Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **2**, Number **3&4**, July & October **2017**

doi:10.3934/bdia.2017014

pp. **209–217**

PREDICTION MODELS FOR BURDEN OF CAREGIVERS APPLYING DATA MINING TECHNIQUES

Sunmoo Yoon*

School of Nursing, Columbia University Medical Center
New York, NY, 10032, USA

Maria Patrao and Debbie Schauer

Department of Nursing, Columbia University Medical Center
New York, NY, 10032, USA

Jose Gutierrez

Department of Neurology, Columbia University Medical Center
New York, NY, 10032, USA

Abstract. **I**ntroduction: Caregiver stress negatively influences both patients and caregivers. Predictors of caregiver difficulty may provide crucial insights for providers to prioritize those with the highest risk of stress. The purpose of this study was to develop a prediction model of caregiver difficulty by applying data mining techniques to a national behavioral risk factor data set.

Methods: Behavioral data including 397 variables on 2,264 informal caregivers, who provided any care to a friend or family member during the past month, were extracted from a publicly available national dataset in the U.S (N = 451,075) and analyzed. We applied several classification algorithms (J48, RandomForest, MultilayerPerceptron, AdaboostM1), to iteratively generate prediction models for caregiving difficulty with 10-fold cross validation.

Results: 44.7% of informal caregivers answered that they faced the greatest difficulties while they took care of patients. Among those who faced the greatest difficulties, the reasons were creating emotional burden (45%). Patient cognitive alteration (e.g. cognitive changes in thinking or remembering during the past year), care hours, and relationship with a caregiver appeared as the main predictors of caregiver stress (classified correctly 63%, difficulty AUC = 65%, no difficulty AUC = 65%).

Conclusions: Data mining methods were useful to discover new behavioral risk knowledge and to visualize predictors of caregiver stress from a multidimensional behavioral dataset.This study suggests that health professionals target dementia family caregivers who are anticipated to experience patients neuro-cognitive changes, and inform the caregivers about importance of limiting care hours, burn out and delegation of caregiving tasks.

1. **Introduction.** Americans live longer but are sicker [17]. Many patients with chronic disease such as stroke, cancer, heart attack, or dementia live at home after the acute phase of their disease. Their physical, psychosocial, and cognitive impairments are significant challenges to family caregivers [1]. Supporting caregivers are a critical element of health care. Whereas 'identifying risks and management

2010 *Mathematics Subject Classification.* Primary: 97R50; Secondary: 97R71.

Key words and phrases. Data mining, neural network, visualization, caregiver, dementia, stress.

*Corresponding author: Sunmoo Yoon, RN, PhD, Associate Research Scientist, Columbia University, sy2102@columbia.edu.

of family caregivers appears frequently in a nursing diagnosis text book, in reality those concepts are rarely translated during patient care. In spite of the amount of family caregivers demands, providers are often forced to rush in discharging patients for management reasons. The unmet needs of family caregivers led to an increased patient readmission rate [6].

Studies have reported that caregiver stress negatively impacts both patients and caregivers [1]. For patients, caregiver stress is associated with admission to nursing home [19]. Moreover, patients at home have often been victimized by domestic violence [23]. For caregivers, their stress altered the capacity of proinflammatory cytokines inducing anxiety, fatigue, sleep alteration, and increased sensitivity to pain. In addition, caregiver stress affected hippocampal dependent cognitive function, suppressed neurogenesis, and caused dendrite shrinkage leading depressive illness and cognitive decline [13]. Accordingly, several studies have reported the importance of interventions to support caregivers [8].

Although most providers are able to foresee the burden of patient care for families, providing intervention for family caregivers may prove challenging in a hectic clinical environment. Therefore, predictors of caregiver difficulty may provide crucial insights for providers to prioritize those with the highest risk of stress. Predictors of caregiver burden have rarely been explored. In addition, as the proven accuracy of novel data science methods in the 2016 U.S. presidential election, data mining methods may overcome the limitation of traditional statistical methodologies (e.g. multivariate regression) which cannot analyze hundreds of variables at once [2, 18, 21]. Thus, the purpose of this study was to develop a prediction model of caregiver difficulty by applying data mining techniques to a national behavioral risk factor data set.

