Front. Eng. Manag. 2022, 9(4): 577–591
https://doi.org/10.1007/s42524-022-0203-7

RESEARCH ARTICLE

Nazmus SAKIB, Xuxue SUN, Nan KONG, Chris MASTERSON, Hongdao MENG, Kelly SMITH, Mingyang LI

Heterogeneous length-of-stay modeling of post-acute care residents in the nursing home with competing discharge dispositions

Abstract Post-acute care (PAC) residents in nursing homes (NHs) are recently hospitalized patients with medically complex diagnoses, ranging from severe orthopedic injuries to cardiovascular diseases. A major role of NHs is to maximize restoration of PAC residents during their NH stays with desirable discharge outcomes, such as higher community discharge likelihood and lower re/hospitalization risk. Accurate prediction of the PAC residents' length-of-stay (LOS) with multiple discharge dispositions (e.g., community discharge and re/hospitalization) will allow NH management groups to stratify NH residents based on their individualized risk in realizing personalized and resident-centered NH care delivery. Due to the highly heterogeneous health conditions of PAC residents and their multiple types of correlated discharge dispositions, developing an accurate prediction model becomes challenging. Existing predictive analytics methods, such as distribution-/regression-based methods and machine learning methods, either fail to incorporate varied individual characteristics comprehensively or ignore multiple discharge dispositions. In this work, a data-driven predictive analytics approach is considered to jointly predict the individualized re/hospitalization risk and community discharge likelihood over time in the presence of varied residents' characteristics. A sampling algorithm is further developed to generate accurate predictive samples for a heterogeneous population of PAC residents in an NH and facilitate facility-level performance evaluation. A real case study using large-scale NH data is provided to demonstrate the superior prediction performance of the proposed work at individual and facility levels through comprehensive comparison with a large number of existing prediction methods as benchmarks. The developed analytics tools will allow NH management groups to identify the most at-risk residents by providing them with more proactive and focused care to improve resident outcomes.

Keywords nursing home, predictive analytics, individualized prediction, competing risks, health outcomes

Received November 30, 2021; accepted March 1, 2022

Nazmus SAKIB, Mingyang LI (✉)
Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA
E-mail: mingyangli@usf.edu

Xuxue SUN
College of Media Engineering, Communication University of Zhejiang, Hangzhou 310019, China

Nan KONG
Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN 47907, USA

Chris MASTERSON
Greystone Health Network, Tampa, FL 33610, USA

Hongdao MENG, Kelly SMITH
School of Aging Studies, University of South Florida, Tampa, FL 33620, USA

This work was supported in part by National Science Foundation under Grant Nos. 1825761 and 1825725.

1 Introduction

Nursing homes (NHs), or skilled nursing facilities, are mainly responsible for caring for the frail and vulnerable population of older adults with 24/7 personal care and assistance. Historically, NHs have been considered as a major healthcare setting for providing custodial care for long-term care (LTC) residents. In recent decades, while maintaining their conventional role as LTC providers, modern NHs have increasingly been responsible for caring for post-acute care (PAC) residents who are recently hospitalized patients and require extended rehabilitation and recovery after an acute care hospital stay. Recent studies reported that the percentage of residents (within NHs) admitted from the hospital increased from 67% in 2000 to 85% in 2015 (Fashaw et al., 2020). Several policy and market changes contributed to the

shifts of the NH population composition. First, the rising trend toward rapid hospital discharge with the reduced hospital length-of-stay (LOS) generated a growing population of quicker-and-sicker patients and drove many NHs to expand their sub-acute care and rehabilitation services (Murad, 2011). Second, Medicare covers qualified PAC residents with a higher reimbursement rate than Medicaid in NHs, the latter of which is the primary payer for qualified LTC residents who mainly require custodial care. In 2017, the US average reimbursement rate of Medicaid was $206 per resident day, which is less than half the rate paid by Medicare, $503 per resident day (NIC, 2018). Thus, a strong financial incentive for NH providers to accept more PAC Medicare beneficiaries was observed. Finally, due to the impact of a series of laws and initiatives (Moon et al., 1997; Ginsburg and Supreme Court of the US, 1998; Eiken et al., 2014) in recent decades, there has been a considerable increase in home and community-based services for the overall LTC population to advocate age-in-place and divert or delay the expensive NH placement.

With the growing demand of PAC residents in NHs, the major goal of NH in caring PAC residents is to return them to the community successfully and efficiently with lower re/hospitalization rate. Thus, successful prediction on how long every individual PAC resident will stay in the NH and what will be his/her discharge disposition is greatly important for both NH administrators and individual residents (and their family). To NH administrators, successfully predicting LOS and the discharge disposition of each resident at individual level with individual risk factors identified will help administrators identify the most at-risk residents (e.g., residents with shorter LOS and higher re/hospitalization risk, and residents with longer LOS and lower community discharge likelihood), and more targeted care can be provided to improve the care quality of the overall facility. At the facility level, an accurate predictive model of LOS with incorporation of varied individual characteristics will further allow accurate evaluation of the NH utilization (measured in average LOS) of a heterogeneous population of PAC residents with varied individual characteristics. To an individual resident and his/her family, accurate prediction of how long he or she will stay in an NH will improve the communication between caregivers and care recipients. It will further assist the family of residents to prepare the informal care resources to accommodate the needs of residents to be discharged.

Accurate prediction of LOS and discharge disposition of PAC residents is challenging. First, PAC residents admitted in NHs are often medically complex with a high level of functional dependence and with a variety of clinical diagnoses, ranging from severe orthopedic injuries (e.g., hip and pelvic fractures) to cardiovascular diseases (e.g., stroke and myocardial infarction). It is unclear which individual characteristics will affect LOS and discharge disposition. Many of the existing LOS models in literature are distribution-based methods and considered various distributions, such as Exponential, Phase-type, Log-normal and Gamma distributions (Xie et al., 2005; Faddy et al., 2009), to characterize LOS of patients. They failed to take into account and quantify the influence of various possible individual characteristics for improving LOS prediction. Second, PAC residents have multiple possible discharge dispositions. They may be discharged to residential community for further recovery or transferred to hospital due to occurrence of critical events (e.g., infection and fall). Community discharge and re/hospitalization are mutually exclusive events, wherein whichever comes first will terminate the NH dwelling duration of a resident. Existing LOS modeling approaches, such as regression-based methods (Carey, 2002; Kelly et al., 2010; Kramer and Zimmerman, 2010) and machine learning methods (Hachesu et al., 2013; Pendharkar and Khurana, 2014; Turgeman et al., 2017), mainly focused on predicting time-to-discharge without differentiating the disposition difference and overlooked the complexity arising in the competing risks between the dispositions. Thus, there is a need to develop an advanced LOS modeling approach for PAC residents that incorporates both individual characteristics and considers multiple competing discharge dispositions.

After realizing superior individual prediction of LOS and discharge disposition with relevant factors identified, there is still need to evaluate the facility-level LOS and discharge outcome performance for a population of residents with varied individual characteristics. This will better inform the NH on resource preparedness and evaluate the facility-level quality outcome. Computer simulation, such as discrete event simulation and agent-based simulation, are often considered in healthcare system engineering by modeling and simulating each individual patient as a discrete event or agent, as well as further evaluating the system level performance (e.g., average LOS of a population of individuals at the facility level) (Hoot et al., 2008; Taboada et al., 2011; Wang et al., 2012). Among these simulation approaches, a key step is to develop sampling algorithm for simulating LOS observations for a population of individuals. Existing sampling algorithms are only applicable to simulating LOS observations characterized by distribution-based models, which ignored various resident characteristics influencing the LOS (Cappanera et al., 2014; Zhang et al., 2019), or regression models (Austin et al., 2002) which did not take the multiple discharge dispositions into account. There is a need to develop a sampling algorithm, which allows simulations of LOS observations for a population of individuals with multiple competing discharge dispositions, as well as varied individual characteristics.

