Mining Interpretable and Predictive Diagnosis Codes from Multi-source Electronic Health Records.

Sanjoy Dey [*] Gyorgy Simon [†] Bonnie Westra [‡] Michael Steinbach [*]
Vipin Kumar [*]

Abstract

Mining patterns from electronic health-care records (EHR) can potentially lead to better and more cost-effective treatments. We aim to find the groups of ICD-9 diagnosis codes from EHRs that can predict the improvement of urinary incontinence of home health care (HHC) patients and also are interpretable to domain experts. In this paper, we propose two approaches for increasing the interpretability of the obtained groups of ICD-9 codes. First, we incorporate prior information available from clinical domain knowledge using the clinical classification system (CCS). Second, we incorporate additional types of clinical information for the same patients, such as demographic, behavioral, physiological, and psycho-social variables available from survey questions during the hospital visits. Finally, we develop a hybrid framework that can combine both prior information and the data-driven clinical information in the predictive model framework. Our results obtained from a large-scale EHR data set show that the hybrid framework enhances clinical interpretability as compared to the baseline model obtained from ICD-9 codes only, while achieving almost the same predictive capability.

1 Introduction

Health-care costs in the US are becoming unsustainable, reaching 18% of the gross domestic product (GDP) in 2011 and headed for 20% by 2020 [9]. The terabytes or even petabytes of health data available in EHRs present new opportunities and challenges for research that aims to effectively use these data to discover new knowledge to improve health-care. For example, half of the waste in health care spending (up to $425 billion) has been attributed to a failure of appropriate care delivery, a lack of coordination between different health care plans, and over-treatment [4, 15]. Mining significant patterns from EHRs can help elucidate such knowledge for potential new care plans and enable more coordination between different health care plans.

We collected a large set of EHRs from 581 home healthcare (HHC) agenies for 270,068 patients. In particular, our data contains the diagnosis (ICD-9) codes during patients' admission into HHC. After admission, the patients received several nursing interventions designed to improve their health status. However, all patients are not equally likely to improve in their health status. For example, patients with poor memory are less likely to improve with respect to urinary incontinence. In general, the nursing interventions are designed mostly based on patients' initial health condition during their admission in the homecare agency. However, the original ICD-9 diagnosis codes for which they were admitted in the HHC at first can help to stratify patient groups for more customized homecare interventions and thus, an increased likelihood of improved health status. Finding the important groups of ICD-9 codes is also valuable for enhancing the interpretability of the final models. In this paper, we aim to find the ICD-9 groups that help in improving the health status as measured by urinary incontinence.

Unlike conventional predictive models which mainly focus on improving the predictive power of a target variable such as urinary incontinence, we are primarily interested in finding interpretable risk factors which can be used by the domain expert for further clinical purposes. Moreover, most classification approaches provide only one final set of biomarkers that are applicable for the overall population. Instead, we are interested in finding relatively homogeneous groups of ICD-9 diagnosis codes that are targeted to specific, homogeneous sub-populations. Indeed, this is the main goal of the paper. For example, Table 1 shows two groups of ICD-9 codes. The first group is more interpretable since they are more related and represent the patient group with diabetes and dementia. On the other hand, the second group of ICD-9 codes is not clinically interpretable, although they can be more predictive than the first group.

To find such homogeneous predictive groups of ICD-9 codes, we explored two distinct approaches: data-driven and prior knowledge driven. The data-driven

[*] {sanjoy,steinbac,kumar}@cs.umn.edu, Department of Computer Science, University of Minnesota;

[†] simo0342@umn.edu, Institute of Healthcare Informatics;

[‡] westr006@umn.edu, School of Nursing, **University of Minnesota.**

Group1 ICD-9	Group1 survey features	Group2 ICD-9	Group2 survey features
250.61	Diabetes with neurological manifestation	401.1	Benign hypertension
294.20	Dementia	817	Multiple fractures of hand bones
272.4	hyperlipidemia	692.71	Sunburn

Table 1: Two groups of ICD-9 codes

models incorporate various clinical information such as demographic, behavioral, physiological, and psychosocial factors which were collected as survey question during patient's admission in homecare agencies. The goal of using such survey data is that the auxiliary information collected for same patient will provide more natural groupings of ICD-9 codes. On the other hand, clinical classification software (CCS) provides systematic grouping of ICD-9 codes into a hierarchical tree structure by prior knowledge. We tried to incorporate such prior knowledge into the predictive models.