2. Method.

2.1. Data and tool.

Our observational study used a national data set, the Behavioral Risk Factor Surveillance System (BRFSS) by Centers for Disease Control and Prevention [7]. The BRFSS is a publicly available dataset which is an annual ongoing health survey tracking health conditions and risk behaviors in the United States. PSAW SPSS©Version 18 was used for extracting records of caregivers, cleaning, and conversion of the data file format 'XTP' to 'csv', which is readable in Weka software [11, 14]. Weka is an open source data mining software used in the data mining and knowledge discovery process. Weka V 3.7 was used for our data analysis to build the predictive models and to assist in adequate selection of the independent variables for prediction modeling.

2.2. Outcome.

We extracted data on 2,264 caregivers who provided any care to a friend or family member during the past month from BRFSS (N = 451,075). We were interested in investigating caregiver difficulty. BRFSS asked caregivers their greatest difficulty they have faced as a caregiver; whether or not they experienced such difficulties as financial burden, time pressure for themselves for their families, interfering with their work, emotional stress, aggravating health problems, and the affect on family relationships. Caregiver difficulty variable was dichomized as no difficulty and difficulty as our outcome variable. "Dont know/not sure" or "refused to answer" the caregiver difficulty questionnaire were treated as missing values because the portion was less than 5% (1.28%).

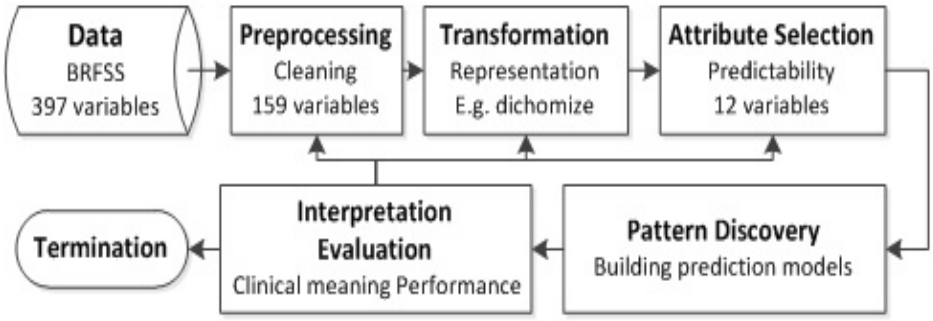

FIGURE 1. Iterativesteps of the data mining process to build a prediction model from a large dataset

2.3. **Analysis.** In order to build a prediction model for caregiver difficulty, we followed the iterative steps of data mining process which consists of preprocessing, transformation, attribute selection, pattern discovery and interpretation (Figure 1).

Preprocess: First, we reduced the number of variables by removing redundant or irrelevant variables (e.g., phone number, disaster preparation, dental cleaning) leaving 159 variables from 397 variables. Next, we used CFS attribute evaluator, a machine learning algorithm, to select variables that were strongly related to the outcome variable; this resulted in 12 predictive variables including relationship with a caregiving patient, years of caregiving, hours of caregiving per week, patient status change, age of caregiver, quality of rest, emotional support, satisfaction with life, diabetes, and patient needs.

Prediction Model: In order to avoid algorithm dependency, we applied several classification algorithms (J48, RandomForest, MultilayerPerceptron, AdaboostM1), to iteratively generate models. C4.5 and Adaboost (J48 and Adaboost M1 in Weka) built based on an accurate sound theory are selected as top 10 data mining algorithm among experts [21, 22]. Deep learning and neural network (MultilayerPerceptron in Weka) is known as a powerful technique,theoretically well suited to non-linear processes like complex stress outcome [18, 3, 12, 24] However, the model by neural network is not transparent [16]. In fact, the model by neural network is technically difficult to communicate and visualize due to its hidden layer. We also chose an ensemble classifier, Random Forest algorithm which is known to be accurate and efficient on large data base [4].

