilt University, Nashville, TN

[7]Dept. of Electrical Engineering & Computer Science, School of Engineering, Vanderbilt University, Nashville, TN

To Whom Correspondence Should Be Addressed:

You Chen, Ph.D.
2525 West End Ave, Suite 1030
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37203 USA
Email: you.chen@vanderbilt.edu
Phone: +1 615 343 1939
Fax: +1 615 322 0502

Appendix A1: Weighting a Distance Function to Account for Outliers

In the manuscript, we use a regression (Equation 5) to assess the similarity of the rate at which concepts (e.g., phenotypic topics, ICD-9 codes and PheWAS codes) occur in disparate patient populations. It is important to consider the impact of outlies when evaluating the transferability of concepts across disparate sites. For instance, Figure A1, depicts a hypothetical characterization of the rates at which concepts are realized in the NMH and VUMC patient populations. In this scenario, outliers in the upper left and lower right sections should have a significant influence on the transferability of concepts.

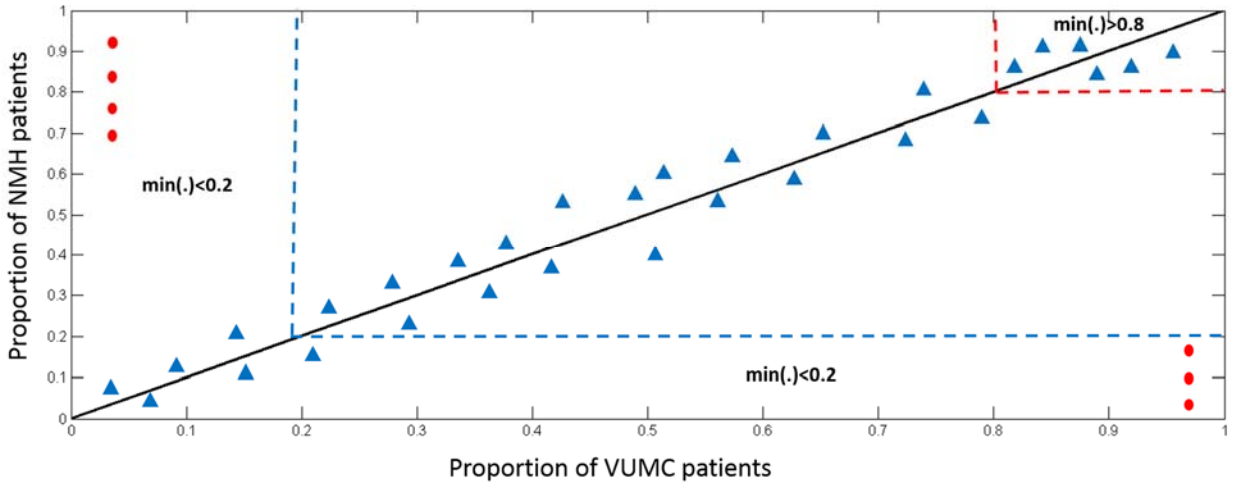

Figure A1. A hypothetical depiction of the rates at which clinical concepts are realized by patients in the NHM and VUMC patient populations. The circles constitute outliers, while the triangles are within an expected range of variation from the regressed line.

There are various ways by which deviance from a regression can be weighted. Several strategies are illustrated in Figure A2. Notice that the distance score grows linearly with increasing $max(\cdot)$ values for the functions $log(max(\cdot)\text{-}min(\cdot)+1)$ and $max(\cdot)\text{-}min(\cdot)$. By contrast, the term $max(\cdot)/min(\cdot)$ ensures that distance grows exponentially as points deviate from the regressed line. To ensure that

the outlying points are weighted more heavily in a distance function and the distance of a point that falls on the regressed line is equal to zero, we integrated the exponential scaling factor *max*(·)/*min*(·) and a logarithmic transformation *log*(*max*(·)-*min*(·)+1) into Equation 6.