To fill the aforementioned research gaps, we propose a heterogeneous LOS modeling approach for PAC residents by taking multiple discharge dispositions into account

and incorporating the varied individual characteristics that may potentially influence each discharge disposition. At individual level, a semi-parametric hazard regression model is considered to characterize the heterogeneous LOS observations of PAC residents with improved prediction accuracy on individual re/hospitalization risk and community discharge likelihood. Various factors affecting re/hospitalization risk and/or community discharge likelihood are also identified with their influence quantified. At facility level, a simulation algorithm is further developed to realize the simulation of both LOS observations and multiple discharge dispositions for a heterogeneous population of PAC residents with accurate predicted samples obtained. A real case study using large-scale NH data is provided to demonstrate the superior prediction performance of the proposed work at both individual and facility levels.

The remaining of the paper is organized as follows. The next section first presents the proposed LOS modeling framework with an individual-level LOS model, which quantifies various influencing factors and considers competing discharge dispositions, as well as subsequently introduces the proposed sampling algorithm for facility-level performance evaluation. Then, a real-data case study is provided to demonstrate the efficacy of the proposed work. Conclusive remarks are provided in the end.

2 Methodology

2.1 Model formulation

Considering a cohort of N PAC residents in an NH facility, each PAC resident may be discharged to residential community for further recovery or transferred to hospital due to critical events (e.g., infections and falls). Community discharge and re/hospitalization are mutually exclusive events, whichever comes first will terminate the NH dwelling duration of a resident. Unlike many of existing LOS modeling works (Faddy et al., 2009; Kelly et al., 2010), which focused on modeling a single discharge disposition, the proposed model formulation aims to take into account the multiple and competing discharge dispositions of PAC residents, namely, re/hospitalization and community discharge. Moreover, unlike many of existing discharge outcome prediction models, such as hospital re/admission models, which focused on predicting the risk of critical outcomes at a fixed time period, e.g., 30-day or 90-day re/hospitalization risk (Incalzi et al., 1992), the proposed model will capture the re/hospitalization risk as well as community discharge likelihood of each individual PAC resident over time. The instantaneous discharge rate of the ith PAC resident with discharge disposition type s (community (C) or hospital (H)) can be characterized as

$$d_{i,s}(t|\boldsymbol{x}_i)=\lim_{\Delta t\to 0}\frac{\Pr(t\leqslant T_{i,s}\leqslant t+\Delta t\,|\,T_{i,\min}\geqslant t,\ \boldsymbol{x}_i)}{\Delta t},\quad i=1,\ ...,\ n;\ s\in\{C,\ H\},\tag{1}$$

where $T_{i,\min}=\min\{T_{i,C},\ T_{i,H}\}$ is the LOS of resident i, and $T_{i,C}$ and $T_{i,H}$ are the latent time-to-discharge quantities with discharge disposition of community and hospital, respectively. $\boldsymbol{x}_i$ is a p_s-dimensional observed vector which contains varied individual covariates that may potentially influence $d_{i,s}(t)$, such as individual demographics, clinical diagnoses, cognitive deficits, and physical functional performance. To associate the individual characteristics explicitly with $d_{i,s}(t)$, the hazard regression is considered as follows

$$d_{i,s}(t)=d_s^b(t)\exp(\boldsymbol{\beta}_s^{\mathrm{T}}\boldsymbol{x}_i),\ i=1,\ ...,\ n;\ s\in\{C,\ H\},\tag{2}$$

where $d_s^b(t)$ is the population average instantaneous discharge rate with disposition s in the absence of the influence of $\boldsymbol{x}_i$. $\boldsymbol{\beta}_s$ is a p_s-dimensional disposition-specific coefficient vector that quantifies the influence of $\boldsymbol{x}_i$ on $d_{i,s}(t)$.

A benefit of the model in Eq. (2) is that, for any time t, individual re/hospitalization risk until time t, i.e., $\Pr(T_{i,H}\leqslant t)$, and community discharge likelihood till time t, i.e., $\Pr(T_{i,C}\leqslant t)$, can be written as

$$\Pr(T_{i,H}\leqslant t)=\int_0^t d_{i,H}(\tau|\boldsymbol{x}_i)\exp\left[-\int_0^\tau\sum_{s\in\{C,H\}}d_{i,s}(y|\boldsymbol{x}_i)\mathrm{d}y\right]\mathrm{d}\tau,\tag{3a}$$

$$\Pr(T_{i,C}\leqslant t)=\int_0^t d_{i,C}(\tau|\boldsymbol{x}_i)\exp\left[-\int_0^\tau\sum_{s\in\{C,H\}}d_{i,s}(y|\boldsymbol{x}_i)\mathrm{d}y\right]\mathrm{d}\tau.\tag{3b}$$

In other words, given a specific time period until time t, the proposed model can always be converted into evaluating both the re/hospitalization risk and community discharge likelihood of resident i at a fixed time period based on Eqs. (3a) and (3b). It bypasses the conventional discharge outcomes modeling approach, which requires the discretization of discharge outcomes data in advance based on a pre-specified time period and then formulates a classification model for outcome prediction. It also allows the comparison of re/hospitalization risk and community discharge likelihood among different individuals with varied individual characteristics $\boldsymbol{x}_i$ over time.

2.2 Model estimation

Given observed data $\boldsymbol{D}=\{t_i,\ z_{i,s},\ \boldsymbol{x}_i\}_{i=1}^n$, where $z_{i,s}=1$ if PAC resident i is discharged to disposition s; 0 otherwise, $s\in\{C,\ H\}$, with unknown parameters/functions $\boldsymbol{\theta}=\cup_{s\in\{C,H\}}\boldsymbol{\theta}_s$, where $\boldsymbol{\theta}_s=\{d_s^b(t),\ \boldsymbol{\beta}_s\}$. The joint likelihood function $L(\boldsymbol{\theta}|\boldsymbol{D})$ can be written as follows

$$L(\boldsymbol{\theta}|\boldsymbol{D}) = \prod_{i=1}^{n} \prod_{s\in\{C,H\}} \left\{ d_{i,s}^{b}(t_i) \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_i) \cdot \exp\left[- \sum_{s\in\{C,H\}} \int_0^{t_i} d_{i,s}^{b}(\tau) \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_i) \mathrm{d}\tau \right] \right\}^{z_{i,s}}. \quad (4)$$

Let index set $A_s = \{i : z_{i,s} = 1\}$, $s \in \{C, H\}$. Then the joint likelihood function can be simplified as

$$L(\boldsymbol{\theta}|\boldsymbol{D}) = \prod_{i\in A_C} d_{i,C}^{b}(t_i) \exp(\boldsymbol{\beta}_C^{\mathrm{T}} \boldsymbol{x}_i) \cdot \prod_{i\in A_H} d_{i,H}^{b}(t_i) \exp(\boldsymbol{\beta}_H^{\mathrm{T}} \boldsymbol{x}_i) \cdot \prod_{i=1}^{n} \exp\left[- \sum_{s\in\{C,H\}} \int_0^{t_i} d_{i,s}^{b}(\tau) \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_i) \mathrm{d}\tau \right]. \quad (5)$$

When $\boldsymbol{\theta}_s$ are mutually exclusive, $L(\boldsymbol{\theta}|\boldsymbol{D})$ can be multiplicatively decomposed into $L(\boldsymbol{\theta}|\boldsymbol{D}) = \prod_{s\in\{C,H\}} L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$, in which $L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$ can be expressed as

$$L_s(\boldsymbol{\theta}_s|\boldsymbol{D}) = \prod_{i\in A_s} d_{i,s}^{b}(t_i) \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_i) \cdot \prod_{i=1}^{n} \exp\left[- \int_0^{t_i} d_{i,s}^{b}(\tau) \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_i) \mathrm{d}\tau \right], s \in \{C, H\}. \quad (6)$$

Thus, maximizing $L(\boldsymbol{\theta}|\boldsymbol{D})$ can be equivalent to maximizing $L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$, separately. To maximize $L_s(\boldsymbol{\theta}_s|\boldsymbol{D})$ by treating $d_s^b(t)$ as an unknown function, we will first maximize the partial likelihood $L_s(\boldsymbol{\beta}_s|\boldsymbol{D})$ written as (Cox, 1972)

$$L_s(\boldsymbol{\beta}_s|\boldsymbol{D}) = \prod_{i\in A_s} \frac{\exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_i)}{\sum_{j\in B(t_i)} \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_j)}, \ s \in \{C, H\}, \quad (7)$$

where $B(t_i)$ is a set of residents who are still in the NH before t_i. Maximum likelihood estimation will be considered by solving $\max_{\boldsymbol{\beta}_s} \log L_s(\boldsymbol{\beta}_s|\boldsymbol{D})$, which can be realized by a numerical optimization algorithm, such as the Newton-Raphson method (Gonzàlez et al., 2008).

