gamAID: Greedy CP Tensor Decomposition for Supervised EHR-based Disease Trajectory Differentiation

Jette Henderson[1] Joyce Ho[2] Joydeep Ghosh[3]

Abstract— **We propose gamAID, an exploratory, supervised nonnegative tensor factorization method that iteratively extracts phenotypes from tensors constructed from medical count data. Using data from diabetic patients who later on get diagnosed with chronic kidney disorder (CKD) as well as diabetic patients who do not receive a CKD diagnosis, we demonstrate the potential of gamAID to discover phenotypes that characterize patients who are at risk for developing a disease.**

I. Introduction

Diabetes can cause kidney damage with varying degrees of severity. This damage, called diabetic nephropathy, is a type of Chronic Kidney Disorder (CKD) and is found in 23% of diabetes patients. The presence of both CKD and diabetes in a patient can result in complications of care. For example, reduced kidney function inhibits the amount of insulin the kidneys can remove from a person's blood, which makes the process of controlling a diabetic patient's glycemic levels more challenging. Being able to identify early signs of CKD in diabetes patients can help mitigate complications of simultaneously managing diabetes and CKD [1].

In this work, we propose gamAID, an exploratory, supervised method for separating diabetes patients into two groups based on their risk of developing diabetic nephropathy. Specifically, gamAID greedily extracts and accumulates phenotypes for each group using nonnegative tensor decomposition on patient data contained in Electronic Health Records (EHRs). While we apply this framework specifically to patients with diabetes, gamAID is general enough to be applied to other sets of conditions. We use gamAID to explore the feasibility of identifying a patient group that diverges from a population of similarly ill patients and compare our method to Fisher's Linear Discriminant Analysis, another supervised method. Our results demonstrate the potential of gamAID to characterize patients who are risk of developing diabetic nephropathy. More broadly, it highlights the advantages of tensor-based analysis vs. vector-based representations of complex medical data.

II. Preliminaries

Our method uses nonnegative tensor decomposition to greedily discover phenotypes within a population of patients represented by their EHRs. gamAID is supervised in that it performs tensor factorization based on the future CKD status of the patient. First, we briefly describe tensors and the process of decomposing them but refer the reader to [2] for a more thorough introduction. A tensor is an n-way or n-mode array that is used to represent n-dimensional relationships. In this exploratory work, we focus on 3-mode tensors, with patients, diagnoses, and procedures as our three modes. Each element captures the multidimensional relationship of the number of times a patient has experienced a medical diagnosis and procedure in a given period of time. We note that our work can be extended to higher dimensions (e.g., patient, diagnosis, procedure, and time).

Much like, but not identical to matrices and their factorizations, tensors can be decomposed or factored into a product of matrices or a combination of matrices and smaller tensors or as the sum of tensors. There are multiple tensor decomposition models, but we focus on the CANDECOMP / PARAFAC (CP) decomposition [3], [4].We first define the notion of a rank one tensor.

Definition 1: A tensor $\boldsymbol{\mathcal{W}}$ is an N-way rank one tensor if it can be written as the outer product of N vectors, $\mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(N)}$, where each element $w_{\vec{i}} = x_{i_1,i_2,\cdots,i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \cdots a^{(N)}_{i_N}$.
CP decomposition factorizes the original tensor $\boldsymbol{\mathcal{X}}$ as a sum of R rank one tensors and can be expressed as follows:

$$\boldsymbol{\mathcal{X}} \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \ldots \circ \mathbf{a}_r^{(N)} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!]. \quad (1)$$

The representation $[\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \ldots; \mathbf{A}^{(N)}]\!]$ is shorthand notation with the weight vector $\boldsymbol{\lambda} = [\lambda_1 \cdots \lambda_R]$ and the factor matrix $\mathbf{A}^{(n)} = [\mathbf{a}_1^{(n)} \cdots \mathbf{a}_R^{(n)}]$, where $\mathbf{a}_r$ denotes the r^{th} column of $\mathbf{A}^{(n)}$. CP decompositions are usually fit using a loss function that makes assumptions about the underlying distribution of the data contained in the tensor (e.g., Gaussian or Poisson). In general, tensor factorization utilizes information in the multiway structure to produce factors that are concise and potentially more interpretable, and it is also able to identify components even with relatively small amounts of observations [2].