However, taking such diverse datasets into account creates a number of computational challenges. First, the three datasets (ICD-9 codes, survey questions, CCS prior knowledge) vary in terms of their innate properties such as type, format, and sparsity. Second, the relationships present between the ICD-9 codes and survey questions may be important, although not necessarily discriminative. Therefore, regular predictive models may overlook them. Third, there is a trade-off between data-driven grouping and prior-knowledge driven groupings, which should be taken into account by the model.

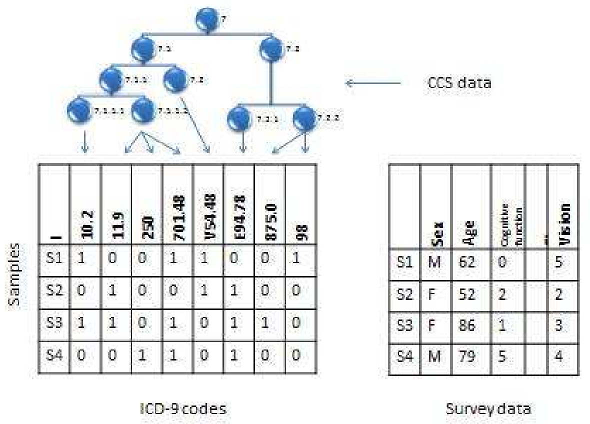
In this paper, we propose an integrative framework to address the above issues in a systematic way. Integration of multiple datasets for biomarker discovery techniques can be broadly classified into two groups: 1) Predictive models ([6], [7] provides a good survey on several kernel fusion methods) and 2) Feature extraction based biomarker discovery techniques [11, 5]. The goal of the predictive model based approaches is to build classification models with high accuracy, but often such techniques do not yield easily interpretable results. In contrast, biomarkers (that are constructed using a small number of features) can be directly useful in diagnosis, treatment or prevention, but equally as important; they can also provide insights into the underlying nature of the disease or related biomedical processes. Hence we focus only on such techniques in this paper to find interpretable ICD-9 code groups.

Among the feature extraction based techniques, canonical correlation analysis (CCA) [8] is one of the most popular technique for data integration, because it can find natural grouping (by components) in each dataset and the potential relationships among those components is measured by correlation. It has been also shown that CCA has fewer model assumptions than other integration techniques [5]. Recently, CCA has been extended for handling high-dimensional data using different type of regularization including sparsity [18]. CCA has further been generalized to integrate more than two datasets [10]. However, none of these methods can take prior knowledge that is available from the CCS tree into account. Moreover, the existing CCA based techniques are unable to hanlde datasets of different types because of their assumptions that all datasets have vector-based records which have been collected for same set of samples.

Our proposed framework further extends the CCA to incorporate the prior knowledge available from the CCS tree into model development which is different than the vector-base data format. Moreover, it can also trade-off between the data-driven knowledge from survey data and prior-knowledge driven CCS framework. We also build a classification model to assess the predictive capability of the obtained components.

1.1 Contributions We aim to find the groups of ICD-9 codes that are responsible for improvement of urinary incontinence. Therefore, we want to build a model that is both predictive and interpretable. To enhance the interpretability of the predictive model, we incorporated several types of knowledge into the model development process as described below:

- To enhance the interpretability of the model, we use the clinical classification system (CCS) as the prior knowledge in the predictive model (baseline model).
- We want to find the relationship between the ICD-9 codes and the clinical survey variables to enhance interpretability. To find such relationships, we use sparse-CCA first to find the relationships present among the two datasets and then use those features along with other useful discriminative features in a predictive model. We further develop a hybrid model called sparse hierarchical CCA (SHCCA) which can take both prior knowledge(CCS) and clinical survey data into account to enhance clinical interpretability.
- To assess the interpretability of the obtained ICD-9 features, we propose a novel metric called I-score



I	10.2	11.9	250	701.48	V54.48	E94.78	875.0	98
S1	1	0	0	1	1	0	0	1
S2	0	1	0	0	1	1	0	0
S3	1	1	0	1	0	1	1	0
S4	0	0	1	1	0	1	0	0

	Sex	Age	Cognitive function	..	Vision
S1	M	62	0		5
S2	F	52	2		2
S3	F	86	1		3
S4	M	79	5		4

Figure 1: SHCCA framework containing three types of data: survey, ICD-9 codes and CCS hierarchy.

based on search on PubMed articles. Our components are more interpretable than the individual ICD-9 and CCS codes with similar prediction capability.

- SHCCA can extract relatively more homogeneous groups of ICD-9 codes, each representing a distinct subgroup of patients, in contrast to finding a global set of ICD-9 codes similar to the baseline predictive models.