Based on the estimated $\boldsymbol{\beta}_s$, we will estimate $d_s^b(t)$ by maximizing the profile likelihood $L_s(d_s^b|\boldsymbol{D})$ written as

$$L_s\left(d_s^b|\boldsymbol{D}\right) \propto \prod_{i\in A_s} d_{s,i}^{b} \exp\left[-d_{s,i}^{b} \sum_{j\in B(t_i)} \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_j) \right], \ s \in \{C, H\}, \quad (8)$$

where $d_{s,i}^b = d_s^b(t_i)$, $i \in A_s$, and $\hat{d}_s^b = 0 \forall t \notin \{t_i\}_{i\in A_s}$. The profile maximum likelihood estimator can be obtained as (Cole et al., 2014)

$$\hat{d}_{s(t_i)}^{b} = \frac{1}{\sum_{j\in B(t_i)} \exp(\boldsymbol{\beta}_s^{\mathrm{T}} \boldsymbol{x}_j)}, \ s \in \{C, H\}. \quad (9)$$

2.3 Sampling algorithm

Based on the proposed model formulation and developed estimation algorithms, as illustrated in Sections 2.1 and 2.2, the re/hospitalization risk and community discharge likelihood over time of any individual i with individual characteristics $\boldsymbol{x}_i$ can be predicted. To further investigate the service utilization of the sample of PAC residents in an NH facility and evaluate the facility-level performance for a heterogeneous population of PAC residents with varied individual characteristics, it becomes important to utilize computer simulation to mimic the patient flow of individual PAC residents and evaluate the system performance at the aggregate level. An essential basis of such computer simulation requires simulating the realization of LOS for each PAC resident. Existing simulation algorithm for LOS models mainly focused on simulating realization based on distribution-based models (El-Darzi et al., 1998; McGuire, 2007; New et al., 2015), such as Weibull distribution and Log-normal distribution. For the developed semi-parametric regression models with multiple competing discharge dispositions in Section 2.1, existing sampling algorithms are not applicable and there is a need to develop the corresponding simulation algorithm to facilitate generating predictive samples of LOS realizations given a heterogeneous population of PAC residents with varied individual characteristics $\boldsymbol{x}_i$. $T_{i,s}$ of resident i with disposition s would be simulated based on the developed sampling algorithm summarized as follows.

3 Case study

3.1 Data description

To demonstrate the performance of the proposed model

Proposed Sampling Algorithm

Step 1: Compute $\phi_i(t|\boldsymbol{x}_i) = \sum_{\forall s\in\{C,H\}} \hat{d}_s^b(t) \exp\left(\hat{\boldsymbol{\beta}}_s^{\mathrm{T}} \boldsymbol{x}_i\right)$, where $\phi_i(t|\boldsymbol{x}_i)$ is the instantaneous probability of resident i with $\boldsymbol{x}_i$ being discharged at time t;

Step 2: Compute $S_i^{(l)} = \exp\left[-\sum_{p=0}^{l} \phi_i\left(t_{(p)}|\boldsymbol{x}_i\right) \right]$, $l = 0, \ldots, N, N+1$, where $S_i^{(l)}$ is the probability of resident i still residing in NH at time $t_{(l)}$; $t_{(0)} < t_{(1)} < \ldots < t_{(l)} < \ldots < t_{(N)} < t_{(N+1)}$ and $t_{(0)} = 0$, $t_{(N+1)} = +\infty$; $\left\{t_{(l)}\right\}_{l=1}^{n}$ are ordered distinct historical LOS observations;

Step 3: Randomly generate $\mu = Unif(0, 1)$;

Step 4: Compute $l_1 = \max\left\{l : \mu \leqslant S_i^{(l)}\right\}$, $l_2 = \min\left\{l : \min S_i^{(l)} \leqslant \mu\right\}$, and get simulated LOS, T_i as $T_i = \frac{S_i^{(l)} - \mu}{S_i^{(l_1)} - S_i^{(l_2)}} \cdot \left(t_{(l_2)} - t_{(l_1)}\right) + t_{(l_1)}$;

Step 5: Determine disposition state s by drawing for the categorical distribution as $w_s = \frac{\hat{d}_s^b(t) \exp\left(\hat{\boldsymbol{\beta}}_s^{\mathrm{T}} \boldsymbol{x}_i\right)}{\phi_i(t|\boldsymbol{x}_i)}$, $s \in \{C, H\}$, i.e., $s \sim Categorical(w)$, where $w = [w_C, w_H]^{\mathrm{T}}$

and sampling algorithm, the Minimum Data Set (MDS) 3.0 of a certified NH in Tampa Bay Area, Florida, is considered. The MDS 3.0 is a rich data set containing comprehensive assessment of clinical and functional status of all residents in a Medicare/Medicaid-certified NH during their stays. The dataset is mandated federally and is required by the Centers for Medicare and Medicaid Services (CMS) (CMS, 2017). Each resident is assessed upon admission, periodically during the stay, and upon discharge or in case of any event causing significant change in their functional status. Each assessment contains over 680 data covariates representing information on identification, admission and discharge dates, socio-demographics, financial details, various functional performance metrics, diseases and chronic conditions, medication and therapy information, discharge outcomes/dispositions of each resident, and facility administrative details.

The data collected includes stays of all residents admitted to the NH in a one-year period. For this case study, a sub-cohort of PAC residents are selected according to "short-stay" criteria defined by the CMS (CMS, 2013). The CMS differentiates between "short-stay" and "long-stay" residents by examining episodes of care of the resident. An episode consists of one or more consecutive stays with breaks no more than 30 days. If the cumulative LOS(s) in the NH is equal to or less than 100 days, the resident is labelled as a "short-stayer", otherwise considered as a "long-stayer". Most recent episode coinciding with the end-date of the consideration time period is used for the categorization. In the selected data, each data instance refers to the LOS observation of a short-stay resident with his/her individual characteristics.

A total of 710 LOS observations with complete information from 611 individual residents are selected for analysis. 98.02% of the LOS observations may be considered post-acute, which meant that the resident was either admitted directly from the hospital to the NH, and/or covered under Medicare Part A insurance plan (Holup et al., 2017). LOS observations with community discharge or re/hospitalization are included. LOS observations with other discharge dispositions, such as death and transfer to another facility, are excluded because they form a very small portion in the dataset with negligible influence on the overall model building. Left-truncated and/or right-censored observations are also neglected due to their negligible portions.

Table 1 provides a summary of descriptive statistics of the selected cohort and stays, which includes socio-demographics (e.g., age, gender, ethnicity, marital status, and so on), care utilization details of the stay (e.g., LOS, admission origin, discharge disposition, payment source, and so on), and health characteristics (e.g., body measures, various functional performance scores, disease and chronic conditions, and so on). The calculated mean LOS of the short-stay residents was 20.33 days, with a

Table 1 Descriptive summary statistics of the selected resident cohort

Characteristics	Mean (SD) or %
Number of stays	710
Number of residents	611
Demographics	
Age at admission (years)	76.68 (10.66)
Gender: Female	64.40%
Race	
Black or African American	6.90%
White	90.30%
Others	2.80%
Marital status	
Never married	14.20%
Married	35.60%
Widowed	33.20%
Divorced	15.60%
Others	1.40%
Care utilization	
LOS (days)	20.33 (15.72)
Admission origin	
Community	2.00%
Hospital	97.30%
Others	0.70%
Discharge disposition	
Community	79.90%
Hospital	20.10%
Primary payer: Medicare Part A	59.00%
Health characteristics	
Height (inches)	65.36 (4.13)
Weight (pounds)	170.95 (54.45)
ADL score	6.48 (3.85)
Mood/Depression score	0.91 (1.24)
Cognitive score	13.17 (2.96)
Visual impairment	15.90%
Hearing impairment	22.70%
Incontinence – urinary	54.50%
Incontinence – fecal	46.20%
Fall within past 180 days	32.70%
Fracture within past 180 days	18.00%
Diseases	
Cancer	8.00%
Heart/Circulation	79.40%
Gastrointestinal	38.60%
Genitourinary	25.20%
Metabolic	73.80%
Musculoskeletal	42.00%
Neurological	26.20%
Psychiatric/Mood disorder	45.90%
Pulmonary	34.60%

Notes: SD: Standard deviation; ADL: Activities of daily living.

majority of 97.3% being admitted from hospital, a majority of 79.9% being discharged to community and the rest 20.1% being readmitted to hospital. The discharge disposition, socio-demographics, and health characteristics form possible covariates influencing LOS of the resident and consequentially, their care service utilization.