When CP decomposition is applied to Electronic Health Records (EHRs), each rank one tensor can be thought of a medical phenotype, where a phenotype is a set of clinically meaningful characteristics used to describe a group of patients. Using a Poisson assumption, Ho et al. showed CP decomposition has the potential to produce clinically relevant phenotypes in a high-throughput, unsupervised manner [5],

[1] J. Henderson is the corresponding author. She is with the Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX 78712, USA. Email: `jette@ices.utexas.edu`

[2] J.C. Ho is with the Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA. Email: `joyce.c.ho@emory.edu`

[3] J. Ghosh is the Schlumberger Centennial Chair Professor with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA. Email: `jghosh@utexas.edu`

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

[6]. Others have used a Gaussian assumption on the data to decompose tensors into sets of rank one phenotypes, which can be solved using an alternating least squares approach. For example, Wang et al. used a CP decomposition fit on mean–squared loss and orthogonality constraints to produce diverse sets of phenotypes [7]. We draw on the work of [5], [6] because it uses a Poisson assumption appropriate to count data (e.g., diagnosis/procedure counts) and carries out the decomposition via the non–negative CP alternating Poisson regression (CP-APR) model developed in [8]. However, instead of fitting all R rank-one tensors simultaneously, we build up the decomposition by fitting rank one tensors one at a time, setting them as constants, and then fitting the next rank one tensor until we have reached the desired rank.

III. gamAID

Our algorithm, Greedy Angular Multiway Array Iterative Decomposition (gamAID), is an exploratory, supervised non–negative tensor factorization method for uncovering distinctive phenotypes that can differentiate patients with or without a disease. Our goal is to accumulate computational phenotypes that are representative of each patient class that are "different" from phenotypes discovered in the other patient class. Given the binary labels representing whether or not a diabetic patient is diagnosed with CKD in the year following the observed records, we construct three types of tensors to which gamAID will fit decompositions. The first tensor, $\boldsymbol{\mathcal{X}}_{(01)}$ contains EHR count data from both classes of patients. We then split $\boldsymbol{\mathcal{X}}_{(01)}$ along the patient mode to form $\boldsymbol{\mathcal{X}}_{(0)}$ and $\boldsymbol{\mathcal{X}}_{(1)}$ so that they only count data specific to the class in question (i.e., class 0 or class 1). gamAID fits one of three tensors $\boldsymbol{\mathcal{Z}}_{(01)}$, $\boldsymbol{\mathcal{Z}}_{(0)}$, and $\boldsymbol{\mathcal{Z}}_{(1)}$ based on what step it is in. These fit tensors are the same size as their respective observation tensor, and each element $z_{\vec{i}}$ contains the optimal Poisson parameter for the observed tensor $x_{\vec{i}}$. We constrain the fit tensors to share all but one of the same factor vectors along the non–patient factors (i.e., diagnosis, procedure). Thus, the discovered phenotypes can be used to uncover higher–order interactions, which can then be used as distinguishing characteristics for improved prediction and understanding. Given the patient classes are similar to one another, the decomposition $\boldsymbol{\mathcal{Z}}_{(01)}$, fit on $\boldsymbol{\mathcal{X}}_{(01)}$, captures some features held in common between the two classes.