2 Method

2.1 The Integrative Predictive Model Framework The main goal of the paper is to utilize as much information available from other clinical domain to enhance the interpretability of the obtained ICD-9 groups without loosing the baseline predictive power of the models. In particular, we want to take two other types of information such as survey data and CCS prior information into account during grouping ICD-9. However, integrating these three type of information poses some computational challenges. First, the datasets are of very different type. For example, ICD-9 contain binary, while clinical factors contain binary, ordinal and numeric data. The ICD-9 codes are very sparse ($< 2\%$ density) in compare to the dense clinical data. Second, there may be some relationships present among the two types of data. For example, a subgroup of patients with mental disorder may have gone through same set of interventions in the homecare. Although these factors may not be discriminative and thus will be missed by the traditional predictive models. Third, CCS hierarchical tree provide completely different type of information containing relationship present among ICD-9 codes only. Moreover, they are not stored in traditional record based datasets similar to ICD-9 and survey data as shown in Figure 1. To address these three challenges, we will first describe how to leverage survey data to group the ICD-9 and then finally taking CCS prior knowledge into account.

2.1.1 Bringing survey data into grouping ICD-9 The easiest way to integrate the ICD-9 and survey data is to concatenate the two datasets together and then build a predictive model such as LASSO as described earlier. However, this will not be able to handle the disparate nature of the two datasets as described earlier. Thus sparse ICD-9 data are more likely to be lost, since the co-efficients of dense survey data will dominate the results. Also, such predictive model will not be able to find relationships present among the types of features. Moreover, such predictive model only focuses on providing one global set of biomarkers. Thus, it cannot provide information about disease heterogeneity, where different set of biomarkers affect different set of population. We want to leverage canonical correlation analysis (CCA) based approach to handle all these issues. Instead of merging the two datasets before performing the analysis, CCA confines in finding components from each of the two datasets such that the components are maximally correlated. This correlation can help find relationships between two datasets. Moreover, each component can correspond to one homogeneous subgroup of the dataset. We used sparse CCA (SCCA) approach for our analysis, because this will perform feature selection from both dataset as well, which will enhance the interpretability. We will describe the SCCA algorithm briefly as follows. Let $\mathbf{X}$ be a $n \times p$ matrix containing p sparse ICD-9 codes and $\mathbf{Y}$ be the $n \times q$ matrix containing q survey questions observed on same n observations. CCA tries to find the linear combination of $\mathbf{X}$ and $\mathbf{Y}$ such that they are maximally correlated. Therefore, we want to find coefficient vectors w_x and w_y from $\mathbf{X}$ and $\mathbf{Y}$ respectively such that

$$(2.1) corr(w'_x X, w'_y Y) = \frac{w'_x C_{XY} w_y}{\sqrt{w'_x C_{XX} w_x}\sqrt{w'_y C_{YY} wy}}$$

is maximized where C_{XX},C_{XY} and C_{YY} are the variance matrix of X, covariance matrix for X and Y and variance matrix of Y, respectively. We can easily see that the correlation is invariant to the any arbitrary scaling of w_x and w_y (by replacing w_x by $a * w_x$). Therefore, equation 2.1 can be re-written as

$$\begin{aligned} &\max_{w_x,w_y} && w_x' C_{XY} w_y \\ &\text{subject to} && w_x' C_{XX} w_x = 1, \\ & && w_y' C_{YY} w_y = 1 \end{aligned} \tag{2.2}$$

Lets define a change of basis $u = C_{XX}^{1/2} w_X$ and $v = C_{YY}^{1/2} w_Y$. Substituting them in equation 2.2, we get

$$\max_{u,v} u' C_{XX}^{-1/2} C_{XY} C_{YY}^{-1/2} v \tag{2.3}$$

such that $u'u = v'v = 1$.

Among many such solutions of equation 2.3, we follow [12], where the solution u and v can be computed as the singular valued decomposition of sample correlation matrix $K = C_{XX}^{-1/2} C_{XY} C_{YY}^{-1/2}$ and then substituting them back on the original basis to get w_x and w_y. However, performing linear combination on all the features of $\mathbf{X}$ and $\mathbf{Y}$ will lead to too many features which will lack biological interpretation. To perform feature selection along with finding the co-efficient vectors, we perform sparse canonical analysis (SCCA) [18] where additional L_1 constraints were imposed on w_x and w_y as below.

$$\begin{aligned} &\max_{w_x,w_y} && w_x' C_{XY} w_y \\ &\text{subject to} && w_x' C_{XX} w_x = 1, \\ & && w_y' C_{YY} w_y = 1, \\ & && \|w_x\|_1 \leq \lambda_x, \\ & && \|w_y\|_1 \leq \lambda_y \end{aligned} \tag{2.4}$$

After the change of basis, the sparseness is imposed on the loading vectors u and v controlled by λ_x and λ_y which determine how many parameters to be selected. The SCCA solutions obtained by integrating icd-9 dataset and the survey data will lead to finding groups of ICD-9 that are not related among themselves but also correlated with the the selected survey data.