3.2 Feature selection

Because the MDS dataset contains numerous elements of data, consider information directly relevant to the LOS. Guided by domain knowledge in NH care, a subset of data most related to care utilization, i.e., socio-demographics, functional performance scores, disease diagnoses and chronic conditions observed on admission, is considered. Although MDS data monitors the stay over time, only assessments upon admission are relevant to LOS prediction of an unknown cohort of residents in the facility, which are also known as baseline observations.

After summarizing the data (i.e., calculating LOS, deriving various functional performance scores and converting categorical covariates into dummy variables) and preprocessing (i.e., removing low-frequency covariates, removing one from each highly correlated pair, and checking for multicollinearity), 68 covariates relevant to LOS are selected. It is still a large number of covariates and incorporating all of them for developing a predictive model may yield model overfitting. To reduce the dimensionality of the input variables, a collection of 10 popular feature selection methods are applied to the dataset.

The feature selection methods include four linear feature selection methods, which are stepwise regression (e.g., Stepwise Akaike Information Criterion (AIC)), recursive feature elimination (RFE), simulated annealing (SA), and regularized linear regression (e.g., LASSO). Each feature selection method implements different algorithm to determine the best subset of covariates. For example, Stepwise AIC trains linear regression models by progressively adding covariates and evaluating model performance with AIC. By contrast, RFE ranks all covariates, progressively removes unimportant ones, trains, and reevaluates a linear model at each step. SA performs a random heuristic search for best combination in the covariate space. LASSO regression penalizes unimportant covariates to zero coefficient value with L1-norm regularization.

Moreover, six nonlinear feature selection methods are applied, including Filtering with Random Forest, RFE with Bagged Trees, RFE Random Forest, Genetic algorithm with Random Forest, SA with Random Forest, and Boruta. Filtering uses a preprocessing step to test strength of individual relationship between each covariate and the response variable before training a predictive model. RFE, Genetic and SA use a subset selection heuristic similar to that applied in training linear models. However, tree-based algorithms are trained instead of linear models at each step. In each case, the tree-based model with the highest accuracy evaluated identifies the best subset of covariates.

The most significant covariates influencing LOS are identified by each of the above feature selection methods. To keep most of the information without missing any relevant covariates, the union of all feature selection results, namely, 35 covariates in total, are considered for further predictive modeling. Table 2 displays the final selected covariates, while Table 3 shows the significant covariates identified by each feature selection algorithm.

3.3 Prediction performance comparison

To compare the prediction performance between the proposed model and alternative prediction methods in the literature, the dataset is split randomly into 90% training and 10% test datasets of observations with stratified sampling to preserve the proportion of discharge dispositions. The previous section of feature selection is conducted based on the training data set without touching the independent test dataset. The proposed model is compared to others by evaluating the C-index values of training and test datasets for each discharge dispositions, which are community and hospital (Harrell et al., 1996; D'Agostino and Nam, 2003). A C-index value beyond 0.5 indicates that the model is consistently satisfactory in predicting discharge risks, rather than making random predictions. A higher C-index value indicates the better predictive capability. The proposed model takes about 0.726 seconds by fitting all the LOS observations and making the predictions, which is efficient for real applications. Several semi-parametric and parametric survival models are compared under the competing risk framework, where the characteristic of competing discharge dispositions, i.e., community discharge and re/hospitalization, is incorporated. Semi-parametric models include the Cox regression with LASSO, or Elastic Net regularization, where the baseline hazard is non-parametric, while regularization attempts to avoid over-fitting. Parametric models comprise Weibull, Logistic, Log-normal, Log-logistic, and Exponential hazard regression, where the baseline hazards are parameterized based on the named distributions. Furthermore, several alternative regularized/unregularized linear and non-linear machine learning methods independent of competing risk are considered, including the linear regression, LASSO regression, Ridge regression, Tobit regression, Decision Tree, Boosting Tree and Random Forest. Thus, a total of 15 different models are evaluated for each of the discharge dispositions based on the same 35 covariates identified in the previous section. Table 4 provides the list of models considered with their abbreviations and corresponding training and test C-index values. Figure 1 further visualizes the results.

As observed in Fig. 1(a), for predicting LOS before

Table 2 List of 35 selected covariates with the associated descriptions

Covariate short name	MDS 3.0 covariate description
Age	Age at admission (years)
ADL score	Activities of daily living score at admission
Cognitive score	Brief interview for mental status (BIMS) summary score at admission
Mood score	Resident mood interview patient health questionnaire (PHQ)-9 total severity score at admission
	Active diagnose indicator of resident at admission
Cancer	Cancer (with or without metastasis)
Anemia	Anemia (e.g., aplastic, iron deficiency, pernicious, and sickle cell)
Atrial fibrillation	Atrial fibrillation or other dysrhythmias (e.g., bradycardias and tachycardias)
Coronary artery disease	Coronary artery disease (CAD) (e.g., angina, myocardial infarction, and atherosclerotic heart disease (ASHD))
DVT, PE, PTE	Deep venous thrombosis (DVT), Pulmonary embolus (PE), or Pulmonary thrombo-embolism (PTE)
Heart failure	Heart failure (e.g., congestive heart failure (CHF) and pulmonary edema)
Hypertension	Hypertension
GERD or Ulcer	Gastroesophageal reflux disease (GERD) or Ulcer (e.g., esophageal, gastric, and peptic ulcers)
UC, CD, IBD	Ulcerative colitis (UC), Crohn's disease (CD), or Inflammatory bowel disease (IBD)
BPH	Benign prostatic hyperplasia (BPH)
Renal disease	Renal insufficiency, Renal failure, or End-stage renal disease (ESRD)
Neurogenic bladder	Neurogenic bladder
Obstructive uropathy	Obstructive uropathy
MDRO	Multidrug-resistant organism (MDRO)
Pneumonia	Pneumonia
Septicemia	Septicemia
Diabetes mellitus	Diabetes mellitus (DM) (e.g., diabetic retinopathy, nephropathy, and neuropathy)
Hyponatremia	Hyponatremia
Hyperlipidemia	Hyperlipidemia (e.g., hypercholesterolemia)
Arthritis	Arthritis (e.g., degenerative joint disease (DJD), osteoarthritis, and rheumatoid arthritis (RA))
Hip fracture	Hip fracture (e.g., sub-capital fractures, and fractures of the trochanter and femoral neck)
Other fracture	Other fracture
Alzheimer's disease	Alzheimer's disease
Non-Alzheimer's dementia	Non-Alzheimer's dementia (e.g., Lewy body dementia, vascular or multi-infarct dementia; mixed dementia; frontotemporal dementia such as Pick's disease; and dementia related to stroke, Parkinson's or Creutzfeldt-Jakob diseases)
Hemiplegia or Hemiparesis	Hemiplegia or Hemiparesis
Malnutrition	Malnutrition (protein or calorie) or at risk for malnutrition
Anxiety disorder	Anxiety disorder
Depression	Depression (other than bipolar)
Schizophrenia	Schizophrenia (e.g., schizoaffective and schizophreniform disorders)
PTSD	Post-traumatic stress disorder (PTSD)
Respiratory failure	Respiratory failure

community discharge, the proposed model outperforms other models with the training and test C-index values of 0.75 and 0.76, respectively. The regularized Cox regression models, i.e., LASSO and Elastic Net, yielded lower C-index values. Both values are still above 0.5, indicating that Cox baseline hazard is flexible in representing the LOS data with improved prediction performance. The reduced performance with regularization suggests that penalization of covariate coefficients in the Cox model is unnecessary, probably because an optimal set of covariates has been chosen based on the previous step of feature selection. Parametric survival models have poorer performance than the Cox model family with C-index values ranging from 0.21 to 0.25, which are much lower than 0.5, indicating that the models are consistently poor at prediction than a random guess. The Cox model family outperforms parametric survival models because its baseline hazard is non-parametric and is able to capture LOS data