Validation: For model validation, Weka software allows us to randomly partition a dataset for training and testing. We applied 10-fold cross validation, meaning 90% of cases to be a training set and 10% of cases to be a validating set. Once the prediction model was generated from the training set, then it was validated on the testing dataset. During the iterative modeling process, we evaluated the models performance each time using proportion correctly classified and the area of under the receiver operating characteristic curve (AUC). We selected a final model based on predictive ability and clinical meaningfulness of variables. Models were visualized in a simple tree form to enhance communication with providers at bedside.

3. Results.

3.1. Characteristics of study population.

Characteristics of caregivers (n = 2,264) were described in Table 1 and 2. While mean age of patients whom the caregivers took care of was 70 (SD = 20.5), mean age of caregivers was 56 (SD = 15.5). Most of caregivers (91%) were white and women (64%). More than half of them were employed for wages or self-employed,and about a fourth were retired (26%). Approximately 40% of caregivers made less than $50,000 per year. Forty percent of caregivers took care of their parents or grandparents, 22% of siblings or child, 16% spouses, and 16% friends. In more than half of the cases, cognitive status of patients (e.g. worse in remembering, decision making) had changed during the past year. Seventy four percent of caregivers have been taking care of patients for less than 5 years. Most of them spent less than 30 hours for care for the patients per week.The most need of patients were 1) taking care cleaning, managing money, or preparing meals (27%), transportation outside of the home (22%), self-care including eating, dressing, bathing (13%), and miscellaneous including communicating with others, moving around within the home, seeing or hearing, getting along with people (14%) and relieving anxiety or depression (8%). Whereas fifty four percent of caregivers answered that they did not experience the greatest difficulty as a caregiver, 44.7% of them answered that they faced the greatest difficulties while they took care of patients. Among those who faced the greatest difficulties, the reasons were creating emotional burden (45%), not enough time for themselves (14%), creating financial burden (8%), altering family relationships (7%), interfering with their work (6%), creating or aggravating their health problems (3%), or others difficulties (10%).

Patient cognitive alteration (e.g. cognitive changes in thinking or remembering during the past year), care hours, and relationship with a caregiver appeared as the main predictors of caregiver stress (classified correctly 63%, difficulty AUC = 65%, no difficulty AUC = 65%). More than half of the patients experienced cognitive changes. When a patient did not experience changes in thinking or remembering, caregivers were more likely to have less difficulty. Most of them (60%) care for less than 14 hours per week. Among those (¡14 hrs/week), if a patient was a parent, a child, a spouse or sibling, caregivers were more likely to have difficulty when the status of patient became worse. In contrast, if a patient was a friend or grandparent, caregivers were more likely to have less difficulty when the condition of patient declined Figure 2, left).

Due to its clinical implication, we further investigated the predictability of caregiving difficulties by acaregivers medical condition: diabetes, heart attack (myocardial infarction), coronary artery disease (angina), stroke, asthma, use of assistant device, pregnancy, glaucoma, macular degeneration, cataract, cancer, insulin use, or snoring. Chances of caregivers having the greatest difficulties and less difficulty were similar (caregiver with disease: caregiver with no disease 50%:50%) regardless medical conditions which the caregivers had. In fact, perceived difficulties were slightly lower among those with medical conditions compared to without diseases. For example, 40% of caregivers with myocardial infarction history answered that they experienced difficulty, whereas 53% of caregivers with no myocardial infarction history answered that they experienced difficulty. In a similar way, rate of having difficulty was lower among the caregivers with diabetes (46%) than among the caregivers with no diabetes (53%) Figure 2, right).