gamAID introduces the use of an angular constraint to encourage diversity between factor vectors of each mode by penalizing any pair of vectors that are "similar" to previous discovered phenotypes. The algorithm represents similarity between two factors via the cosine similarity, $\frac{\mathbf{a}^\intercal \mathbf{b}}{||\mathbf{a}||_2||\mathbf{b}||_2}$. Under this measure, two vectors $\mathbf{a}, \mathbf{b}$ that are orthogonal will have a score of 0, while two exact same vectors will have a score of 1. We add this angular regularization term, Equation (3), to the objective function of a count tensor decomposition which uses KL divergence, Equation (2). It is important to note that since $\boldsymbol{\mathcal{X}}_{(01)}$ consists of count data, it is not possible to standardize the tensor by subtracting off the mean and dividing by the standard deviation. Thus, a bias term, $\mathbf{u}^{(n)}$, is added in Equation (4) to capture the baseline state of the data. Each factor matrix $\mathbf{A}^{(n)}$ can be projected onto a sparse simplex denoted by s (shown in Equation (5)), which provides a tunable parameter to alter the number of elements in the resulting factors. The optimization problem that is solved for each separate tensor $\boldsymbol{\mathcal{X}}_{(d)}$, where $d \in \{0, 1, 01\}$, is the following:

$$f(\boldsymbol{\mathcal{Z}}_{(d)}) = \min(\sum_{\vec{i}} (z_{\vec{i}(d)} - x_{\vec{i}(d)} \log z_{\vec{i}(d)}) \tag{2}$$

$$+ \frac{\beta}{2} \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{p=1}^{r} (\frac{(\mathbf{a}_p^{(n)})^\intercal \mathbf{a}_r^{(n)}}{||\mathbf{a}_p^{(n)}||_2 ||\mathbf{a}_r^{(n)}||_2})^2 \tag{3}$$

$$\text{s.t } \boldsymbol{\mathcal{Z}}_{(d)} = [\![\sigma; \mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!] + [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \cdots; \mathbf{A}^{(N)}]\!] \tag{4}$$

$$||\mathbf{a}_r^{(n)}||_1 = s, \; 0 < s \leq 1, \; \forall n \tag{5}$$

$$||\mathbf{u}^{(n)}||_1 = 1, \; \forall n. \tag{6}$$

This is a complex objective that cannot be solved exactly. Instead, gamAID uses a greedy algorithm to iteratively build up a tensor decomposition of size R by fitting rank one tensors only using one class at a time. The algorithm fits a rank one tensor that is "best fit" relative to the class and the rank-one tensors we have already accumulated. The first step is to fit the best rank one tensor $\boldsymbol{\mathcal{Z}}_{(01)}$ to $\boldsymbol{\mathcal{X}}_{(01)}$ based on the optimization problem described above ($\boldsymbol{\mathcal{Z}}_{(01)}$ $= \lambda_1 \mathbf{a}_1^{(1)} \circ \mathbf{a}_1^{(2)} \circ \mathbf{a}_1^{(3)}$). We then choose one class, $\boldsymbol{\mathcal{X}}_{(1)}$, and minimize the optimization problem with respect to $\boldsymbol{\mathcal{X}}_{(1)}$ to fit a rank-two decomposition, with the first rank one tensor set to the one learned in the previous steps ($\boldsymbol{\mathcal{Z}}_{(1)} = \lambda_1 \mathbf{a}_1^{(1)} \circ \mathbf{a}_1^{(2)} \circ \mathbf{a}_1^{(3)} + \lambda_2 \mathbf{a}_2^{(1)} \circ \mathbf{a}_2^{(2)} \circ \mathbf{a}_2^{(3)}$). gamAID then switches classes and minimizes the optimization problem with respect to $\boldsymbol{\mathcal{X}}_{(0)}$ to fit a rank-three decomposition based on the two rank one tensors learned previously. gamAID continues to switch classes until the user-specified number of phenotypes R. The patient mode for each class needs to be refit each time as the membership to phenotypes might be redistributed for a given patient, given a new set of phenotypes. The pseudocode for the algorithm is shown in Algorithm 1.

At the end of the gamAID process, the diagnosis and procedure modes are fixed and the combined patient factor matrix is learned by minimizing the objective function once more. The final step is to normalize across the rows of the patient factor matrix. We can interpret the normalized values as a patient's membership to or loading on a phenotype.

IV. Data

We demonstrate the potential of the gamAID framework on the publicly available *CMS Linkable 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)* that the Centers for Medicare and Medicaid Services (CMS) provides.[1] It contains claim records spanning 3 years of data. The records have been synthesized from 5% of the 2008 Medicare population to protect the privacy of the patients.