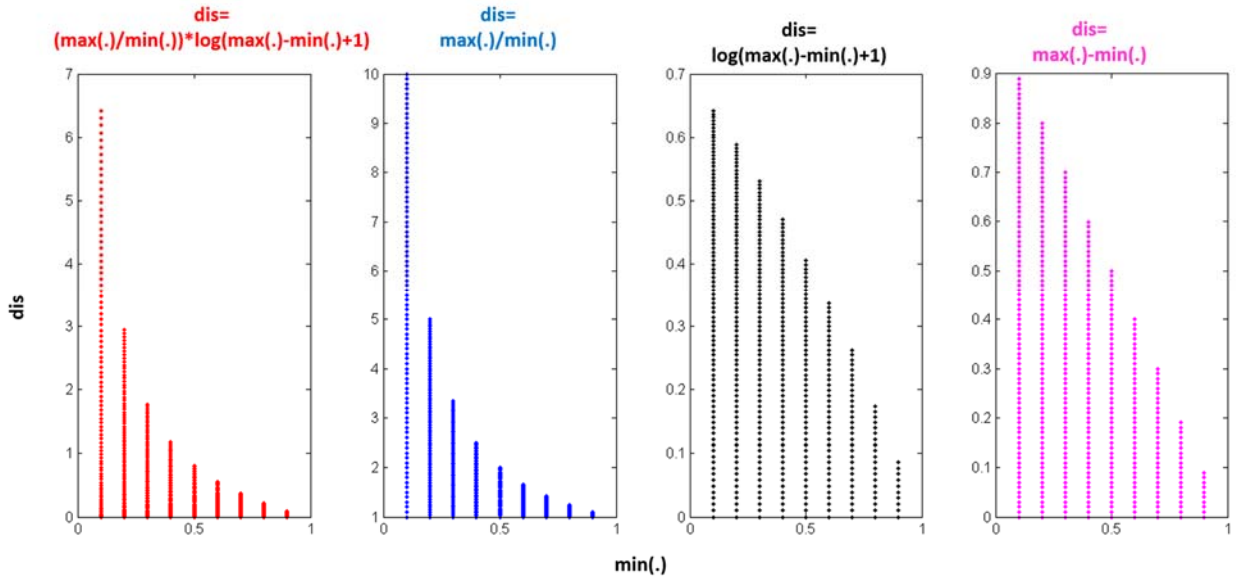

Figure A2. Influence of four measurements on the distance of a point to the regressed line. From left to right: the influence of *max*(·)/*min*(·) decreases exponentially with the increasing *min*(·) values, while *max*(·)-*min*(·) and *log*(*max*(·)-*min*(·) +1) decreases lineally with the increased value of *min*(·). The first subfigure is an integration of exponential scaling factor *max(·)/min(·)* and a logarithmic transformation *log(max(·)-min(·)+1)*, which is used in our method to measure the distance of a point to the regressed line.

Appendix A2: Setting the Number of Phenotypic Topics

To parameterize the number of phenotypic topics for the LDA model, we minimize 1) the perplexity score and 2) the average similarity of the topics within a site.

The perplexity analysis for both datasets is depicted in Figure A3. We adopted a 10-fold cross-validation strategy to obtain the average perplexity scores and corresponding standard deviations for models learned according to a varying number of phenotypic topics. It can be seen that the perplexity score stabilizes when the number of topics is larger than 40 and 50 for the NMH and VUMC data, respectively.

