Predictive Modeling of Inpatient Fall of Stroke Patients Using electronic medical records data

Ying Liu * Cindy Jin † Jeffrey Yangang Zhang ‡

Abstract

Predictive modeling of inpatient fall of stroke patients is challenging. Traditionally, logistic regression or survival models could be applied, but . This study presents a framework of analyzing discrete-time data as a binary classification problem. With its great flexibility of the novel framework, the potential relationship between clusters can be incorporated. Comparisons in experiments suggest that the proposed models consistently yield better predictive performances than classical statistical modeling for survival data. The proposed method is applied to electronic patient records data from 2007 to 2014 collected by Kessler foundation.

Key Words: Discrete-time survival model, multi-task learning, ensemble algorithm, electronic medical record, patient fall

1. Introduction

To stroke survivors, fall is one of the most commonly occurring adverse events. Studies found that falls occurred more frequently in stroke patients than in other populations [Yates et al., 2002, Weerdesteyn et al., 2008]. In acute care setting, fall incidence of stroke patients ranges from 14% to 64.5%. In rehabilitation centers, fall incidence does not improved much. 24% - 47% stroke patients reported fall experience in the rehabilitation setting [Weerdesteyn et al., 2008]. Falls usually result in serious, sometimes fatal, consequences, and thus it is one of the high-priority safety risks. Therefore, preventing falls is an important strategy for improving care among stroke patients.

The Joint Commission International accreditation standards for hospitals states that the first step to fall prevention is to assess patients for risk of falling within 24 hours of admission, and identify and educate patients at risk of falls [JCI, 2013]. Different tools have been developed, including well-known instruments such as Morse Fall Scale (MFS) [Morse, 2006], St. Thomas's Risk Assessment Tool in Falling Elderly Inpatients (STRATIFY) [Oliver et al., 1997], the Hendrich Fall Risk Model [Hendrich et al., 1995, 2003], and new approaches like Lee et al. [2016] proposed an automated fall risk assessment system Auto-FallRAS based on electronic medical records (EMR) data. But these tools do not always show high accuracy across institutions [Milisen et al., 2007, Chow et al., 2007]. Thus, Oliver et al.

*iMEDacs, 1531 Natalie Ln, Ann Arbor, MI 48105

†Princeton Pharmartech, 43 Princeton Hightstown Rd, Princeton Jct, NJ 08550

‡Princeton Pharmartech, 43 Princeton Hightstown Rd, Princeton Jct, NJ 08550

[1997] recommended to use an institutional and patient-specific validated risk assessment tool.

The keystone followed by almost all fall-risk assessment tools is the simplicity to use in practice. For example, MFS and STRATIFY consist of only 5-10 items, which enable nurses to score patients' conditions on each item easily. While useful as demonstrated by studies, these paper-based tools still increase workloads of care providers. Ideally, it is better to develop an automated risk assessment system to quickly assess fall risk, improve patient safety and reduce the heavy workload [JCI, 2013].

The main goals of this study are to (1) investigate a new learning algorithm which can incorporate cluster effects, and (2) evaluating the performance of an ensemble algorithm. As observed by many studies, there exists institution-specific performance when applying risk assessment tools, which may be explained by institutional-level factors. This observation motivates us to consider a learning algorithm which can incorporate such prior cluster information. With the proposed novel method as a sub-learner, we also aim to examine an ensemble algorithm to see what we can gain with a meta-learner in predicting inpatient fall.

The rest of the paper is organized as follows. Section 2 describes details of the new algorithm starting with its connection to discrete-time hazard model. In section 3, the method will be applied to predict inpatient fall event based on EMR data which combines data from Inpatient Rehabilitation Facility Prospective Payment System (IRF PPS) with information from incident report system in each rehabilitation site. While comparing with other learning algorithm, one ensemble algorithm will also be examined in section 3. In the final section, we will provide a brief discussion on learning algorithms in patient fall prediction.