[1] The dataset can be downloaded at `https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/index.html`

Algorithm 1: Pseudocode for the gamAID algorithm

```
Data: $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}_{(1)}, \boldsymbol{\mathcal{X}}_{(0)}, K$
Result: $[\![\sigma; \mathbf{u}^{(1)}; \cdots; \mathbf{u}^{(N)}]\!], [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}; \mathbf{A}^{(2)}; \mathbf{A}^{(3)}]\!]$
for $r = 1, 2, \cdots, R$ do
    if $r == 1$ then
        Solve the optimization problem (Equations (2)-(6)) for $\boldsymbol{\mathcal{X}}_{(01)}$, the tensor corresponding to both class 0 and class 1 patients
    end
    if $r$ is even then
        Solve the optimization problem (Equations (2)-(6)) for $\boldsymbol{\mathcal{X}}_{(1)}$, the tensor corresponding to class 1 patients
    end
    if $r$ is odd and $r > 1$ then
        Solve the optimization problem (Equations (2)-(6)) for $\boldsymbol{\mathcal{X}}_{(0)}$, the tensor corresponding to class 0 patients
    end
end
```

DE-SynPUF contains inpatient, outpatient, carrier, and prescription drug event claims in addition to the beneficiary files. Although the relationships between some of the variables have been altered to minimize re-identification risk, due to the very large size and coverage of the data, the conclusions obtained by population level models are expected to closely represent those obtained from the unaltered dataset, and thus still provided very valuable clinical insights.

We extracted two classes based on values for different disease flags in the Beneficiary file. Class 1 consists of patients flagged as diabetic in 2009 and 2010, who did not have a chronic kidney disease (CKD) flag in 2009 but did have a CKD flag in 2010. We also refer to this class as "diabetes-CKD." Class 0, which we also refer to as "diabetes only," consists of patients with a diabetes flag in 2009 and 2010 and no CKD flag in 2009 or 2010.

The extracted cohort consists of $1,492$ diabetes-CKD patients and $1,625$ diabetes-only patients. Figure 1 shows the difference between the diagnosis counts between diabetes–only and diabetes–CKD patients. For reference, the negative side of the x-axis are diagnoses that appeared more in diabetes–CKD than diabetes–only patients. Our analysis also showed that some diagnoses appear in one class but not the other. To build our patient×diagnosis×procedure tensor, we use the 50 diagnosis with the highest counts for each class as well as the diagnoses that appeared much more in one class than the other. We included all procedures associated with these diagnoses.

V. Results

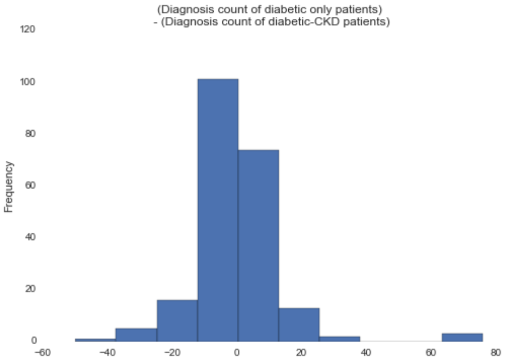


Fig. 1. Histogram of difference between diagnosis counts between classes.

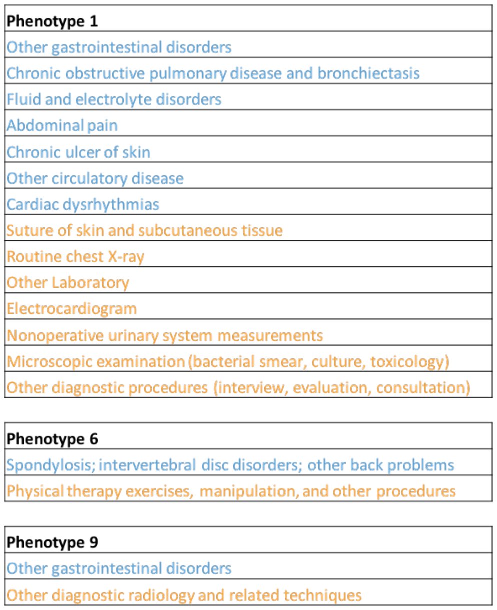
TABLE I
Percentages of Class Membership by Phenotype