2.1.2 Taking CCS Prior Knowledge into Account

In this section, we will describe how bring prior CCS information will lead to more interpretable solutions. Lets consider the example of Figure 1. To take the prior CCS tree structure into account, we need to penalize less for grouping the ICD-9 codes that are closer to each other in the CCS tree. Lets consider H be such a matrix that contain the similarity matrix between each pair of ICD-9 codes to represent the closeness in the CCS tree. Intuitively, this prior knowledge is parallel to the covariance matrix C_{XX} computed from the data as in equation 2.2. Intuitively, we want to tradeoff between these two matrices: the prior knowledge based similarity H and data-driven similarity matrix C_{XX}. The tradeoff is imposed by introducing a new parameter $\lambda_h = [01]$ in equation 2.4. When $\lambda_h = 0$ the solution is exactly equal to those of SCCA, while $\lambda_h = 1$ leads to ICD-9 codes that are purely similar based on the CCS tree H. We will call these as Sparse Hierarchical CCA(SHCCA).

$$\begin{aligned} &\max_{w_x,w_y} && w_x' C_{XY} w_y \\ &\text{subject to} && w_x' [(1-\lambda_h) C_{XX} + \lambda_h H] w_x = 1, \\ & && w_y' C_{YY} w_y = 1, \\ & && \|w_x\|_1 \leq \lambda_x, \\ & && \|w_y\|_1 \leq \lambda_y \end{aligned} \tag{2.5}$$

After the change of basis similar as described in equation 2.4, the solution can be obtained from the new sample correlation matrix $K_h = [(1-\lambda_h)C_{XX} + \lambda_h H]^{-1/2} C_{XY} C_{YY}^{-1/2}$. Note that matrices C_{XX}, C_{YY} and H have to be non-singular, which is ensured by computing them from the data with regularization if required, as mentioned in [3]. In this section we will first describe how to calculate the similarity matrix H from CCS tree followed by the detailed algorithm for computing the solution of equation 2.5.

Computing CCS similarity: The similarity between any two ICD-9 codes was determined based on the depth of their lowest common ancestor(LCA) in the tree. However, some of the ICD-9 codes are not labeled upto the 4th level if the tree(e.g.,875.0 and 95 in Figure 1). Therefore, we normalize that metric by the maximum depth of individual ICD-9 codes. Note that similar type was edge-based similarity measure has also been applied in other biological ontologies such as gene ontology [14]. More formally, it is defined as below:

$$H_{ij} = \frac{depth(LCA(X_i, X_j))}{\max(depth(X_i), depth(X_j))} \tag{2.6}$$

Finding the solution of SHCCA: Finding the solution of SHCCA relies on finding the SVD of the sample correlation matrix K_h approximated by the first singular vectors. We used the two parameter λ_x and λ_y as soft-thresholding parameters to perform feature selection on the datasets X and Y, which is similar to LASSO [13]. In addition, we have the third parameter λ_h which is used to incorporate the prior knowledge measured by H into account. We used an iterative soft-thresholding algorithm for performing SHCCA similar to [13]. This will lead to the first component of u_1 and v_1, where $u_1 = [(1-\lambda_h)C_{XX} + \lambda_h H]^{1/2} w_X$ and $v_1 = C_{YY}^{1/2} w_Y$ from equation 2.5. The second canonical component u_2 and v_2 can be computed such that they are orthogonal to the other components. This can be computed as below from the SVD solution of K_h.

$$K_h = \sum_{i=1}^{k} u_i * d_i * v_i'. \tag{2.7}$$

Therefore, the successive components of u_i and v_i can be computed as the SVD of the remaining sample correlation matrix $\{K_h\}_i = K_h - \sum_{i=1}^{k-1} d_i u_i v_i'$. The algorithm is given in the supplementary section.