2. Discrete-time hazard model

2.1 Notation

Let's denote by $T_i \in \{1, 2, ...\}$ a discrete survival outcome with covariate vector $\mathbf{x}_i$. Correspondingly, $C_i \in (1, 2, ...)$ is a random censoring time. Here it is assumed that C_i is independent of τ_i. Thus, the time period during which Subject i is under observation is denoted by $\tau_i := \min(T_i \leq C_i)$. The random variable $\delta_i := I(T_i \leq C_i)$ indicates whether τ_i is right censored.

The hazard function of a discrete time survival model is defined as the conditional probability of an event at time point j experienced by a subject, $h_{ij} = P(T_i = j | T_i \geq j, X = x)$. Then, the survival function is given by $S(t|x) = P(T_i > j | X = x)$, where T_i is the true time to event, which is unobserved for censored subjects. The hazard h_{ij} is conditional probability. Individual only contribute information at time J_i if they experienced/didn't experience the event at the period J_i. Thus, the likelihood of his entire event history, represented by the vector $\mathbf{y}_i = (y_{i1}, ..., y_{iJ_i})$ with $y_{ij} = \mathbf{I}(T_i < j)$, is given by

$$L = \prod_{i=1}^{n} \left[\left(h_{iJ_i} \prod_{j=1}^{J_i - 1} (1 - h_{ij}) \right)^{\delta_i} \left(\prod_{j=1}^{J_i} (1 - h_{ij}) \right)^{1-\delta_i} \right] \quad (1)$$

with the assumption of ignorable missingness, *i.e.* no information contribution from censored subjects. As noted by Allison [1982] and Singer and Willett [1993], if represent the event history in terms of response $\mathbf{y}_i$, the likelihood can also be written as

$$L = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \left\{ h_{ij}^{y_{ij}} (1 - h_{ij})^{1-y_{ij}} \right\}. \quad (2)$$

Subject's hazard is affected by various factors. Thus, we have the hazard function of the form

$$g(h_{ij}) = \alpha_j + f(\boldsymbol{\beta}^T \mathbf{x}_i), \quad (3)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$, g and f are pre-specified link functions. $f(z)$ is usually set to be an identity link, with which we assume linear relationship in the covariates. Options for g can be logistic function or complementary log-log function.

Equation (2) is identical to the likelihood function for a sequence of $J_1 + J_2 + \cdots + J_n$ independent Bernoulli trials with parameter h_{ij} [Allison, 1982, Singer and Willett, 1993]. As such, the event response for an individual at each discrete time period can be treated as a separate, independent observation. This allows application of likelihood-based algorithm for standard binary classification. However, it assumes a common set of parameter $\boldsymbol{\beta}$ in Eq (3) for all time periods, which may not reflect the reality in which factors can potentially have heterogeneous effect on patients over time due to changes of disease status. We thus consider to relax the restriction.

2.2 Relation to multi-task learning framework

The completely unconstrained discrete time model is to model each time point separately. Instead of letting time-specific $\boldsymbol{\beta}$ completely random, a more sensible, and intuitive as well, approach is to smoothen the transition between consecutive time points in the unconstrained model. Thus constructing a model with all time points together is more attractive, since we can attach certain controls for achieving various purposes, such as parameter selection and transition smoothening. This leads to a case of multi-task model, in which each time point is treated as an individual task [Zhou et al., 2013]. The multi-task model provides us more flexibility as well, including incorporating cluster structure of observed data.

