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 are also 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 their stay in HHC. 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) agencies for 270,068 patients. In particular, our data contains the diagnosis (ICD-9) codes during patients’ admission into HHC. After admission, the patients received 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. The ICD-9 diagnosis codes recorded during the admission into HHC 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 models incorporate various clinical information such
as demographic, behavioral, physiological, and psycho-social factors which were collected as routine assessment in homecare EHRs through OASIS survey questions\(^1\). 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 using 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 into account prior knowledge that is available from the CCS tree into account. Moreover, the existing CCA based techniques are unable to handle 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 related with the 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.

<table>
<thead>
<tr>
<th>Group1 ICD-9</th>
<th>Group1 survey features</th>
<th>Group2 ICD-9</th>
<th>Group2 survey features</th>
</tr>
</thead>
<tbody>
<tr>
<td>250.61</td>
<td>Diabetes with neurological manifestation</td>
<td>401.1</td>
<td>Benign hypertension</td>
</tr>
<tr>
<td>294.20</td>
<td>Dementia</td>
<td>817</td>
<td>Multiple fractures of hand bones</td>
</tr>
<tr>
<td>272.4</td>
<td>hyperlipidemia</td>
<td>692.71</td>
<td>Sunburn</td>
</tr>
</tbody>
</table>

Table 1: Two groups of ICD-9 codes

---

\(^1\) We will denote these assessments as survey data in rest of the paper.
Figure 1: SHCCA framework containing three types of data: survey, ICD-9 codes and CCS hierarchy.

- To assess the interpretability of the obtained ICD-9 features, we propose a novel metric called I-score 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 as possible 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 types. For example, ICD-9 contain binary, while clinical factors contain binary, ordinal and numeric data. The ICD-9 codes are very sparse (<2% density) compared 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. However, these factors may not be discriminative and thus will be missed by the traditional predictive models. Third, the CCS hierarchical tree provides a completely different type of information containing relationship present among ICD-9 codes. Moreover, they are not stored in traditional record based datasets similar to the ICD-9 and survey data as shown in Figure 1. To address these three challenges, we will first describe how to leverage the survey data to group the ICD-9 codes and then finally take CCS prior knowledge into account.

2.1.1 Bringing survey data into grouping ICD-9 codes

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 coefficients of dense survey data will dominate the results. Also, such a predictive model will not be able to find relationships present among the types of features. Moreover, they only focus on providing one global set of biomarkers. Thus, it cannot provide information about disease heterogeneity, where different set of biomarkers affect different subsets of the population. We want to leverage canonical correlation analysis (CCA) based approaches to handle all these issues. Instead of merging the two datasets before performing the analysis, CCA finds 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 the sparse CCA (SCCA) approach for our analysis, because this will perform feature selection for both datasets as well, which will enhance the interpretability. We will describe the SCCA algorithm briefly as follows. Let \( X \) be a \( n \times p \) matrix containing \( p \) sparse ICD-9 codes and \( Y \) be the \( n \times q \) matrix containing \( q \) survey questions observed on same \( n \) observations. CCA tries to find the linear combination of \( X \) and \( Y \) such that they are maximally correlated. Therefore, we want to find coefficient vectors \( w_x \) and \( w_y \) from \( X \) and \( Y \) respectively such that

\[
\text{corr}(w'_x X, w'_y Y) = \frac{w'_x C_{XY} w_y}{\sqrt{w'_x C_{XX} w_x} \sqrt{w'_y C_{YY} w_y}}
\]

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 \( a \cdot w_x \)).
Therefore, equation 2.1 can be re-written as:

$$\max_{w_x, w_y} \ w_x^T C_{XY} w_y$$

(2.2)

subject to \( w_x^T C_{XX} w_x = 1, \)

\( w_y^T C_{YY} w_y = 1 \)

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:

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

(2.3)

such that \( u'u = v'v = 1 \).

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

$$\max_{w_x, w_y} \ w_x^T C_{XY} w_y$$

(2.4)

subject to \\

\( w_x^T C_{XX} w_x = 1, \)

\( w_y^T C_{YY} w_y = 1, \)

\( \|w_x\|_1 \leq \lambda_x \)

\( \|w_y\|_1 \leq \lambda_y \)

After the change of basis, the sparseness is imposed on the loading vectors \( u \) and \( v \) controlled by \( \lambda_x \) and \( \lambda_y \) which determines how many parameters will 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 the utility of bringing prior CCS information to provide more interpretable solutions. Let us 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. Let \( H \) be a matrix that contains the similarity between each pair of ICD-9 codes in terms of the closeness in the CCS tree. Intuitively, this prior knowledge is parallel to the pair of ICD-9 codes in terms of the closeness in the CCS tree. However, some of the ICD-9 codes are not related among themselves but also correlated with the the selected survey data.