3 Experimental Setup

3.1 Dataset: We collected a large EHR data for 270068 patients from 281 home healthcare (HHC) agencies. In particular, we collected 6800 distinct ICD-9 diagnosis codes from HHC for each of those patient as shown in Figure 1. A clinician manually labeled each patient with at most twelve primary and secondary ICD-9 diagnosis codes during his admission in HHC. Moreover, the patients were assessed based on several survey questions related to thier demographic, behavioral, physiological, and psycho-social factors during their admission and discharge in the homecare agencies. These survey questions were summarized into 184 variables guided by domain expert and they were used as an auxiliary dataset for grouping the ICD-9 diagnosis codes. The class label was also created based on whether the urinary incontinence improved during their discharge in compare to the baseline level during admission in HHC. Furthermore, prior information is also available for the ICD-9 codes in the form of clinical classification software(CCS) [2]. CCS has been developed and maintained by Agency for Healthcare Research and Quality (AHRQ) to systematically manage the relationship among several ICD-9 diagnosis codes into a multi-level hierarchical tree. The root contains very generic terms while leaves contain the most specific terms. Therefore, a CCS term is a summarization of several correlated ICD-9 codes. In this paper, we used 4-level tree containing 15073 CCS terms which finally contain all the 6800 ICD-9 codes downloaded from [2].

A few preprocessing steps were performed on the datasets guided by domain experts. For example, the samples with no scope of improvement (highest urinary incontinence score during admission in the homecare) were dropped from the analysis, which led to ultimately 121956 samples. The very rare ICD-9 codes (occurrence in fewer than 10 samples) were removed which ultimately led to 2705 ICD-9 codes. The categorical variables in the survey questions were converted into binary variables, each corresponding to one category. Finally, we ended up with 184 survey questions.

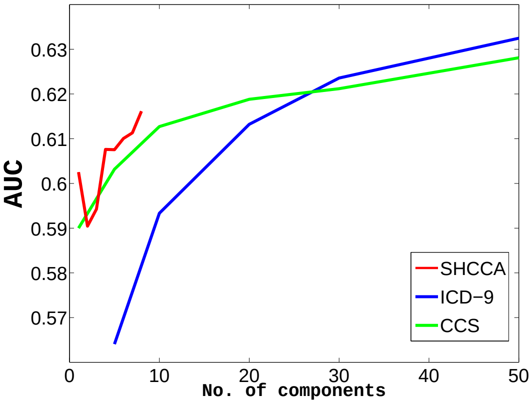


Figure 2: AUC scores for the three methods.

3.2 Evaluation : We evaluated the obtained ICD-9 codes by two metrics: the prediction power and the interpretability. The prediction power was assessed by the area under the ROC curve(AUC) score [16]. We first describe two baseline predictive models which were built on ICD-9 and CCS codes. Then, we describe how the components obtained from SHCCA will be used to build the final predictive model. Finally, we will discuss the techniques for assessing the interpretability of the ICD-9 codes.

3.2.1 Baseline Predictive Models: We created two baseline models for evaluating the prediction power of SHCCA. First, we only considered the basic ICD-9 diagnosis codes which are of lowest level of granularity in the CCS hierarchy. Second, we used all internal nodes of the CCS hierarchy. CCS provides a systematic clusters of the related ICD-9 codes and thus, provides a natural summarization of ICD-9 codes. Therefore, if we build the predictive model on the CCS terms, it can provide more correlated ICD-9 codes. We converted ICD-9 feature space into CCS feature space by taking the most conservative approach. In particular, we created a binary data set with 650 CCS codes (the internal nodes of CCS tree), where we denoted the presence of a CCS code for a particular patient if any of the ICD-9 codes belonging to the subtree rooted at that CCS node was present in that sample. Among different predictive models, we choose LASSO based regularized model [17] because of its inbuilt feature selection technique using L_1 penalty on the coefficients of the solution. Selecting a few most important features in that way will particularly help in interpreting the obtained features by domain expert, which is the main goal of the paper. Furthermore, we used adaptive LASSO [19] to increase the stability of the obtained co-efficients of both ICD-9 and CCS baseline models.

3.2.2 Building Predictive Model on SHCCA Components: To assess the prediction power of the SHCCA components, we first transform the original ICD-9 data into the newly formed K components of ICD-9 codes. We multiply the original data matrix with component vector obtained from SHCCA into a new matrix of $n \times k$, where k is the number of components obtained from SHCCA. To evaluate the prediction power of SHCCA method, we used two cross-validation (CV) frameworks. The external CV was used to find the prediction error of the predictive method built on the SHCCA components using a logistic regression model. For each of the training dataset, a 5-fold internal CV was further used to tune the parameters of SHCCA namely λ_u and λ_v (described later in the parameter selection section). We treated λ_h as an independent parameter since that is not related to the feature selection process from the two datasets.