Still, we consider a logistic model for the probability of event at a time point: $P(y_{ij} = 0) = \{1 + \exp(\gamma_j + \boldsymbol{\theta}_j^T \mathbf{x}_i)\}^{-1}$. For case i with an event at time j, instead of coding as $(y_{i1}, ..., y_{ij}) = (0, 0, ..., 1)$, we expand his history to the maximal observed time J in the population, *i.e.* $(y_{i1}, ..., y_{ij}, ..., y_{iJ}) = (0, .., 1, ..., 1)$, that is, all $(y_{i,(j+1)}, .., y_{iJ})$ are a series of 1, which reflects the scenario that an occurred event cannot change back, i.e. from $y_{ij} = 1$ to

$y_{i,j+1} = 0$ (in modeling non-recurrent event). Then,

$$P(y_{i1}, ..., y_{iJ}|\mathbf{x}_i) = \frac{\exp\left\{\sum_{j=1}^{J} y_{ij}(\gamma_j + \boldsymbol{\theta}_j^T \mathbf{x}_i)\right\}}{\sum_{k=0}^{J} \exp\left\{\sum_{l=k+1}^{J}(\gamma_l + \boldsymbol{\theta}_l^T \mathbf{x}_i)\right\}}.$$

For censored cases, we depend on a matrix W with $w_{ij} \in \{0, 1\}$ being patients' observational status experienced by patients. $w_{ij} = 1$ indicates a patient is censored, while $w_{ij} = 0$ indicates that a patient is still observed in the study. Then the likelihood for all uncensored and censored patients is:

$$\begin{aligned} L(\boldsymbol{\Theta}) = & \sum_{i=1}^{n_e} \sum_{j=1}^{J} y_{ij}(\gamma_j + \boldsymbol{\theta}_j^T \mathbf{x}_i) \\ & + \sum_{i=n_e+1}^{n} \log \left[\sum_{j=1}^{J} w_{ij} \exp\left(\sum_{l=j}^{J} (\gamma_l + \boldsymbol{\theta}_l^T \mathbf{x}_i) \right) \right] \\ & - \sum_{i=1}^{n} \log \left(\sum_{k=0}^{J} \exp\left\{ \sum_{l=k+1}^{J} (\gamma_l + \boldsymbol{\theta}_l^T \mathbf{x}_i) \right\} \right) \end{aligned} \quad (4)$$

where $\boldsymbol{\Theta} = (\Theta_1, ...\Theta_J) = (\gamma_1, \boldsymbol{\theta}_1, ..., \gamma_J, \boldsymbol{\theta}_J)$.

The parameter $\boldsymbol{\Theta}$ can be estimated through minimizing the following objective function:

$$\min_{\boldsymbol{\Theta}} \quad L(\boldsymbol{\Theta}) + \xi_1 \sum_{j=1}^{J} ||\Theta_j||_2^2 \quad (5)$$

The penalty term is for controlling generalization error, with $\xi_1 > 0$ as a regularization parameter. The model (5) is also known as the ridge regression.

As discussed before, assuming parameter sets for time points are completely independent between each other may not reflect the longitudinal scenario in which a close connection between time j and $j + 1$ may exists. To capture the smooth variation of parameters across time points, we introduce the following regularization term:

$$\Omega_j = \xi_2 \sum_{j=1}^{J-1} ||\Theta_{j+1} - \Theta_j||_2^2 \quad (6)$$

where $\xi_2 > 0$ is the regularization parameter for controlling the temporal smoothness. In matrix format, the term can be expressed as $||\boldsymbol{\Theta}F||_2^2$ where $F \in \mathbb{R}^{J\times(J-1)}$ is defined as $F_{ij} = 1$ if $i = j$, $F_{ij} = -1$ if $i = j + 1$, and $F_{ij} = 0$ if otherwise.

2.3 Modeling relationship between sites

Since patient records were built by individual rehabilitation sites, it is reasonable to assume that data have site-specific characterizations [Yokota and

Ohe, 2016]. To capture such information, site-specific tasks are constructed in the multi-task model. Considering potential relationships between sites, especially when the sites are under the same management system or from the same geographical area, we build the model with additional parameters for modeling the relationship between sites.