4. Discussion.

In this study, we demonstrated data mining and neural network technologies, to investigate predictors of caregiver burden among caregivers. Data

Patient age (mean, SD)	69.87	20.53
Caregiver age (mean, SD)	56.14	15.46
Race/Ethnicity		
White	2,049	90.50%
Black	61	2.69%
Hispanic	69	3.05%
Others	56	2.47%
Patient Gender		
Male	795	35.11%
Female	1,455	64.27%
Employment		
Employed for wages	1,035	45.72%
Self-employed	220	9.72%
Unemployed	423	18.69%
Retired	582	25.71%
Income		
<$35,000	577	25.49%
<$50,000	299	13.21%
<$75,000	344	15.19%
≥$75,000	734	32.42%
Relationship		
(Grand) Parents	915	40.41%
Spouse	371	16.39%
Child, sibling, relatives	504	22.26%
Friends	451	19.92%
Patient status		
Cognitive changes	1,156	51.06%
No cognitive changes	1,038	45.85%
Not sure	29	1.28%

TABLE 1. Characteristics of Caregivers (n=2,264)

mining processes provided strategies to overcome challenges using common traditional methods such as multivariable regression analysis. Two thousand and two hundred sixty four caregiver records were extracted from the BRFSS (N = 451,075). In almost half of cases, caregivers expressed that they experience the greatest difficulties during caregiving. Patient mental status change, hours of caregiving, and relationship with patients appeared as the main predictors of caregivers stress (accuracy 63%, AUC = 65%) among 397 variables including demographics, behaviors, medical conditions, and environmental conditions. Medical condition of caregivers which is commonly assumed as a stress factor appeared to have low predictability.

Our study may contribute state of science with surprising findings due to their clinical meaningfulness;

1. to understand conditions of being vulnerable to stress among caregivers,and
2. to be aware that traditionally known factors (e.g. socioeconomic status, medical conditions) which have been commonly assumed as caregiverstress factors are not associated with caregiver stress among caregivers.

Caregiving duration		
$\leq$ 1 year	769	33.97%
$\leq$ 5 years	907	40.06%
$>$ 5 years	497	21.95%
Caregiving frequency		
$\leq$ 10 hours/week	1,344	59.36%
$\leq$ 30 hours/week	380	16.78%
$\leq$ 100 hours/week	201	8.88%
$>$ 100 hours/week	92	4.06%
Most needs		
Cleaning, managing $, prepare meals	614	27.12%
Transportation outside of the home	503	22.22%
Something else	317	14.00%
Self care - eating, dressing, bathing	302	13.34%
Relieving anxiety or depression	184	8.13%
Caregiving difficulties		
No difficulty	1,013	54.0%
Difficulty	1,178	44.7%
Not sure/ Dont know	28	1.25%
Refused	24	1.07%
Greatest difficulties having difficulties		
Creates emotional burden	528	44.82%
Not enough time for yourself	165	14.01%
Other difficulty	113	9.59%
Creates financial burden	95	8.06%
Affects family relationships	85	7.22%
No enough time for your family	84	7.13%
Interferes with your work	71	6.03%
Aggravates health problems	37	3.14%

TABLE 2. Characteristics of Caregivers - Cont'd (n=2,264)

First, understanding predictors of caregiver difficulties will be helpful to prepare providers how to guide caregivers and how to choose the right candidate for intervention. Although more than half of caregivers faced cognitive changes of patients, in reality providers rarely understand how much of the caregivers might experience difficulties with patients condition changes. Based on our finding, providers can augment caregiver education, which may prepare caregivers better when cognitive changes occur.