Phenotype	% Class 1	% Class 0	% Population Captured
1	0.52	0.48	0.08
2	0.49	0.51	0.80
3	0.48	0.52	0.10
4	0.48	0.52	0.21
5	0.48	0.52	0.17
6	0.54	0.46	0.09
7	0.00	0.00	0.00
8	0.48	0.52	0.08
9	0.62	0.38	0.01

We used the gamAID process to accumulate 9 phenotypes from $\boldsymbol{\mathcal{X}}_{(01)}$ (to fit phenotype 1), $\boldsymbol{\mathcal{X}}_{(1)}$ (to fit phenotypes 2, 4, 6, 8), and $\boldsymbol{\mathcal{X}}_{(0)}$ (to fit phenotypes 3, 5, 7, 9). After finishing the gamAID process, we fixed the diagnosis and procedure modes and fit the patient mode to learn the membership of the patients across the phenotypes. Table I shows the percentage of patients by class per phenotype and the percentage of patients the phenotype captured overall. Interestingly, phenotypes not fit on one class are dominated by that class (e.g., phenotypes 4 is mostly diabetes-only patients though it was fit on diabetes-CKD patients).

This implies that the patients in both classes are quite similar, which makes intuitive sense. Figure 2 depicts a selection of phenotypes for which diabetes-CKD patients were the majority. In the future, we plan to consult domain experts on the clinical relevance of the extracted phenotypes, but based on a literature search, many of the elements of the phenotypes in diabetes-CKD majority phenotypes have been documented as being related to chronic kidney disease. For example, gastrointestinal disorders (phenotypes 1 and 9), heart dysrhythmias (phenotype 1), and abdominal pain (phenotype 1) are commonly found in patients with chronic kidney disorder [9], [10], [11]. Additionally, back issues (phenotype 6) are a symptom of chronic kidney disorder [12]. While we can say nothing about causation, it is interesting to see that these phenotypic elements were present in the diabetic-CKD majority phenotypes.

In comparison, we applied Fisher's Linear Discriminant Analysis (LDA) to a matricized $\boldsymbol{\mathcal{X}}_{(01)}$ and to the first 30 components of a PCA decomposition of the matricized $\boldsymbol{\mathcal{X}}_{(01)}$[13], [14]. We then used 5-fold cross-validation to fit the projected vector. Figure 3 shows a distribution of observations projected onto the linear discriminant. When Fisher's LDA is fit on the raw matricized tensor (top left), it appears there is good separation between the classes. However, when



Phenotype 1
Other gastrointestinal disorders
Chronic obstructive pulmonary disease and bronchiectasis
Fluid and electrolyte disorders
Abdominal pain
Chronic ulcer of skin
Other circulatory disease
Cardiac dysrhythmias
Suture of skin and subcutaneous tissue
Routine chest X-ray
Other Laboratory
Electrocardiogram
Nonoperative urinary system measurements
Microscopic examination (bacterial smear, culture, toxicology)
Other diagnostic procedures (interview, evaluation, consultation)

Phenotype 6
Spondylosis; intervertebral disc disorders; other back problems
Physical therapy exercises, manipulation, and other procedures

Phenotype 9
Other gastrointestinal disorders
Other diagnostic radiology and related techniques

Fig. 2. A subset of phenotypes resulting from the gamAID process.