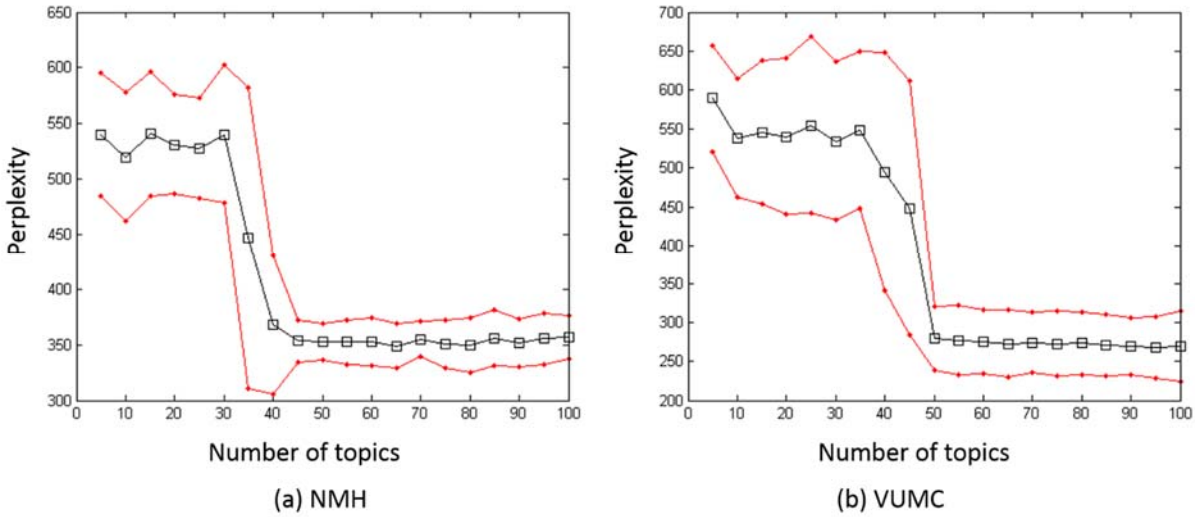

Figure A3. Perplexity of LDA models for (*a*) NMH and (*b*) VUMC.

If we were to rely solely on perplexity, the number of topics for the models associated with the two datasets would be set to 40 and 50 respectively. However, the smallest perplexity score does not necessarily indicate the best model. As such, we search for a model that minimizes the phenotypic topic similarity within a site. Figure A4 depicts the average (and one standard deviation) of the average similarity for the datasets using LDA models learned over 15, 25, 40, and 50 phenotypic topics.

Based on Figure A4, it can be seen that the average similarity of VUMC phenotypic topics is smaller than NMH, ranging from 1.76×10^{-6} to 1.59×10^{-4} (Figure A4(b)) By contrast, the average similarity for NMH topics ranges from 0.031 to 0.057 (Figure A4(a)). When setting the number of topics to 25, the phenotypic topics learned from the NMH dataset exhibit the lowest average mean and standard deviation.

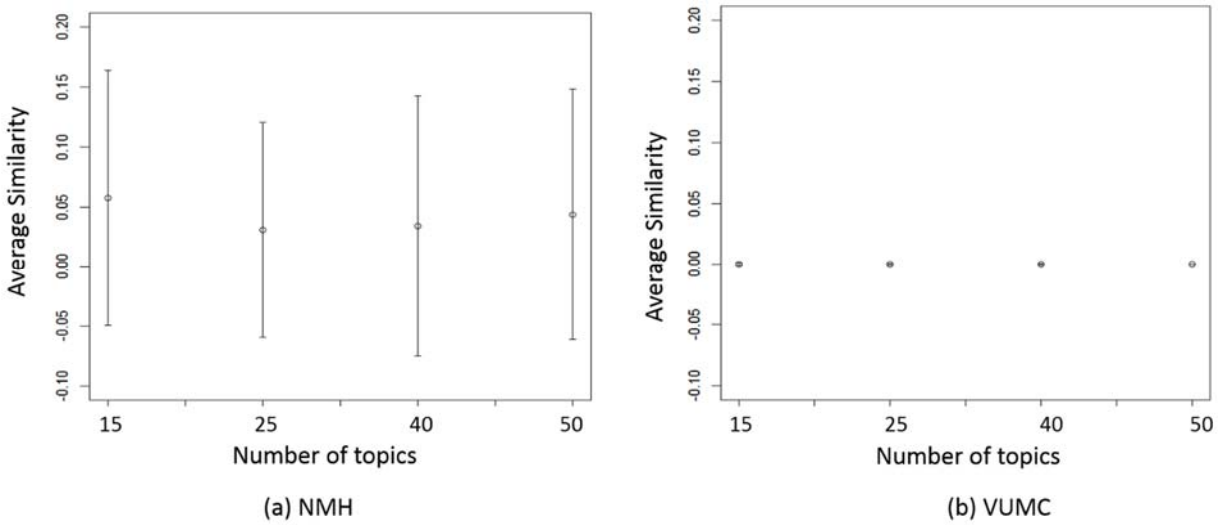

Figure A4. Average similarity (+/- 1 standard deviation) over various phenotypic topic models, derived via LDA, for (*a*) NMH and (*b*) VUMC.