One of the findings which surprised the authors was that those three factors (mental cognitive change, care hours, relationship) of predicting caregiver burden were overriding any other socioeconomic status (SES) factors. The findings of patient cognitive change and hours of care providing as primary predictors of caregiving difficulties are specifically relevant to the family caregivers of dementia, most common illness of caregiving [1]. Patient with dementia including Alzheimers disease progressively show cognitive changes due to irreversible brain changes. Caregivers of dementia patients often spend nine or more hours per day providing care for five

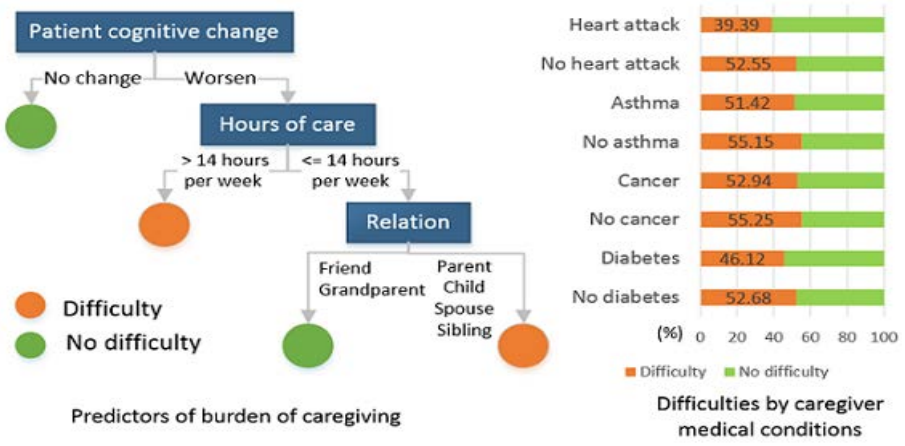

FIGURE 2. Burden of caregivers

years or more [5]. National surveys report that dementia caregiving is more psychologically stressful and physically exhausting compared to caregiving for other health conditions.20With the rapidly increased aging population in the 21 century, the burden of caregivingis expected to be widespread [1, 5]. Health professionals are encouraged to inform the caregivers of patients who are expected to have neurocognitive changes, regarding importance of limiting care hours, and educate about burn out and delegation of caregiving tasks.

A considerable amount of research has been devoted to economic burden of caregivers [9, 10, 20]. On the contrary, our finding showed that more than half of caregivers (55%) make more than $50,000 per year. In fact approximately 40% of caregivers make more than $75,000, and only 3% of caregivers met the national poverty thresholds ($14,710 for 2 persons in family in 48 states and D.C) [15]. One of the reasons that our study had conflicting results may be from application of new methods. As mentioned earlier, traditional methods to investigate predictors such as multivariate regression are limited in handling hundreds of variables. Our approach was to apply data mining techniques, relatively new methods which allowed us to investigate hundreds of variables at once, and yielded surprising results. Another clinically meaningful result is that caregiving difficulties are not associated with caregivers medical conditions. We looked at common chronic diseases and medical conditions including diabetes, heart attack, coronary artery disease, stroke, asthma, use of assistant device, pregnancy, glaucoma, muscular degeneration, cataract, cancer, insulin use, or snoring. Our study revealed that chances of caregivers having difficulties were similar to caregivers having no difficulty (caregiver with disease: caregiver with no disease 50%:50%) regardless of any caregvier's medical conditions. In fact, perceived difficulties were slightly lower among those with medical conditions compared to without diseases (Figure 2, right).

This study has several potential weaknesses using a self-report telephone survey data. First, these results only reflect population who could answer the survey. It is unknown which population could not answer the long length (85 pages of questionnaires) of survey. Second, the former research studying caregiver has tended to focus on disease-specific caregiver stress (e.g. dementia caregiver stress, stroke

caregiver stress). We only looked at the macro-level phenomenon of caregiver stress, and did not investigate disease specific information.

5. **Conclusion.** Data mining methods were useful to discover new behavioral risk knowledge and to visualize predictors of caregiver stress from a multidimensional behavioral dataset. This study suggests that health professionals target dementia family caregivers who are anticipated to experience patients' neuro-cognitive changes, and inform the caregivers about importance of limiting care hours, burn out and delegation of caregiving tasks.

REFERENCES

[1] R. D. Adelman, L. L. Tmanova, D. Delgado, S. Dion and M. S. Lachs, Caregiver burden: A clinical review, *Jama*, **311** (2014), 1052–1060.