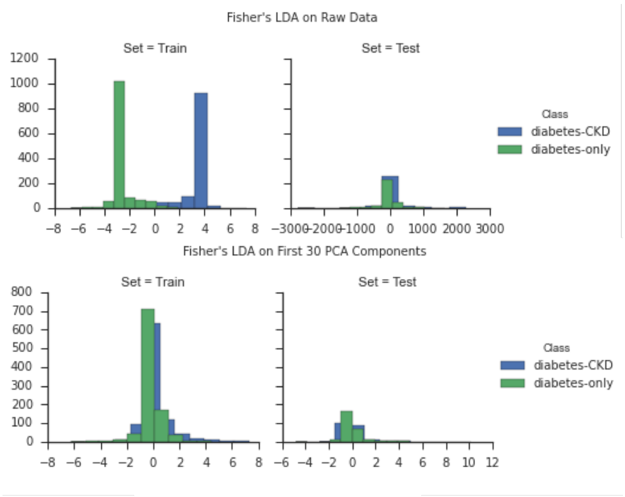


Fig. 3. LDA distribution of projected data (raw and first thirty components of data transformed by PCA)

applied to the test set (top right), the separation quickly disappears, which suggests overfitting. The training and test distributions of Fisher's LDA applied to the first 30 PCA components look similar (bottom left and right, respectively), but the overlap of the two classes suggests the classes are difficult to separate. Finally, we used the linear discriminant to predict the classes of the test set. This resulted in an average f1-score of .4783 on the raw tensor and .3914 on the PCA components of the tensor. In contrast, a SVM model trained on top of the gamAID decomposition resulted in an average f1-score of .5106. Thus, while this is a difficult classification problem, gamAID shows an improvement over other methods.

VI. Conclusions and Future Work

We presented an exploratory greedy, iterative approach called gamAID that extracts phenotypes in a supervised manner from a population consisting of diabetes patients without CKD and diabetes patients who will transition to a CKD diagnosis in the future. We showed that this method has the potential to tease out phenotypes of diverging disease populations and paired with a simple classifier can identify patients at-risk for CKD. In the future, we would like to continue exploring and improving gamAID. One possible way to improve gamAID is to tune the sparsity parameters to capture more of the population as well as increasing the final rank of the gamAID produced tensor. Additionally, we intend to study the effects of fitting the second phenotype using class 0 instead of class 1. We also plan to test this framework on other sets of diseases (e.g., diabetes and hypertension) with the possibility of expanding beyond pairs of diseases.

References

[1] K. L. Cavanaugh, "Diabetes management issues for patients with chronic kidney disease," *Clinical Diabetes*, vol. 25, no. 3, pp. 90–97, 2007. [Online]. Available: http://clinical.diabetesjournals.org/content/25/3/90

[2] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[3] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[4] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[5] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, "Limestone: High-throughput candidate phenotype generation via tensor factorization." *Journal of Biomedical Informatics*, vol. 52, pp. 199–211, Dec. 2014.

[6] J. C. Ho, J. Ghosh, and J. Sun, "Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 115–124.

[7] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun, "Rubik: Knowledge guided tensor factorization and completion for health data analytics," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1265–1274.

[8] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1272–1299, 2012.

[9] A. De Francisco, "Gastrointestinal disease and the kidney." *European journal of gastroenterology & hepatology*, vol. 14, pp. S11–5, 2002.

[10] K. Stadler, I. J. Goldberg, and K. Susztak, "The evolving understanding of the contribution of lipid metabolism to diabetic kidney disease," *Current diabetes reports*, vol. 15, no. 7, pp. 1–8, 2015.

[11] G. Boriani, I. Savelieva, G.-A. Dan, J. C. Deharo, C. Ferro, C. W. Israel, D. A. Lane, G. La Manna, J. Morton, A. M. Mitjans *et al.*, "Chronic kidney disease in patients with cardiac rhythm disturbances or implantable electrical devices: clinical significance and implications for decision making-a position paper of the european heart rhythm association endorsed by the heart rhythm society and the asia pacific heart rhythm society," *Europace*, vol. 17, no. 8, pp. 1169–1196, 2015.

[12] R. D. Hays, J. D. Kallich, D. L. Mapes, S. J. Coons, and W. B. Carter, "Development of the kidney disease quality of life (kdqol?) instrument," *Quality of Life Research*, pp. 329–338, 1994.

[13] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[14] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.