3.2.3 Assessing Interpretability: The main goal of the paper was to improve the interpretability of the ICD-9 groups obtained from predictive models. Therefore, we evaluated the obtained ICD-9 codes rigorously based on their interpretability. First, the ICD-9 groups were analyzed by domain experts (also the co-author of the paper). The main evaluation criteria was whether the obtained groups of ICD-9 codes is coherent representing similar type of pathology or disease symptoms. Second, we propose a novel measure called I-score to quantify the coherence of the obtained ICD-9 groups using the PubMed articles [1]. In particular, we searched each pair of the terms belonging to same group (or components of SHCCA) for their co-occurrence in the same article. Intuitively, the higher the terms co-occur in PubMed article, the more coherently they represent an underlying disease. Let t_i and t_j be two sets of PubMed articles containing i-th and j-th ICD-9 terms, respectively. Then, a Jaccard similarity measure [16] was defined to assess the semantic similarity of the two terms based on the intersection and the union of two terms. Finally, all such semantic similarities between each possible pairs were summarized as the final similarity of the cluster. Note that the co-occurrence (thus the union of the two terms) of two terms is very rare and therefore, the I-score is low in general.

$$I - score(C) = \sum_i \sum_j (t_i \cap t_j)/(t_i \cup t_j) \tag{3.8}$$

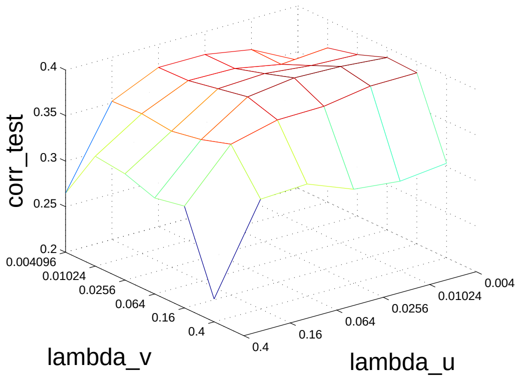


Figure 4: Effect of the sparseness parameter on the average test correlation.

4 Results

Initially, we will show the predictive power of SHCCA components in compare to the two baseline methods built on CCS and ICD-9 codes. Figure 2 represents the area under the ROC curve (AUC) of the three methods for different number of features (components for SHCCA) selected by the LASSO model. Among the three models, ICD-9 provides best overall prediction power. The prediction power of both ICD-9 and CCS model improves when the number of selected features increases. On the other hand, CCS provides better predictive power in the beginning, but saturates as the number of features increases. On the other hand, SHCCA (with best parameter of $\lambda_h = 0.8$, $\lambda_h = 0.16$, and $\lambda_v = 0.0016$) performs slightly better than both of the baseline methods. The AUC score of SHCCA does not vary too much on the λ_h value with range between 0.59 and 0.62 (Supplementary section). It is quiet natural for ICD-9 codes to have the best predictive power, because that is the lowest level of granularity in terms of feature selection and the LASSO model only picks the ICD-9 codes that have best predictive power. However, as the number of features go beyond more than 20, the interpretability of the ICD-9 codes becomes harder since the ICD-9 codes seem very disparate in nature (Supplementary section for full list of ICD-9 codes). In contrast, each of the CCS and SHCCA components represents a cluster of ICD-9 codes, which may not be necessarily best predictive features. However, Figure 2 shows that even those groups are almost equally predictive as the raw ICD-9 codes. Note that the main purpose of this study is to group ICD-9 codes into more interpretable clusters rather than only solely developing predictive model.

The interpretability of the SHCCA method is greatly enhanced in compare to the two baseline methods. Figure 3 represents the interpretability score (I-score) of SHCCA in compare to two baseline methods for $\lambda_h = 0$, i.e., without bringing any prior information. The left subfigure of this figure shows the I-score