Denote clusters by $h = 1, ..., H$, and let $\mathbf{R}$ be a matrix with $R_{h,k}$ describing a relation between site h and k. Then, we define a penalty based on the graph Laplacian regularizer for incorporating the prior cluster knowledge:

$$\Omega_r = \sum_{h,k=1}^{H} R_{h,k}||\Theta_h - \Theta_k||_2^2 \tag{7}$$

Put terms (6) and (7) into the model, the objective function becomes

$$\min_{\mathbf{\Theta}} \sum_{h=1}^{H} L(\mathbf{\Theta}_h) + \xi_1 \sum_{h=1}^{H}\sum_{j=1}^{J} ||\Theta_{h,j}||_2^2 + \xi_2 \sum_{h=1}^{H}\sum_{j=1}^{J-1} ||\Theta_{h,j+1} - \Theta_{h,j}||_2^2 + \xi_3 \sum_{h,k=1}^{H} A_{h,k}||\Theta_h - \Theta_k||_2^2 \tag{8}$$

2.4 Optimization algorithm

This optimization problem can be efficiently solved using an alternatinge approach which is similar to the block coordinate descent method [Nocedal and Wright, 2006]. Parameter $\mathbf{\Theta}$ and relation matrix $\mathbf{R}$ are optimized alternatively with the other one fixed. Given $\mathbf{R}$, the proximal gradient descent method can be applied for optimizing $\mathbf{\Theta}$ [Parikh and Boyd, 2014]. The proximal gradient descent is a commonly applicable algorithm when solving optimization problems containing non-differentiable components. It also has faster convergence compared to other methods such as subgradient descent. Minimizing a corresponding upper bound function of the original objective function is the basic idea of this algorithm. Given $\mathbf{\Theta}$, the relation matrix $\mathbf{R}$ can be optimized using similar techniques in Zhou et al. [2011].

3. Simulation

We illustrate the proposed method with simulation studies. Generated data includes 5 covariates: $x_1 \sim uniform(2, 4)$, $x_2 \sim N(0, 1)$, $x_3 \sim uniform(0, 3)$, $x_4 \sim uniform(0, 3)$, and $x_5 \sim N(0, 1)$. Parameters are set at $(\alpha, \beta_1, ..., \beta_5) = (-1.5, -2, -3, 0.9, -1.2, -0.8)$. The hazard function is defined as

$$h_{ki} = h_0 \exp \alpha_i + \beta_{1i}x_1 + \beta_{2i}x_2 + \beta_3\sqrt{x_{3i}} + \beta_4 \log x_{4i} + \beta_5 x_{5i}^2,$$

with h_0 be the baseline hazard rate leading to the simulated event rate which is close to that in the observed data of inpatient fall. Three clusters (sites) are formed with intra-cluster correlation $(\rho_1, \rho_2, \rho_3) = (0.08, 0.1, 0.15)$ to construct a covariance matrix Σ. Event time of subject i in cluster k is then calculated using $T_{ki} = -h_{ki}^{-1} \log(1 - U_{ki})$ where $U_{ki} \sim N(0, \Sigma)$. The generated sample size is $n = 1000$ in each run. All evaluation metrics are averaged over 1000 runs.

The novel method will be compared with discrete-time logistic model (GLM), discrete-time survival tree (DSTree), and Bayesian discrete-time logistic model (BGLM). Evaluation metrics include (a) Root mean squared error (RMSE) or Brier Score: $\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - p_{ij})^2$; (b) Misclassification error rate (MER) using cut-off 0.5: $\frac{1}{n}\sum_{i=1}^{n}|y_{ij} - I(p_{ij} > 0.5)|$; (c) Area under the ROC curve (AUC); (d) False positive rate(FPR) using cut-off 0.5: $\frac{1}{n_{y=0}}\sum_{i=1}^{n}(1 - y_{ij})I(p_{ij} > 0.5)$; (e) False negative rate(FNR) using cut-off 0.5: $\frac{1}{n_{y=1}}\sum_{i=1}^{n}(1 - y_{ij})I(p_{ij} \leq 0.5)$;

Simulation results indicates the new method improves performance in terms of evaluated metrics when comparing with several ordinary methods (Table 1), though some has only slightly improved. This indicates that incorporating cluster information can benefit model performance.