Therefore, equation 2.1 can be re-written as:

$$\max_{w_x, w_y} \ w_x^T C_{XY} w_y$$

(2.2)

subject to \\

\( w_x^T C_{XX} w_x = 1, \)

\( w_y^T C_{YY} w_y = 1 \)

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 this Sparse Hierarchical CCA (SHCCA).

$$\max_{w_x, w_y} \ w_x^T C_{XY} w_y$$

(2.5)

subject to \\

\( w_x^T [(1 - \lambda_h) C_{XX} + \lambda_h H] w_x = 1, \)

\( \|w_x\|_1 \leq \lambda_x, \)

\( \|w_y\|_1 \leq \lambda_y \)

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 up to the 4th level in 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 of edge-based similarity measure has also been applied in other biological ontologies such as the gene ontology [14]. More formally, it is defined as below:

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

(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 \( \lambda_x \) and \( \lambda_y \) 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 \( \lambda_h \) 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 \( w_1 \) and \( v_1 \), where \( w_1 = [(1 - \lambda_h) C_{XX} + \lambda_h H]^{1/2} w_x \) and \( v_1 = C_{YY}^{1/2} w_y \) from equation 2.5. The second canonical components, \( w_2 \) 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 \( K_h \).
(2.7) \[ K_h = \sum_{i=1}^{k} u_i \cdot d_i \cdot v'_i. \]

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 set of EHR data for 270,068 patients from 281 home healthcare (HHC) agencies. In particular, we collected 6,800 distinct ICD-9 diagnosis codes from HHC for those patients 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 their 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 a 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 at discharge compared 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 ICD-9 diagnosis codes as 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 a 4-level tree containing 15,073 CCS terms which finally contain all the 6,800 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 into homecare) were dropped from the analysis, which led to ultimately 121,956 samples. The very rare ICD-9 codes (occurrence in fewer than 10 samples) were removed leading to 2,705 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.

3.2 Evaluation: We evaluated the obtained ICD-9 codes using 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 were used to build the final predictive model. Finally, we 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 at the lowest level of granularity in the CCS hierarchy. Second, we used all internal nodes of the CCS hierarchy. CCS provides a systematic clustering of 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 a 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 this way helps in the interpretation of the obtained features by a domain expert, which is the main goal of the paper. Furthermore, we used adaptive LASSO [19] to increase the stability of the obtained coefficients of both ICD-9 and CCS baseline models.
3.2.2 Building a 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 creating a new matrix of size \( n \times k \), where \( k \) is the number of components obtained from SHCCA. To evaluate the prediction power of the 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 datasets, a 5-fold internal CV was further used to tune the parameters of SHCCA namely \( \lambda_u \) and \( \lambda_v \) (described later in the parameter selection section). We treated \( \lambda_h \) as an independent parameter since it 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 are coherent, representing similar types 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 more frequently 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 the \( i \)-th and \( j \)-th ICD-9 terms, respectively. Then, a Jaccard similarity measure [16] is 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 are 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.

\[
I - \text{score}(C) = \sum_i \sum_j \frac{(t_i \cap t_j)}{(t_i \cup t_j)}
\]

4 Results
Initially, we report the predictive power of SHCCA components compared 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 models 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 \( \lambda_h = 0.8 \), \( \lambda_u = 0.16 \), 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 \( \lambda_h \) value with a range between 0.59 and 0.62 (Supplementary section). It is quite 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 20, the interpretability of the ICD-9 codes becomes less since the ICD-9 codes are 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 solely developing a predictive model.

The interpretability of the SHCCA method is greatly enhanced compared to the two baseline methods. Figure 3 represents the interpretability score (I-score) of SHCCA compared to two baseline methods for \( \lambda_h = 0 \), i.e., without bringing any prior information.

![Figure 4: Effect of the sparseness parameter on the average test correlation.](image-url)
The left subfigure of this figure shows the I-score 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 effectiveness of bringing survey data into account. Note that, PubMed contains a large number of articles for any of the two terms being searched. However, finding co-occurrence of two disease codes (referred as co-morbidity in medical domain) is very rare. Therefore, even though the absolute value of the I-score is low compared 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 coefficients are shown for ICD-9 and CCS codes) and the SHCCA as shown in Table 2 and Table 3, respectively. Then, our domain experts evaluated the results 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 represent codes from several diseases such as heart disease, radiological procedure, Alzheimer’s 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 the 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 match 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 third 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 $\lambda_h$ was varied independently between the range $[0 : 0.2 : 1]$, with $\lambda_h = 0$ indicating no prior information included, while $\lambda_h = 1$ means only prior information is included. We found that $\lambda_h = 0.8$ provided the best predictive power as shown in Supplementary Figure 2. We also checked how the interpretability varies with the increment of $\lambda_h$, but only considered the first component for this analysis. It turned out that the I-score remains almost same to 0.075 for all $\lambda_h$. 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 opposed 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 disorders. Note that, since we
computed the I-score using all of a component’s pairwise I-score, a component is more likely to have lower I-score as the component become larger. In our case, using prior CCS information provides larger but very coherent ICD-9 without any loss of I-score. Therefore, using 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 coefficient can have is 1. Therefore, the maximum value of $\lambda_X$ and $\lambda_Y$ is 2. However, we found that if these parameters are set too high ($\geq 0.5$) no variable selection is performed. Therefore, we searched exponentially within the range of $[0,0.4]$ to tune these parameters using the k-fold CV framework as mentioned earlier. For each CV run, SHCCA was computed for each combination of the two parameters on the training dataset and then, the obtained coefficients from the training dataset were used to compute the correlation on the test dataset as defined below [18].

\begin{equation}
\text{corr} = \frac{1}{k} \sum_{j=1}^{k} [\text{cor}(X_j u^{-j}, Y_j v^{-j})]
\end{equation}

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

5 Conclusion

In this paper, we incorporated 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 (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 codes 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

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

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

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

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