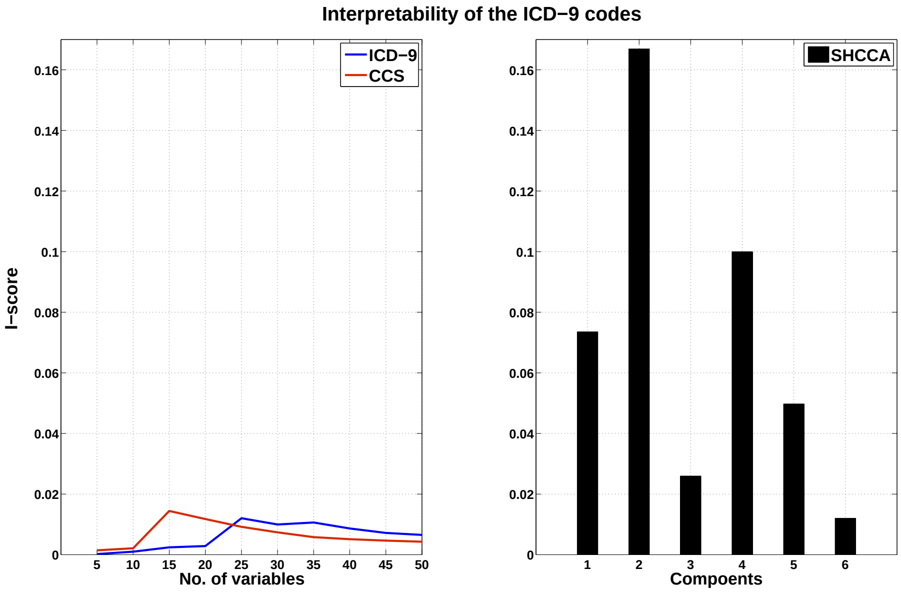


Figure 3: I-score of baseline methods

of the two baseline methods when built by successively adding features into the model. On the other hand, the I-score of each component of SHCCA is shown on the right subfigure. The I-score is greatly enhanced by SHCCA (from 0.015 to 0.165), which shows the effectivity of bringing survey data into account. Note that, PubMed contains a large number of articles containing any of the two terms being searched. However, finding co-occurrence of two disease codes (referred as comorbidity in medical domain) is very rare. Therefore, even though absolute value of the I-score is low in compare to the perfect score of 1, the improvement of 0.15 is very significant. We also examined the ICD-9 groups selected by the two baseline methods (top 20 features with highest LASSO co-efficients are shown for ICD-9 and CCS codes) and the SHCCA as in Table 2 and Table 3, respectively. Then, our domain experts evaluated the ICD-9 obtained from three methods. It turned out that the components selected by SHCCA are more coherent representing one underlying socio-psychological status of the patients. The ICD-9 codes shown in Table 2 represents codes from several diseases such as heart disease, radiological procedure, Alzheimerś disease, paralysis and so on. CCS terms are more interpretable in terms of representing only three major types of disease such as disease related to nervous system, several congenital anomalies, decubitus ulcer, and so on. On the other hand, Table 3 represents three top components from ICD-9 codes and survey data both. The first components represent the ICD-9 codes that are only related to several neurological disorders. More interestingly, the corresponding survey features also exactly matches with socio-psychological functions such as old age, poor cognitive function, speech, prior memory loss, memory deficiency and higher confusions. Similarly, the second component is more related to dysphagia, gastronomy, and blindness which lead to poor self-management skill. The component consists of several aftercare therapies, which is confirmed by prior surgical wound observed in survey data.

We also studied the effect of bringing prior knowledge into SHCCA. Therefore, the λ_h was varied independently between the range $[0 : 0.2 : 1]$, with $\lambda_h = 0$ means no prior information included, while $\lambda_h = 1$ means only prior information is included. We found that $\lambda_h = 0.8$ provided best predictive power as showed in Supplementary Figure 2. We also checked how the interpretability varies with the increment of λ_h, however we only consider the first component for this analysis. It turned out that the I-score remains almost same to 0.075 for all λ_h. However, the size of the components (number of the ICD-9 codes selected) becomes larger as more prior information is included. For example, 37 ICD-9 codes (Supplementary section) were selected for $\lambda_h = 0.6$ in the first component as oppose to only four ICD-9 codes selected when no prior was included (Table 3). Actually, the 37 components comprise most of the ICD-9 codes represented by three subtrees rooted in three CCS level-3 codes representing dementia, transient mental disorders and persistent mental disorders, which are very related mental disorders. Note that, since we computed I-score by all of its pairwise I-score, as the components become larger, it is more likely to have lower I-score. In our cases, bringing prior CCS information provides larger but very coherent ICD-9 without any loss of I-score. Therefore, bringing CCS prior information is important for both increasing the prediction power and interpretability of the SHCCA.

4.1 Effect of the Parameters: We also studied the effect of the sparseness parameters of the two methods. Since, we normalize the canonical vectors u and v in each step of the algorithm, the maximum value that any individual canonical co-efficients can have is 1. Therefore, the maximum value of λ_X and λ_Y is 2. However, we found that if these parameters are set too high(≥ 0.5) no variable selection is performed. Therefore, we searched exponentially within the range of $[0, 0.4]$ for tuning these parameters using a k-fold CV framework as mentioned earlier. For each of the CV run, SHCCA was computed for each combination of the two parameters on the training dataset and then, the obtained co-efficients from the training dataset were used to compute the correlation on the test dataset similar to [18] as defined below:

$$corr = \frac{1}{k}\sum_{j=1}^{k} |cor(X_j u^{-j}, Y_j v^{-j})| \tag{4.9}$$

Here X_j represents the j-th test set and the u^{-j} represents the canonical co-efficients learnt from the corresponding training data. Finally, the test correlation was averaged over the k-fold CV steps and the parameters yielding to largest average correlation was used for building the final predictive model in the outer CV for loop. The average correlation lead to a kind of convex function. In most of the cases, it lead to the $\lambda_u = [0.0016, 0.16]$ for ICD-9 codes and $\lambda_v = [0.0016, 0.1]$ for survey data.