[2] A. Barfar and B. Padmanabhan, Predicting presidential election outcomes from what people watch, *Big Data*, **5** (2017), 32–41.

[3] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford university press, 1995.

[4] L. Breiman, Random forests, *Machine Learning*, **45** (2001), 5–32.

[5] C.-Y. Chiao, H.-S. Wu and C.-Y. Hsiao, Caregiver burden for informal caregivers of patients with dementia: A systematic review, *International Nursing Review*, **62** (2015), 340–350.

[6] G. DePalma, H. Xu, K. E. Covinsky, B. A. Craig, E. Stallard, J. Thomas III and L. P. Sands, Hospital readmission among older adults who return home with unmet need for ADL disability, *The Gerontologist*, **53** (2013), 454–461.

[7] C. for Disease Control and Prevention, Behavioral risk factor surveillance system survey data, atlanta, georgia. u.s.

[8] J. E. Gaugler, D. L. Roth, W. E. Haley and M. S. Mittelman, Can counseling and support reduce burden and depressive symptoms in caregivers of people with Alzheimer's disease during the transition to institutionalization? results from the new york university caregiver intervention study, *Journal of the American Geriatrics Society*, **56** (2008), 421–428.

[9] P. E. Greenberg and H. G. Birnbaum, The economic burden of depression in the us: Societal and patient perspectives, *Expert Opinion on Pharmacotherapy*, **6** (2005), 369–376.

[10] S. Gupta, G. Hawker, A. Laporte, R. Croxford and P. Coyte, The economic burden of disabling hip and knee osteoarthritis (oa) from the perspective of individuals living with this condition, *Rheumatology*, **44** (2005), 1531–1537.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, The weka data mining software: An update, *ACM SIGKDD Explorations Newsletter*, **11** (2009), 10–18.

[12] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444.

[13] S. J. Lupien, B. S. McEwen, M. R. Gunnar and C. Heim, Effects of stress throughout the lifespan on the brain, behaviour and cognition, *Nature Reviews Neuroscience*, **10** (2009), 434–445.

[14] P. C. J. Navas, Y. C. G. Parra and J. I. R. Molano, Big data tools: Haddop, mongodb and weka, in *International Conference on Data Mining and Big Data*, Springer, 2016, 449–456.

[15] U. D. of Health and Human Service., 2011 poverty guideline, *Federal Register*, **76** (2010), 3637–3638.

[16] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge university press, 2007.

[17] J. W. Rowe, T. Fulmer and L. Fried, Preparing for better health and health care for an aging population, *Jama*, **316** (2016), 1643–1644.

[18] J. Schmidhuber, Deep learning in neural networks: An overview, *Neural Networks*, **61** (2015), 85–117.

[19] B. C. Spillman and S. K. Long, Does high caregiver stress predict nursing home entry?, *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, **46** (2009), 140–161.

[20] C. H. Van Houtven, S. D. Ramsey, M. C. Hornbrook, A. A. Atienza and M. van Ryn, Economic burden for informal caregivers of lung and colorectal cancer patients, *The Oncologist*, **15** (2010), 883–893.

[21] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.

[22] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu and P. S. Yu et al., Top 10 algorithms in data mining, *Knowledge and Information Systems*, **14** (2008), 1–37.

[23] E. Yan and T. Kwok, Abuse of older Chinese with dementia by family caregivers: An inquiry into the role of caregiver burden, *International Journal of Geriatric Psychiatry*, **26** (2011), 527–535.

[24] Q. Yang and X. Wu, 10 challenging problems in data mining research, *International Journal of Information Technology & Decision Making*, **5** (2006), 597–604.

E-mail address: Sunmoo Yoon: sy2102@columbia.edu
E-mail address: Maria Patrao: map9063@nyp.org
E-mail address: Debbie Schauer: mschauerd@aol.com
E-mail address: Jose Gutierrez: jg3233@cumc.columbia.edu