Table 3 Summary of covariates identified by various feature selection methods

Covariate	Vote	Feature selection methods									
		Linear				Nonlinear					
		Stepwise AIC	RFE: Linear	SA: Linear	LASSO	Filtering: Random Forest	RFE: Bagged Trees	RFE: Random Forest	Genetic: Random Forest	SA: Random Forest	Boruta
Community											
ADL score	80%	▲		▲	▲	▲	▲	▲	▲		▲
Cognitive score	70%	▲				▲	▲	▲	▲	▲	▲
Mood score	80%	▲		▲	▲	▲	▲	▲	▲		▲
Cancer	70%	▲	▲		▲		▲	▲	▲	▲	
Anemia	70%			▲	▲	▲	▲	▲	▲		▲
Atrial fibrillation	60%			▲		▲	▲	▲	▲	▲	
Coronary artery disease	30%					▲	▲		▲		
DVT, PE, PTE	80%	▲	▲		▲	▲	▲	▲	▲		▲
Heart failure	70%				▲	▲	▲	▲	▲	▲	▲
Hypertension	100%	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲
GERD or Ulcer	40%					▲	▲	▲			▲
UC, CD, IBD	30%	▲	▲		▲						
BPH	50%		▲		▲	▲	▲	▲			
Renal disease	80%	▲	▲	▲	▲	▲	▲	▲			▲
Neurogenic bladder	60%		▲	▲	▲			▲	▲		▲
Obstructive uropathy	50%		▲	▲	▲		▲		▲		
MDRO	40%		▲		▲			▲			▲
Pneumonia	50%		▲		▲	▲	▲		▲		
Diabetes mellitus	40%				▲		▲	▲	▲		
Hyponatremia	90%	▲	▲		▲	▲	▲	▲	▲	▲	▲
Hyperlipidemia	60%				▲	▲	▲	▲	▲		▲
Arthritis	30%					▲	▲	▲			
Hip fracture	90%	▲	▲	▲	▲	▲	▲	▲		▲	▲
Other fracture	60%	▲	▲		▲	▲	▲	▲			
Alzheimer's disease	60%	▲	▲		▲		▲	▲			▲
Non-Alzheimer's dementia	90%	▲	▲		▲	▲	▲	▲	▲	▲	▲
Hemiplegia or Hemiparesis	80%	▲	▲		▲	▲	▲	▲	▲		▲
Malnutrition	30%						▲	▲	▲		
Anxiety disorder	30%			▲			▲		▲		
Depression	40%					▲	▲	▲			▲
Schizophrenia	60%	▲	▲	▲	▲	▲	▲				
PTSD	40%				▲	▲	▲			▲	
Respiratory failure	30%		▲	▲	▲						
Hospital											
Age	30%	▲			▲		▲				
ADL score	50%	▲		▲			▲	▲		▲	
Cognitive score	20%						▲	▲			
Hypertension	50%	▲		▲				▲		▲	▲
GERD or Ulcer	30%			▲			▲			▲	
Obstructive uropathy	20%			▲						▲	
Septicemia	20%								▲		▲
Hyponatremia	20%	▲		▲							
Hyperlipidemia	30%						▲	▲			▲
Arthritis	30%	▲					▲		▲		
Non-Alzheimer's dementia	40%	▲						▲		▲	▲
Hemiplegia or Hemiparesis	20%			▲						▲	
PTSD	30%						▲		▲		▲
Respiratory failure	40%	▲		▲						▲	▲

Table 4 Prediction performance (C-index) comparison between the proposed and alternative models

Model family	Model short name	Model description	Discharge dispositions			
			Community		Hospital	
			Train	Test	Train	Test
With competing risks assumption						
Semi-parametric survival	**SP.Cox**	**Proposed: Cox regression with non-parametric baseline hazard**	**0.758**	**0.768**	**0.707**	**0.699**
	SP.Cox. Lasso	Cox regression with L1-regularization	0.728	0.747	0.500	0.500
	SP.Cox.Elastic Net	Cox regression with L1/L2-mixed regularization	0.669	0.677	0.669	0.677
Parametric survival	P.Exponential	Survival regression with exponential baseline hazard	0.241	0.231	0.292	0.303
	P.Weibull	Survival regression with Weibull baseline hazard	0.248	0.248	0.293	0.319
	P.Logistic	Survival regression with Logistic baseline hazard	0.237	0.206	0.291	0.327
	P.Log.logistic	Survival regression with Log-logistic baseline hazard	0.237	0.204	0.284	0.308
	P.Log.normal	Survival regression with Log-normal baseline hazard	0.239	0.215	0.281	0.313
Independent of competing risks						
Data mining: Linear	ML.Tobit Reg	Tobit regression	0.243	0.224	0.288	0.324
	ML.Linear Reg	Linear regression	0.265	0.229	0.334	0.318
	ML.Lasso	Linear regression with L1-regularization	0.284	0.242	0.500	0.500
	ML.Ridge	Linear regression with L2-regularization	0.266	0.225	0.347	0.352
Data mining: Tree-based	ML.R.Forest	Random Forest regression	0.290	0.220	0.491	0.261
	ML.Boosting Tree	Boosting Tree regression	0.229	0.245	0.238	0.170
	ML.Tree	Decision Tree regression	0.305	0.311	0.308	0.352

with more flexibility. Regularized/Unregularized linear models perform slightly better than survival models with C-index values ranging between 0.22 and 0.28. Conversely, tree-based methods perform better than linear models with Decision Tree and Random Forest producing C-index around 0.3. The improvement achieved from tree-based methods indicate a non-linear relationship between LOS and the covariates.

Similarly, from Fig. 1(b), the proposed model outperforms other models for predicting LOS before transferring to hospital. The performance patterns are similar to those of predicting community discharge likelihood with a few differences. LASSO regularization in Cox regression performs poorly for predicting re/hospitalization risk, because a minimum penalty term was not found to be more effective for improving prediction than a random guess. Within linear models, LASSO regularization produces improved results, but were still inadequate for accurate prediction. Within the nonlinear models, the Decision Tree produces the best result. The test C-index value is generally lower than the training C-index value, because the test dataset is serving as an independent dataset untouched during the model development phase to evaluate the future prediction performance of the model.

3.4 Identification of risk/protective factors

Apart from producing superior prediction performance, the proposed competing risk Cox regression model identifies important risk/protective factors that influence a resident's LOS. Table 5 shows the significant covariates identified by the proposed model for predicting community discharge likelihood and re/hospitalization risk. The significance level α is set at 0.05.

The magnitude and sign of the coefficient values quantify the influence of the covariates on the probability of being discharged/transferred. A higher probability of being discharged/transferred implies a shorter LOS, and vice versa. For a resident being ultimately discharged to community, if s/he has higher ADL score, higher Mood score, or any of the disease diagnoses, her/his discharge likelihood decreases due to the negative sign and the LOS increases. Alternately, if a resident is ultimately transferred to hospital, having a higher ADL, or being diagnosed with anemia, uropathy or diabetes, her/his hospitalization risk will be increased due to her/his positive signs, implying a shorter LOS before being transferred to hospital. For both community and hospital dispositions, ADL is the most significant factor on influencing LOS, confirming the domain knowledge that residents with high dependency for daily living activities (e.g., eating, bathing, toileting, dressing, etc.) require greater NH care. ADL also has an opposite effect on dispositions, indicating the importance of incorporating multiple discharge dispositions. Such identified risk/protective factors are valuable for the healthcare provider to better identify and target on the most "at-risk" NH residents with more focused care and resources.