Table 1: Evaluation metrics estimated using simulated data.

Rate	Learner	RMSE	MER	AUC	FPR	FNR
0.05	GLM	0.0056	0.0063	85.7	0.0065	18.2
	DSTree	0.0150	0.0427	73.4	0.0105	22.1
	BGLM	0.0050	0.1031	87.2	0.0059	19.4
	MTL	0.0043	0.0058	88.1	0.0055	15.3
0.10	GLM	0.0136	0.1062	93.0	0.0107	12.1
	DSTree	0.0218	0.0713	83.5	0.0206	18.2
	BGLM	0.0153	0.0931	93.6	0.0112	12.4
	MTL	0.0093	0.0881	94.2	0.0135	12.1

4. Application in inpatient fall prediction

Electronic patient records from 2007 to 2014 were provided by Kessler foundation, which consists of three campuses. Required by CMS as part of the Inpatient Rehabilitation Facility Prospective Payment System (IRF PPS), the electronic records collected information on patient identification, disease diagnoses, functionality at admission, and status at discharge as well using Inpatient Rehabilitation Facility - Patient Assessment Instrument (IRF-PAI). With IRF-PAI, a set of function modifiers is collected by assessing patients' functionality in bladder level/frequency, bowel level/frequency, tub transfer, shower transfer, distance walked/traveled in wheelchair, etc. The function modifiers serve to assist in the scoring of related FIM instrument items, and to provide explicit information as to how a particular FIM item score has been determined. Patients are assessed at admission and discharge time. Only information at admission is used for modeling in this study.

Kessler campuses have incident report system. The system recorded every incident event patients experienced during their stay, including event date, type (fall, burn, etc.), and related diagnoses. Since all fall events in incident reports were judged by medical staff in rehabilitation facilities, we define a fall as "an event judged to be a fall for which a fall report was created by medical staff at a rehabilitation center" in this study. No further review is performed on each individual fall event by comparing fall reports with medical records. The incident report data is merged with patients' admission data which were cleaned first. Incident dates are the basis to identify legitimate fall events, i.e. a fall event is considered as legitimate only when the event date can be landed within an appropriate interval determined by admission and discharge date. For patients with multiple fall events, the first event is considered and identified based on event date.

Polychoric correlation is applied on ordinal function modifiers and FIM item scores [Olsson et al., 1982]. On obtained correlation matrix, principal component analysis (PCA) with PROMAX rotation is applied to find a simple structure of factors. Factor loadings are computed based on the structure of factors and are used in predictive model building.

Based on ICD-9-CM comorbidity diagnosis codes, Elixhauser comorbidity measures are created: separate indicator variables for 30 different diseases, including cancer, cardiovascular/blood disease, endocrine disorders, gastrointestinal disease, musculoskeletal/pain-related disease, neurologic disorders, nutritional disease/obesity, psychological/behavioral disease, and miscellaneous disease [Elixhauser et al., 1998, Austin et al., 2015].

The data will be split into training set and test set. Predictive models, including penalized logistic model(GLMp), random forest, multi-layer perceptron (MLP), multivariate adaptive regression spline (MARS), support vector machines with radial basis function kernel (SVM), and generalized additive model using splines (GAM), are developed on the training set. Repeated 10-fold cross validation will be applied to tune parameters for each model. Table 2 lists evaluation metrics of prediction with the validation data set. The proposed method is at top of all learner in terms of AUC.