5 Conclusion

In this paper, we incorporated several clinical information available from a survey survey data and prior information to group ICD-9 diagnosis codes into more coherent groups. In particular, we proposed a novel method to incorporate prior information into a sparse hierarchical canonical component analysis. The proposed method enhances the interpretability of ICD-9 codes greatly when assessed by both a novel score call I-score based on search in PubMed articles and clinical interpretation by domain experts. The proposed SHCCA method can further be extended to take the class label into account during method development in our future work. A more systematic score can be also developed to search PubMed articles by mapping ICD-9 code into Mesh terms for assessing interpretability.

Acknowledgement This work was supported by NSF grant IIS-1344135. The first author was also funded by the Doctoral Dissertation Fellowship 2013-14 awarded by the University of Minnesota.

References

[1] http://www.ncbi.nlm.nih.gov/pubmed/advanced.

[2] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ccs/.

[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[4] D. M. Berwick and A. D. Hackbarth. Eliminating waste in us health care. *JAMA: the journal of the American Medical Association*, 307(14):1513–1516, 2012.

[5] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun. Canonical correlation analysis for data fusion and group inferences. *Signal Processing Magazine, IEEE*, 27(4):39–50, 2010.

[6] T. Diethe, D. Hardoon, and J. Shawe-Taylor. Constructing nonlinear discriminants from multiple data views. *Machine Learning and Knowledge Discovery in Databases*, pages 328–343, 2010.

[7] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[9] S. P. Keehan, A. M. Sisko, C. J. Truffer, J. A. Poisal, G. A. Cuckler, A. J. Madison, J. M. Lizonitz, and S. D. Smith. National health spending projections through 2020: economic recovery and reform drive faster spending growth. *Health Affairs*, 30(8):1594–1605, 2011.

[10] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[11] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, and V. Calhoun. Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. *Human brain mapping*, 30(1):241–255, 2009.

[12] K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. 1980.

[13] E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[14] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443, 2009.

[15] R. Steinbrook. Health care and the american recovery and reinvestment act. *New England Journal of Medicine*, 360(11):1057–1060, 2009.

[16] P.-N. Tan et al. *Introduction to data mining.* Pearson Education India, 2007.

ICD-9 terms	CCS terms
Malignant hypertensive heart disease with heart failure	Delirium
Radiological procedure and radiotherapy	Congenital hip deformity
Alzheimer's disease	Other paralysis
Attention to Cystostomy	Decubitus ulcer
Other paralytic syndromes	Other congenital anomalies of urinary system
Neurogenic Bladder Nos	Benign neoplasm of uterus
Senile dementia with delusional or depressive features	Psychogenic disorders
Multiple sclerosis	Other lower gastrointestinal congenital anomalies
Other cerebral degenerations	Other nervous system congenital anomalies
Aftercare following surgery of the genitourinary system	Other aftercare

Table 2: Top 20 features selected by two baseline models based on ICD-9 and CCS terms.

SHCCA survey components-1	SHCCA ICD-9 terms-1	SHCCA survey components-2	SHCCA ICD-9 terms-2	SHCCA survey components-3	SHCCA ICD-9 terms-3
Age	Alzheimer's disease	Poor vision	Legal blindness	Fully granulating surgical wound	Aftercare for healing traumatic fracture of hip
Prior memory loss	Persistent mental disorders	Poor speech	Dysphagia, other	Missing surgical wound	Encounter for change or removal of surgical wound dressing
Poor Speech	Dementias	Worst Speech	Dysphagia		Knee joint replacement
Frequent Behavioral problem	Cerebral degenerations	Partially granulating surgical wound	Degeneration of macula and posterior pole		Hip joint replacement
Poor Cognitive Function		Not healing surgical wound	Non-healing surgical wound		Aftercare following surgery of the musculoskeletal system
Medium Confusion		Average feeding condition	Attention to gastrostomy		Aftercare following joint replacement
High Confusion		Poor feeding condition	Hemiplegia affecting dominant side		Aftercare following surgery for neoplasm
Highest Confusion		Worst feeding condition	Hemiplegia or hemiparesis		Aftercare following surgery of the circulatory system
Memory deficiency					

Table 3: The main three components of SHCCA with $\lambda_h = 0, \lambda_u = 0.3$, and $\lambda_v = 0.3$.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[18] D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[19] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.