To enable a fair comparison of the phenotypic topics learned from NMH and VUMC, we fixed the number of topics to 25 for each site. Specifically, we selected the 25 topics with the smallest average similarity for NMH data. Figure A5 depicts the similarity of the phenotypic topics for the two sites. It is apparent that the VUMC topics generally exhibit a smaller similarity than NMH topics.

Figure A6 displays a network of the similarity of the 25 topics derived from the NMH (N) dataset after triaging relations smaller than 0.2. It is clear there are two communities of phenotypic topics. The first pertains to topics N_1, N_6, N_8, N_{14}, N_{21}, and N_{22}, which are strongly related to

hypertension, hyperlipidemia, hypothyroidism and type 2 diabetes, the details for which are presented in Appendix A4. The second consists of phenotypic topics N_2, N_4, and N_{17}, which are related pregnancy issues (e.g., obstetrical and/or birth trauma), the details for which are provided in Appendix A4.

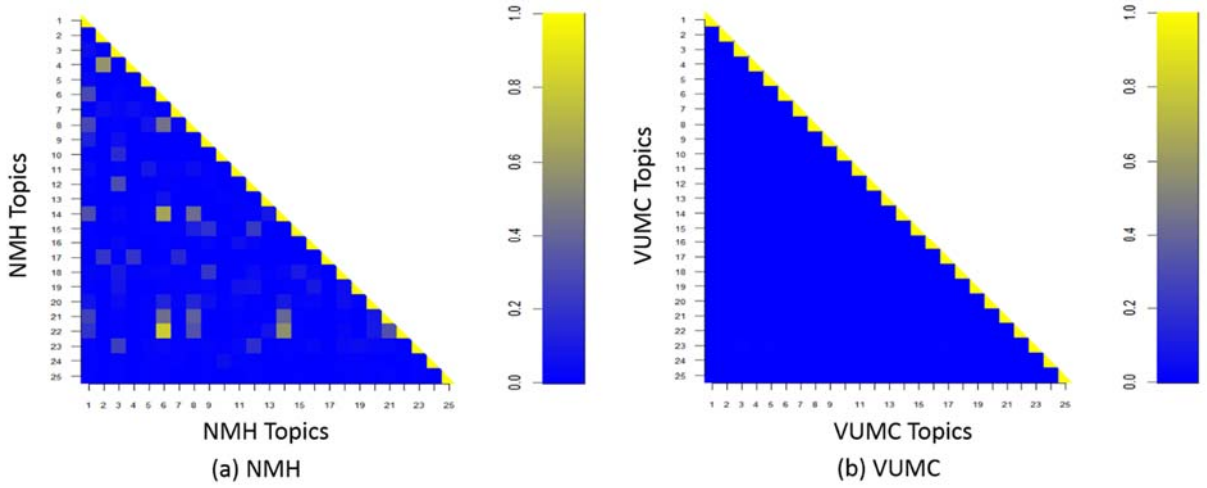

Figure A5. Similarity for the 25 phenotypic topics learned from (*a*) NMH and (*b*) VUMC.