We consider to ensemble these learners with a genetic algorithm (GA). The algorithm assigns weights to each classifier. Weight are proportional to their performances on evaluation metrics. The basic procedure of a genetic algorithm is first to generate a weight distribution vector randomly, and then the performance of its corresponding meta-classifier will be evaluated using the holdout subset. Based on the fitness of each member learner, GA selects members for the next generation. And by crossover and mutation operators, a new generation of weight distribution vectors is created. With the set of new weight, the meta-classifier will be re-evaluated. Repeat the above steps B times. With the ensemble algorithm, we achieved an AUC of 0.874, with a sensitivity of 0.983 and a specificity of 0.299, on the hold-out test data set.

5. Conclusion

This study investigated an extended multi-task learning framework with prior cluster information incorporated as tasks. The flexibility of the algorithm enables us to integrate potential correlation among clusters. Ap-

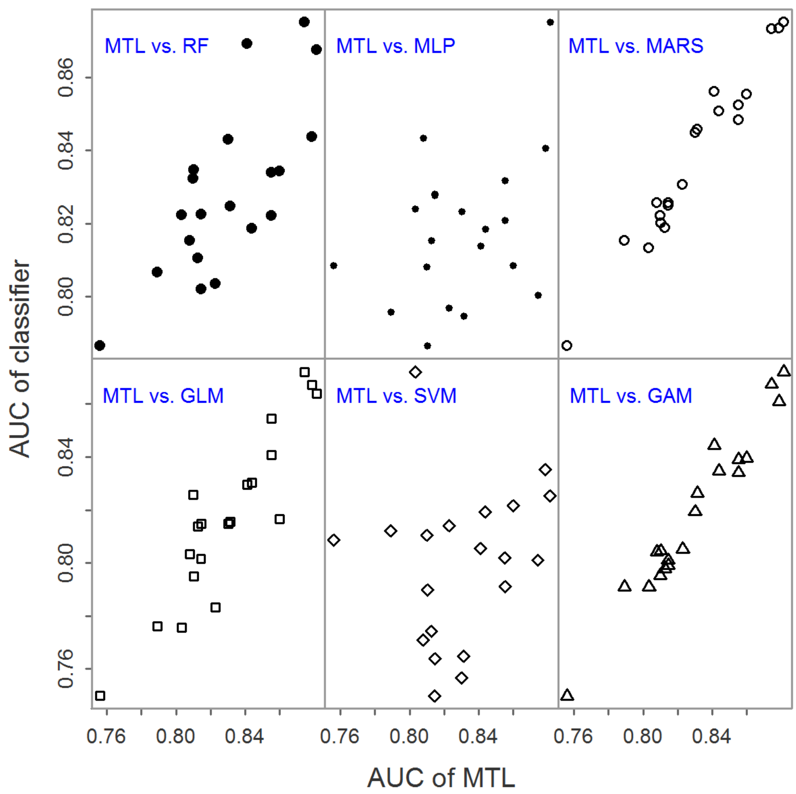

Figure 1: Scatter plots of AUCs from repeated 10-fold cross-validation between MTL and other algorithms.

plication to a real world data set showed that the proposed framework outperformed other algorithms.

As noticed in the experiment on EMR data, the ensemble learner performed well in terms of sensitivity, but have a large space to improve in specificity. This is probability due to the limitation of the data. This set of EMR data did not collect any laboratory measurements. Furthermore, for the functional modifier and FIM score, only measurements at admission can be used. There was no intermittent measuring during patients stay. Though measured at discharge, they are not helpful in terms of predicting fall. Thus, there was no information on change in patients' functionalities. Such a change is found to be an important predictor of fall in our exercise of building models with measurements at discharge included.

Table 2: Performance of sub-learners from impatient fall data.