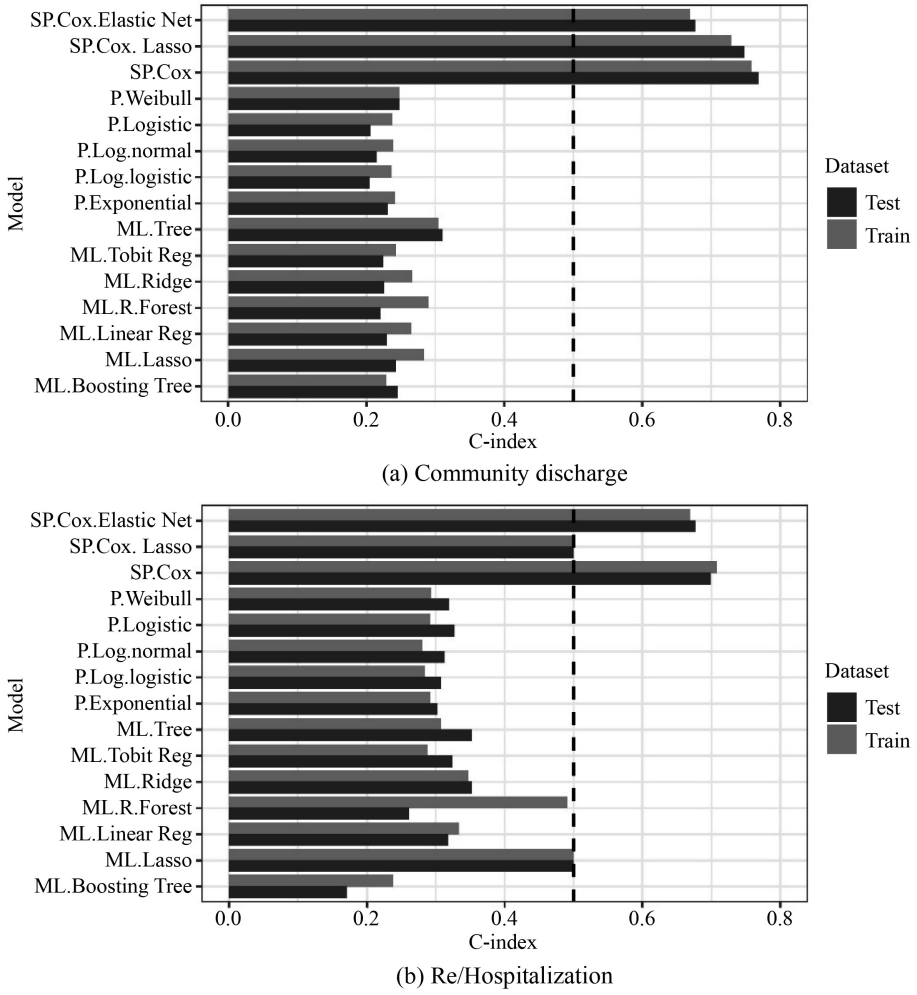


Fig. 1 Prediction performance (C-index) comparison under different discharge dispositions.

3.5 Marginal effects of covariates on community discharge likelihood and re/hospitalization risk

As opposed to a single value to quantify predicted LOS obtained from many existing predictive models, the proposed competing risk model further provides information on disposition-specific probability of being discharged/transferred. Such information can be visualized and compared by plotting the survival probability (i.e., 1-probability of being discharged/transferred) over time. Furthermore, because the proposed model is a proportional hazard model, marginal effects of survival curves can be visualized under different values of covariates. Based on such survival curves, the influence of each individual covariate on the probability of being discharged/transferred can be visualized. The probability of being discharged/transferred over time among different individuals with different individual characteristics can be visualized and compared as well. Figures 2 and 3 provide examples of marginal effects of various baseline ADL values on the LOS of an example resident over time for specific discharge dispositions. All variables other than the ADL score is fixed at the mean level of the observed sample.

As observed in Fig. 2, a community discharge resident with a higher baseline ADL score (more physical functional dependency) has a curve (red) higher than the average (blue) resident. In such a case, the probability to remain in the facility is higher than average at any point in time, which further increases the LOS. By contrast, a resident with a lower baseline ADL score (more functionally independent) has a survival curve (green) lower than average and tends to have a shorter LOS. Figure 3 shows the survival curves for a re/hospitalized resident. ADL score has an opposite effect on the curves, reaffirming the competing risk assumption. A higher baseline ADL score results in a shorter stay, while a lower one increases the stay. Because the curves evolve differently over time, it is possible to assess the probability of being discharged/transferred at any time point during the resident's stay.

Similarly, a resident's disposition outcome may also be examined over time for various combinations of varied individual characteristics. For instance, in Fig. 4, a hypothetical resident with better health conditions (e.g., lower ADL and Mood/Depression scores, and less number of

Table 5 Significant covariates identified by the proposed model for each disposition

Covariate	$\hat{\kappa}_j$	SE ($\hat{\kappa}_j$)	p-value	
Community discharge				
ADL score	−0.113	0.014	1.11E-15	***
Mood score	−0.143	0.041	0.0005	***
Cancer	−0.436	0.162	0.0072	**
Anemia	−0.203	0.097	0.0367	*
Hypertension	−0.559	0.100	0	***
BPH	−0.515	0.141	0.0003	***
Renal disease	−0.330	0.134	0.0136	*
MDRO	−0.628	0.295	0.0331	*
Hip fracture	−0.604	0.233	0.0096	**
Other fracture	−0.421	0.127	0.0009	***
Non-Alzheimer's dementia	−0.448	0.140	0.0014	**
Hemiplegia or Hemiparesis	−0.846	0.239	0.0004	***
Malnutrition	−0.556	0.196	0.0045	**
Re/Hospitalization				
ADL score	0.087	0.024	0.00026	***
Anemia	0.482	0.180	0.00727	**
Obstructive uropathy	1.028	0.307	0.00080	***
Diabetes mellitus	0.503	0.170	0.00311	**

Notes: 1) * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 2) SE: Standard error; 3) 95% confidence intervals for each parameter estimate are calculated by $\hat{\kappa}_j \pm 1.96 \times \mathrm{SE}(\hat{\kappa}_j)$, where $\hat{\kappa}_j$ is the respective estimated covariate coefficient.

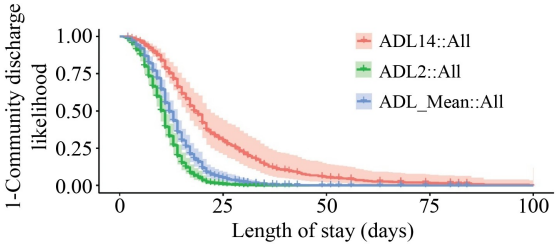


Fig. 2 Marginal effect of ADL score on survival curves for community discharge.

diseases diagnosed at baseline) tends to have a shorter LOS with community discharge (red curve) but a longer LOS with re/hospitalization (blue). By contrast, in Fig. 5, a hypothetical resident with worse health conditions (e.g., higher ADL and Mood/Depression scores, and more diseases diagnosed at baseline) tends to remain in the facility for a long time for recovery before being discharged to the community (red), and a relatively short stay if being transferred to hospital (blue).

3.6 Performance of the proposed sampling algorithm for generating LOS predictive samples

Survival models are different in predicting response compared with conventional machine learning models.

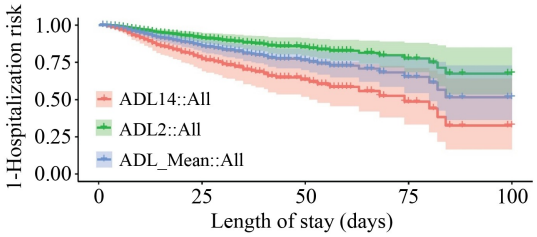


Fig. 3 Marginal effect of ADL score on survival curves for re/hospitalization.