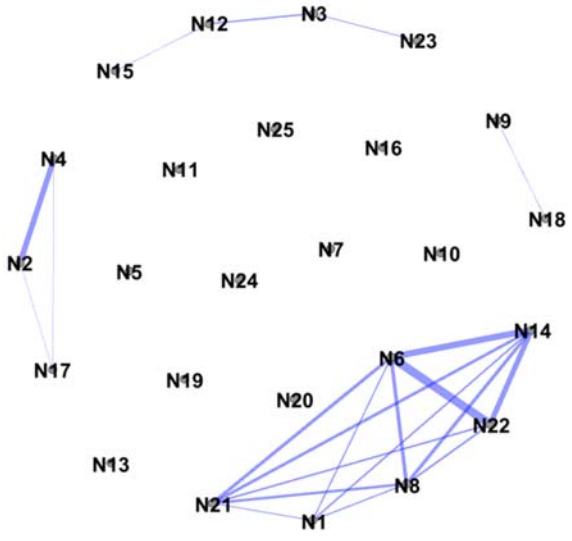

Figure A6. A network view of the similarity for the 25 phenotypic topics learned from NMH (score ≥ 0.2).

Appendix A3: VUMC Phenotypic Topics

Each VUMC topic is shown as the first 5 PheWAS codes, described by their clinical terms. Each PheWAS code was assigned with a probability of that code to the corresponding topic. The probability for a topic corresponds to the sum of the probabilities of the first five codes to that topic.

V_1 0.48173

Other perinatal conditions	0.18682
Short gestation; low birth weight; and fetal growth retardation	0.10986
Infectious & parasitic conditions complicating pregnancy	0.07834
Other respiratory conditions of fetus and newborn	0.05515
Venous embolism & thrombosis	0.05156

V_2 0.32028

severe protein-calorie malnutrition	0.09638
Secondary malignant neoplasm	0.06290
Bacterial pneumonia	0.05704
Secondary malignancy of bone	0.05206
Secondary malignancy of lymph nodes	0.05189

V_3 0.39933

Depression	0.09575
Anxiety disorder	0.08690
Constipation	0.08536
Crohn's disease	0.06925
Deep vein thrombosis	0.06206

V_4 0.43609

Other conditions of the mother complicating pregnancy	0.12554
Obesity	0.10778
Obstetrical/birth trauma	0.09042
Abnormality in fetal heart rate or rhythm	0.06139
Late pregnancy and failed induction	0.05096

V_5 0.48270

GERD	0.23483
Hypothyroidism	0.09983
Other intestinal obstruction	0.05916
Myeloid leukemia, acute	0.05259
Nonrheumatic aortic valve disorders	0.03629

V_6 0.57452

Heart failure	0.37697
Epilepsy	0.05899
Complications of medical procedures NOS	0.04915
Acute appendicitis	0.04650
Methicillin resistant Staphylococcus aureus	0.04291

V_7 0.77800

Coronary atherosclerosis	0.31215
Myocardial infarction	0.21671
Chronic airway obstruction	0.09936
pulmonary heart disease	0.08523
Nonspecific chest pain	0.06455

V_8 0.57402

Morbid obesity	0.17615
Skull fracture and other intercranial injury	0.15827
Substance addiction and disorders	0.11758
Bipolar	0.06340
Major depressive disorder	0.05863

V_9 0.56931

Infection/inflammation of internal prosthetic device, implant or graft	0.13763
Thrombocytopenia	0.13233
Hypotension NOS	0.12370
Postoperative infection	0.10310
Viral hepatitis C	0.07255

V_{10} 0.60926

Hypertensive chronic kidney disease	0.23358
End stage renal disease	0.15419
Chronic renal failure	0.09720
Kidney replaced by transplant	0.07930
Polyneuropathy in diabetes	0.04498

V_{11} 0.83954

Essential hypertension	0.49256
Systolic/diastolic heart failure	0.20779
Ischemic stroke	0.07524
Other specified cardiac dysrhythmias	0.04252
Rheumatoid arthritis	0.02143

V_{12} 0.50107

Cerebral edema and compression of brain	0.21038
Gram negative septicemia	0.08635
Convulsions	0.07536
Bacterial infection NOS	0.06832
Gastrointestinal complications	0.06066

V_{13} 0.89707

Acute renal failure	0.37797
Respiratory failure	0.24952
Shock	0.12333
Septicemia	0.10724
Valvular heart disease/ heart chambers	0.03900