	AUC			Sensitivity			Specificity		
	Mean	95% CI		Mean	95% CI		Mean	95% CI	
MTL	0.830	0.816	0.844	0.947	0.942	0.952	0.366	0.339	0.393
RF	0.829	0.818	0.839	0.982	0.979	0.985	0.207	0.185	0.230
MLP	0.674	0.655	0.693	1.000	1.000	1.000	0	0	0
MARS	0.828	0.815	0.841	0.979	0.976	0.982	0.213	0.188	0.238
SVM	0.766	0.755	0.776	0.985	0.982	0.988	0.162	0.144	0.179
GAM	0.828	0.815	0.841	0.981	0.978	0.983	0.198	0.169	0.227
GLMp	0.820	0.804	0.835	0.998	0.997	0.999	0.007	0.003	0.012

References

P. D. Allison. Discrete-time methods for the analysis of event histories. In S. Leinhardt, editor, *Sociol. Methodol.* Jossey-Bass Inc.,U.S., 1982.

S. R. Austin, Y. N. Wong, R. G. Uzzo, J. R. Beck, and B. L. Egleston. Why Summary Comorbidity Measures Such As the Charlson Comorbidity Index and Elixhauser Score Work. *Med Care*, 53(9):65–72, Sep 2015.

S. K. Chow, C. K. Lai, T. K. Wong, L. K. Suen, S. K. Kong, C. K. Chan, and I. Y. Wong. Evaluation of the Morse Fall Scale: applicability in Chinese hospital populations. *Int J Nurs Stud*, 44(4):556–565, May 2007.

A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey. Comorbidity measures for use with administrative data. *Med Care*, 36(1):8–27, Jan 1998.

A. Hendrich, A. Nyhuis, T. Kippenbrock, and M. E. Soja. Hospital falls: development of a predictive model for clinical practice. *Appl Nurs Res*, 8(3):129–139, Aug 1995.

A. L. Hendrich, P. S. Bender, and A. Nyhuis. Validation of the Hendrich II Fall Risk Model: a large concurrent case/control study of hospitalized patients. *Appl Nurs Res*, 16(1):9–21, Feb 2003.

JCI. Hospital accreditation standards. *Joint Commission International: 5th field review*, 2013.

J. Y. Lee, Y. Jin, J. Piao, and S. M. Lee. Development and evaluation of an automated fall risk assessment system. *Int J Qual Health Care*, 28(2):175–182, Apr 2016.

K. Milisen, N. Staelens, R. Schwendimann, Ln De Paepe, J. Verhaeghe, T. Braes, S. Boonen, W Pelemans, RW. Kressig, and E Dejaeger. Fall prediction in inpatients by bedside nurses using the st. thomas's risk assessment tool in falling elderly inpatients (stratify) instrument: A multicenter study. *Journal of the American Geriatrics Society*, 55(5):725–733, 2007.

J. M. Morse. The safety of safety research: the case of patient fall research. *Can J Nurs Res*, 38(2):73–88, Jun 2006.

J. Nocedal and S. Wright. *Numerical Optimization (Springer Series in Operations Research and Financial Engineering).* Springer, 2006.

D. Oliver, M. Britton, P. Seed, F. C. Martin, and A. H. Hopper. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. *BMJ*, 315(7115):1049–1053, Oct 1997.

U. Olsson, F. Drasgow, and N. Dorans. The polyserial correlation coefficient. *Psychometrika*, 47(3):337–347, 1982.

N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014.

J. D. Singer and J. B. Willett. It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *J. Educ. Behav. Stat.*, 18(2):155–195, January 1993.

V. Weerdesteyn, M. de Niet, H. J. van Duijnhoven, and A. C. Geurts. Falls in individuals with stroke. *J Rehabil Res Dev*, 45(8):1195–1213, 2008.

J. S. Yates, S. M. Lai, P. W. Duncan, and S. Studenski. Falls in community-dwelling stroke survivors: an accumulated impairments model. *J Rehabil Res Dev*, 39(3): 385–394, 2002.

S. Yokota and K. Ohe. Construction and evaluation of find, a fall risk prediction model of inpatients from nursing data. *Japan Journal of Nursing Science*, 13(2):247–255, 2016.

J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 702–710. Curran Associates, Inc., 2011.

J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via multi-task learning. *Neuroimage*, 78:233–248, Sep 2013.