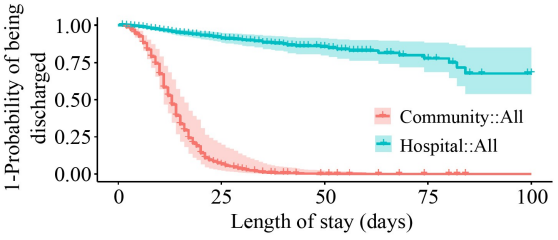


Fig. 4 Survival curves of a healthy resident under different discharge dispositions.

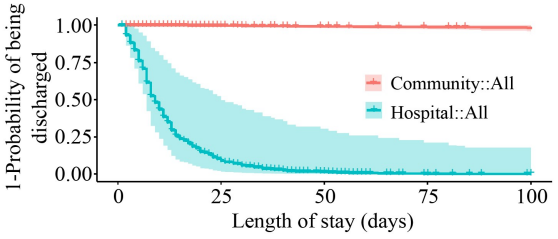


Fig. 5 Survival curves of an unhealthy resident under different discharge dispositions.

The former models characterize the predictive distribution by providing a probabilistic prediction, e.g., the predicted probability of being discharged/transferred over time for each of the resident, while the latter models often provide a single point prediction quantity, e.g., the predicted LOS value. To simulate the residents' flow in a typical NH facility using computer simulation, an important step is to simulate LOS predictive samples accurately. Our proposed sampling algorithm is capable of generating predictive LOS samples accurately and simultaneously providing the corresponding discharge dispositions as well. Sampling performance may be evaluated by comparing survival plots of observed LOS samples and simulated LOS samples. The survival curves are calculated by the Kaplan-Meier curves, which provide the disposition specific observed and simulated survival curves of a sample. In Fig. 6, survival curves are compared for each disposition and the full dataset. The sampling algorithm is very effective in generating predictive samples of LOS, because the simulated (light-colored) curves are very close to their observed ones (dark-colored). The green curves (light and dark) are slightly lower than the full

dataset (black and grey), indicating that residents transferred to hospital tend to have shorter LOSs than the average, as opposed to blue curves, indicating that residents discharged to the community have slightly longer LOSs than the average. Figure 7 shows the performance of the sampling algorithm in predicting discharge dispositions with 100% classification accuracy.

The accuracy of the proposed algorithm is further compared with the simulation results based on several alternative models, such as survival models and machine learning models. Simulation is performed using the Cox Weibull regression and Log-normal regression, where the baseline hazard functions are fitted with Weibull and Log-normal distributions, respectively, and under the competing risk framework. As shown in the survival curves of Figs. 8–10, the Cox Log-normal regression performs poorly in generating samples as compared to the observed data, while Cox Weibull performs better due to its increased flexibility in fitting the LOS data. The prediction performance of the proposed work is also compared with popular linear and non-linear machine learning models, such as linear regression, L1-regularized linear regression (LASSO), Decision Tree, and Random Forest. Overall, the proposed work generates the most accurate LOS samples, when compared with other methods due to the incorporation of non-parametric baseline hazard as well as the proposed simulation algorithm.

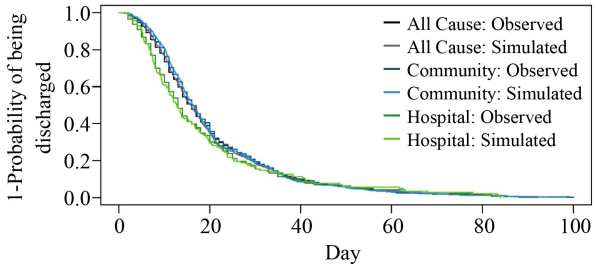


Fig. 6 Sampling algorithm prediction performance.

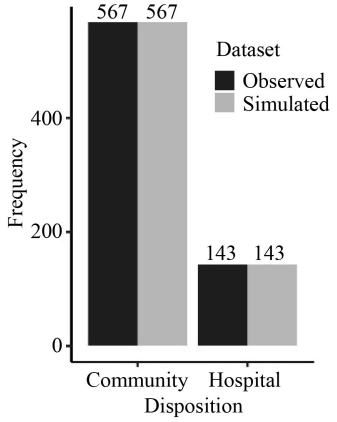


Fig. 7 Discharge disposition prediction.

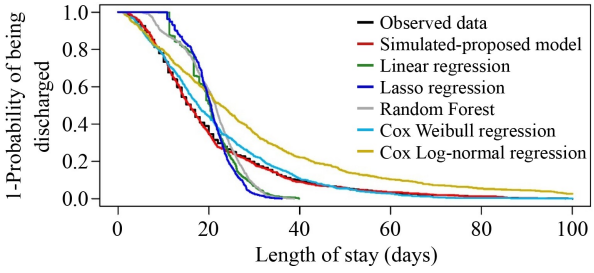


Fig. 8 Comparison of prediction performance between the proposed sampling algorithm and alternative models for all discharge dispositions.

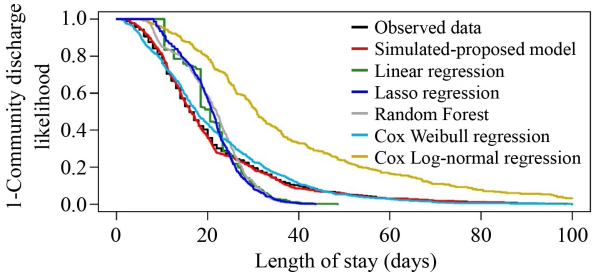


Fig. 9 Comparison of prediction performance between the proposed sampling algorithm and alternative models for community discharge disposition.

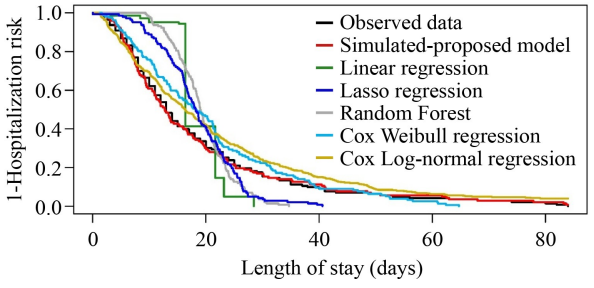


Fig. 10 Comparison of prediction performance between the proposed sampling algorithm and alternative models for hospital transfer disposition.

3.7 Simulation-based facility-level performance evaluation

The proposed sampling algorithm is not only able to predict the probability of being discharged/transferred to a specific disposition for a specific individual resident, it can be further used to generate predictive LOS samples of a heterogeneous population of NH residents with varied individual characteristics. This will allow the users to evaluate the system level performance of an NH facility, given a census composition scenario of a heterogeneous population of NH residents.

To explain the functionality of the proposed work, eight different cohorts of residents are defined in an increasing order of acuity. Simulation data is generated for each cohort using various segmentation and distributions of the significant covariates identified by Cox regression in Table 5. The setting for each cohort is

provided in Table 6. For each acuity scenario, 1000 random admission observations are generated in the following process. First, the ADL score is randomly generated with a truncated normal distribution with mean fixed at desired level corresponding to acuity and upper and lower limits set with range of observed data. Second, the Mood score is randomly generated similarly with a truncated normal distribution. Lastly, for each observation, any of the 13 diseases are randomly selected and binary value (0 or 1) is generated through a Bernoulli distribution, where the rate of success is sampled from a Beta prior distribution with shape parameters set according to the desired acuity level. The sampling process is repeated 20 times for each acuity scenario to account for the stochastic simulation uncertainty. Mean LOS values and disposition-specific discharge rates are estimated with standard errors reported as well. The results obtained are summarized in Table 7 and Fig. 11. The computational cost of generating 1000 predictive LOS samples is around 0.7 seconds under various acuity scenarios, which is quite time efficient for real application.