V_{14} 0.47135

Other tests	0.15737
Fracture of pelvis	0.10241
Primary/intrinsic cardiomyopathies	0.08513
Failure to thrive	0.07079
Cirrhosis of liver without mention of alcohol	0.05566

V_{15} 0.53931

Internal injury to organs	0.26826
Early complications of trauma or procedure	0.08184
Fracture of ribs	0.08000
Paralytic ileus	0.05590
Other conditions of brain	0.05331

V_{16} 0.51065

Urinary tract infection	0.23172
Nausea and vomiting	0.12815
Prostate cancer	0.05365
Hydronephrosis	0.05252
Nonrheumatic mitral valve disorders	0.04461

V_{17} 0.50830

Hypovolemia	0.17984
Pleurisy; pleural effusion	0.14479
Anemia NOS	0.07004
Lung cancer	0.06049
Paroxysmal ventricular tachycardia	0.05314

V_{18} 0.46820

Burns	0.30163
Mechanical complications of cardiac/vascular device, implant, and graft	0.05974
Breast cancer	0.04269
Pneumonia due to fungus	0.03940
Lymphadenitis	0.02474

V_{19} 0.64038

Pneumonia	0.29317
Decubitus ulcer	0.16559
Other cerebral degenerations	0.08883
Asphyxia and hypoxemia	0.04700
Sickle cell anemia	0.04579

V_{20} 0.80005

Respiratory insufficiency	0.36680
Protein-calorie malnutrition	0.17254
Hypopotassemia	0.15145
Empyema and pneumothorax	0.05584
Pulmonary collapse; interstitial/compensatory emphysema	0.05341

V_{21} 0.68398

Acute posthemorrhagic anemia	0.24348
Atrial fibrillation	0.20199
Pneumonitis due to inhalation of food or vomitus	0.11273
Coma; stupor; and brain damage	0.06789
Ascites (non malignant)	0.05789

V_{22} 0.56563

Tobacco use disorder	0.22992
Cardiac shunt/ heart septal defect	0.12256
Asthma	0.09291
Asthma with exacerbation	0.06595
Epilepsy, recurrent seizures, convulsions	0.05429

V_{23} 0.37276

Fever of unknown origin	0.10126
Chemotherapy	0.07136
Dysphagia	0.07101
Aplastic anemia	0.06858
Candidiasis	0.06054

V_{24} 0.64845

Fracture of vertebral column without mention of spinal cord injury	0.21823
Acidosis	0.20105
Hyposmolality and/or hyponatremia	0.12733
Gram positive septicemia	0.06636
Diarrhea	0.03548

V_{25} 0.63491

Type 2 diabetes	0.33217
Hyperlipidemia	0.16443
Infantile cerebral palsy	0.05062
Insulin pump user	0.05049
Other dypnea	0.03719

Appendix A4: NMH Phenotypic Topics

Each NMH topic is shown as the first 5 PheWAS codes, described by their clinical terms. Each PheWAS code was assigned with a probability of that code to the corresponding topic. The probability for a topic corresponds to the sum of the probabilities of first five codes to that topic.

N_1	**0.35282**
Chronic airway obstruction	0.09593
Pneumonia	0.07922
GERD	0.06631
Essential hypertension	0.06201
Asphyxia and hypoxemia	0.04935

N_2	**0.95107**
Obstetrical/birth trauma	0.32306
Other conditions of the mother complicating pregnancy	0.26877
Other tests	0.20353
Umbilical cord complications during labor and delivery	0.11171
Indications for care or intervention related to labor and delivery NEC	0.04401

N_3	**0.36714**
Anemia NOS	0.10629
Other abnormal glucose	0.08176
Other specified cardiac dysrhythmias	0.07026
Atrial fibrillation	0.05697
Pulmonary collapse; interstitial/compensatory emphysema	0.05187

N_4	**0.66990**
Obstetrical/birth trauma	0.24844

Umbilical cord complications during labor and delivery	0.11880
Indications for care or intervention related to labor and delivery NEC	0.10311
Fetal distress and abnormal forces of labor	0.10127
Abnormality in fetal heart rate or rhythm	0.09827

N_5 0.42972

Substance addiction and disorders	0.11843
Tobacco use disorder	0.11592
Alcoholism	0.08071
Depression	0.06035
Anxiety disorder	0.05431

N_6 0.44478

Essential hypertension	0.18047
Hyperlipidemia	0.08107
Type 2 diabetes	0.07480
Cellulitis and abscess of leg	0.05698
Chronic ulcer of leg or foot	0.05146

N_7 0.49109

Asthma	0.18942
Abnormality pelvic soft tissues & organs complicating pregnancy	0.13175
Uterine leiomyoma	0.08468
Indications for care or intervention related to labor and delivery NEC	0.04318
Migraine	0.04206

N_8 0.51154

Hypothyroidism	0.16896

Essential hypertension	0.14391
GERD	0.07318
Osteoporosis	0.06410
Depression	0.06139

N_9 0.33436

Acute renal failure	0.09135
Sepsis	0.06990
Respiratory failure	0.06724
Pleurisy; pleural effusion	0.05518
Septicemia	0.05068

N_{10} 0.58452

Systolic/diastolic heart failure	0.19610
Atrial fibrillation	0.16233
Heart failure	0.11388
Primary/intrinsic cardiomyopathies	0.06061
pulmonary heart disease	0.05159

N_{11} 0.39977

Abdominal pain	0.12018
Nausea and vomiting	0.10161
Pain NEC	0.06055
Nonspecific chest pain	0.05872
Back pain	0.05872

N_{12} 0.54127

Anemia NOS	0.14936

Hypertensive chronic kidney disease	0.13435
End stage renal disease	0.11523
Renal dialysis	0.07141
Kidney replaced by transplant	0.07093

N_{13} 0.86693

Coronary atherosclerosis	0.56342
Myocardial infarction	0.15487
Hyperlipidemia	0.10770
Peripheral arterial disease	0.02158
Nonspecific chest pain	0.01936

N_{14} 0.60177

Essential hypertension	0.16974
Hyperlipidemia	0.14817
Hyperplasia of prostate	0.10372
Prostate cancer	0.09254
GERD	0.08760

N_{15} 0.53511

Chronic renal failure	0.19382
Hypertensive chronic kidney disease	0.12289
Acute renal failure	0.11166
Gout	0.05665
Hypothyroidism	0.05009

N_{16} 0.31399

Breast cancer	0.11182

Secondary malignant neoplasm of liver	0.05421
Lung cancer	0.05421
Secondary malignant neoplasm	0.04826
Colon cancer	0.04549

N_{17} 0.40271

Malposition and malrepresentation of fetus or obstruction	0.14965
Other conditions of the mother complicating pregnancy	0.07310
Indications for care or intervention related to labor and delivery NEC	0.07117
Multiple gestation	0.05676
Problems associated with amniotic cavity and membranes	0.05203

N_{18} 0.25722

Ascites (non malignant)	0.06477
Viral hepatitis C	0.06196
Cirrhosis of liver without mention of alcohol	0.05140
Acute renal failure	0.03990
Liver abscess and sequelae of chronic liver disease	0.03919

N_{19} 0.28474

Fever of unknown origin	0.08922
Non-Hodgkin's lymphoma	0.05787
Chemotherapy	0.05239
Pancytopenia	0.04450
Neutropenia	0.04077

N_{20} 0.26891

Transient cerebral ischemia	0.07341

Late effects of cerebrovascular disease	0.07166
Epilepsy, recurrent seizures, convulsions	0.04669
Ischemic stroke	0.03945
Convulsions	0.03770

N_{21} 0.63354

Essential hypertension	0.16522
Obstructive sleep apnea	0.15379
Osteoarthritis; localized	0.11677
Morbid obesity	0.10509
Obesity	0.09267

N_{22} 0.82490

Type 2 diabetes	0.31242
Essential hypertension	0.26649
Hyperlipidemia	0.19684
Insulin pump user	0.03655
Type 2 diabetic ketoacidosis	0.01260

N_{23} 0.44673

Hypovolemia	0.14589
Hypopotassemia	0.13856
Anemia NOS	0.07845
Magnesium metabolism disorder	0.04765
Other intestinal obstruction	0.03617

N_{24} 0.48362

Venous embolism & thrombosis	0.19268

Deep vein thrombosis	0.10287
Hemorrhagic disorder due to intrinsic circulating anticoagulants	0.09058
Pulmonary embolism and infarction	0.06013
Skull fracture and other intercranial injury	0.03736

N_{25} 0.51642

Urinary tract infection	0.23130
Decubitus ulcer	0.10862
Bacterial infection NOS	0.07603
E. coli	0.05283
Other paralytic syndromes	0.04764

Appendix A5: Cost of Phenotypic Topic Alignment

Based on Equation 3, the smaller the similarity value, the higher its cost. It was found that the total cost for a maximum matching of topics between NMH and VUMC is 15.26. The average cost for each pair of phenotypic topics is 0.61, which implies that the average cosine similarity for a pair of aligned phenotypic topics is 0.39. The cost for each match is provided Table A1.

Table A1 Alignment of NMH and VUMC phenotypic topics and their corresponding cost (range from 0 to 1).

VUMC	NMH	Cost	Random Cost
V_1	N_{17}	0.9551	0.8342
V_2	N_{16}	0.5912	0.9018
V_3	N_8	0.7959	0.8772
V_4	N_4	0.4730	0.8662
V_5	N_{14}	0.5354	0.8905
V_6	N_{10}	0.6501	0.8496
V_7	N_{13}	0.1390	0.8791
V_8	N_{21}	0.7011	0.8822
V_9	N_{11}	0.7579	0.8890
V_{10}	N_{12}	0.3160	0.8674
V_{11}	N_6	0.3146	0.8552
V_{12}	N_{20}	0.7928	0.8943
V_{13}	N_{15}	0.3139	0.8836
V_{14}	N_2	0.7383	0.8753
V_{15}	N_{24}	0.8375	0.8405
V_{16}	N_{25}	0.3050	0.8534
V_{17}	N_{23}	0.5113	0.8513
V_{18}	N_7	0.9697	0.8583
V_{19}	N_1	0.6107	0.8725
V_{20}	N_9	0.9993	0.8457
V_{21}	N_3	0.7693	0.8789
V_{22}	N_5	0.5969	0.8591
V_{23}	N_{19}	0.3713	0.8626
V_{24}	N_{18}	0.9971	0.8836
V_{25}	N_{22}	0.2185	0.8429

Next, we set out to determine if the cost for a pair of aligned phenotypic topics is significantly different than a pair of random phenotypic topics. To do so, we generated 50 random Dirichlet vectors, each of which had the same length and sparsity rate as the learned phenotypic topics. . We then aligned 25 pairs of random topics from the 50 random topics, the cost of which is provided in the final column of Table A1.

We hypothesized that the alignment cost for the learned phenotypic topics would be significantly smaller than that of the random topics. To test this hypothesis, we applied a linear mixed model ((Cost - Random cost) ~ $\alpha+\beta*h$, where h is either 1 (non-random) or 0 (random)) over the costs. The results indicate the phenotypic topic alignment cost was significantly smaller, with $\beta = -0.2574$. Assessing at the two-sided $\alpha=0.05$ significance level, it was observed that the p-value is 4.227×10^{-6}, which validates the hypothesis.

Appendix A6: Regression Results for Transferability of Phenotypic Topics

The results of the regression models (based on Equations 5) are depicted in Table A2. Here, α corresponds to the slope of the regressed line and *I* corresponds to the intercept of the regression parameters.

Table A2. Regression parameters of various phenotype models.

Model	**I**	**α**
ICD-9 codes	-0.06	0.97
PheWAS codes	-0.15	0.90
N-Topic	0.25	1.21
V-Topic	0.47	1.54
N-Topic-Reduced	-0.10	1.03
V-Topic-Reduced	-0.11	1.23