The proposed sampling algorithm generates the LOS and predicts the discharge disposition for each simulated resident in each cohort. As seen in Fig. 11(a), as acuity increases, the mean LOS increases in the samples. More residents are transferred to hospital. Residents being discharged to community have increasingly longer stays, which further increased the mean LOS of the samples across acuity scenarios. The whiskers represent dispersion among samples in each acuity and are mostly non-overlapping, indicating a significant difference of LOSs between acuity scenarios. As shown in Fig. 11(b), increasing acuity has a sharper decreasing effect on community discharge rates over time, while hospital discharge rates increase at a more gradual rate. The phenomenon occurs because a larger number of diseases influenced the community discharge likelihood than re/hospitalization risk. The results further emphasized the competing nature of two dispositions. As the resident acuity increases, LOS tends to increase for residents discharged to community, while LOS tends to decrease for residents transferred to hospital. Depending on the

Table 6 Experimental settings of acuity scenarios in the simulation study

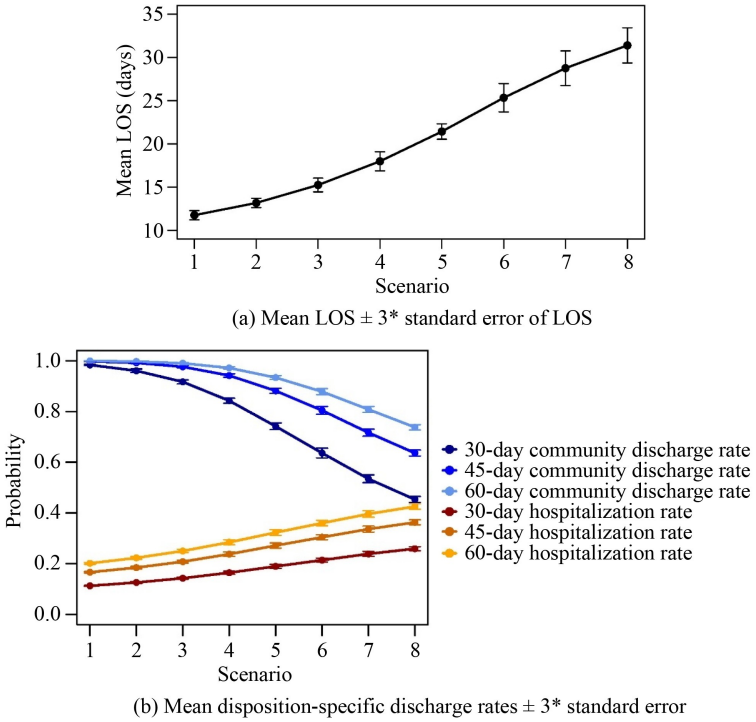
Covariate	Distribution	Parameter	Acuity scenario								Limits	SD
			Less acute				More acute					
			1	2	3	4	5	6	7	8		
ADL score	Truncated normal	Mean	1	3	5	7	9	11	13	15	[0, 16]	4
Mood score	Truncated normal	Mean	1	2	3	4	5	6	7	8	[1, 8]	2
Disease incidence prior	Beta	Mean	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	[0, 1]	
		SD	0.090	0.121	0.138	0.148	0.151	0.148	0.138	0.121		

Notes: SD: Standard deviation; Diseases include: Cancer, Anemia, Hypertension, BPH, Renal disease, MDRO, Hip fracture, Other fracture, Non-Alzheimer's dementia, Hemiplegia or Hemiparesis, Malnutrition, Obstructive uropathy, and Diabetes mellitus.

Table 7 Facility-level performance results under various census acuity scenarios of an NH

Metric		Measure	Acuity scenario							
			1	2	3	4	5	6	7	8
LOS		Mean	11.77	13.21	15.25	18.02	21.37	25.40	28.74	31.36
		SE	0.20	0.18	0.23	0.32	0.33	0.54	0.72	0.66
Disposition: Community	30-day discharge rate	Mean	0.98	0.96	0.92	0.84	0.74	0.64	0.53	0.45
		SE	0.0009	0.0023	0.0024	0.0042	0.0051	0.0066	0.0055	0.0043
	45-day discharge rate	Mean	1.00	0.99	0.98	0.94	0.88	0.80	0.72	0.64
		SE	0.0002	0.0010	0.0012	0.0026	0.0038	0.0056	0.0050	0.0045
	60-day discharge rate	Mean	1.00	1.00	0.99	0.97	0.93	0.88	0.81	0.74
		SE	0.00009	0.0005	0.0007	0.0017	0.0027	0.0045	0.0042	0.0041
Disposition: Hospital	30-day discharge rate	Mean	0.11	0.13	0.14	0.16	0.19	0.21	0.24	0.26
		SE	0.0007	0.0013	0.0015	0.0025	0.0030	0.0029	0.0037	0.0030
	45-day discharge rate	Mean	0.17	0.18	0.21	0.24	0.27	0.30	0.34	0.36
		SE	0.0010	0.0018	0.0021	0.0034	0.0038	0.0037	0.0047	0.0038
	60-day discharge rate	Mean	0.20	0.22	0.25	0.28	0.32	0.36	0.39	0.43
		SE	0.0012	0.0021	0.0024	0.0039	0.0042	0.0040	0.0052	0.0041

Note: SE: Standard error.



(a) Mean LOS ± 3* standard error of LOS

(b) Mean disposition-specific discharge rates ± 3* standard error

Fig. 11 Comparison of facility-level performance results under various census acuity scenarios.

proportion of the residents finally discharged to community or transferred to hospital, the mean LOS varies accordingly. The sampling algorithm successfully mimics the competing phenomenon of two dispositions. Furthermore, the algorithm can also provide disposition-specific probability of being discharged over time in a continuous time scale for the collective cohort and individual resident, allowing for a greater understanding of facility utilization and resident outcome (i.e., re/hospitalization risk) over the course of the stay. Figure 11(b) shows several discharge rate curves at discrete times of 30-, 45-, and 60-days.

4 Conclusions

In this paper, a heterogeneous LOS modeling approach was proposed by considering multiple discharge dispositions and incorporating varied individual characteristics for NH PAC residents. At individual level, several popular predictive models, such as machine learning and survival models, are considered to predict LOS and their performances are compared with the proposed model. The proposed model outperforms other models by jointly predicting the re/hospitalization risk and community discharge likelihood over time. It is also capable of identifying disposition-specific risk/protective factors for influencing the disposition-specific probability of being discharged/transferred over time. Furthermore, to enable the facility-level performance evaluation of the NH, a novel simulation algorithm was proposed for generating LOS predictive samples of residents by incorporating varied individual characteristics and competing discharge dispositions. The proposed algorithm is capable of accurately generating samples for a heterogeneous population of NH residents with varied individual characteristics, which allows the evaluation of facility performance measures, such as facility-level re/hospitalization rate and mean LOS. A real case study based on a large-scale de-identified data from an NH in Tampa Bay area was considered to illustrate the proposed work and demonstrate its superior prediction performance. The proposed approach would allow NH administrators and health practitioners to identify the most at-risk residents and design more targeted care delivery, facilitate optimal resource allocation strategies at the facility level for achieving greater quality outcomes at reduced costs, and further improve communication of prognostic information among everyone involved in the care delivery process.

References

Austin P C, Rothwell D M, Tu J V (2002). A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. Health Services and Outcomes Research Methodology, 3(2): 107–133

Cappanera P, Visintin F, Banditori C (2014). Comparing resource balancing criteria in master surgical scheduling: A combined optimisation-simulation approach. International Journal of Production Economics, 158: 179–196

Carey K (2002). Hospital length of stay and cost: A multilevel modeling

analysis. Health Services and Outcomes Research Methodology, 3(1): 41–56

Centers for Medicare and Medicaid Services (CMS) (2013). MDS 3.0 Quality Measures: User's Manual. Research Triangle Park, NC: RTI International

Centers for Medicare and Medicaid Services (CMS) (2017). Long-Term Care Facility Resident Assessment Instrument 3.0 User's Manual

Cole S R, Chu H, Greenland S (2014). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. American Journal of Epidemiology, 179(2): 252–260

Cox D R (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B, Methodological, 34(2): 187–202

D'Agostino R B, Nam B H (2003). Evaluation of the performance of survival analysis models: Discrimination and calibration measures. Handbook of Statistics, 23: 1–25

Eiken S, Sredl K, Gold L, Kasten J, Burwell B, Saucier P (2014). Medicaid Expenditures for Long-Term Services and Supports in FFY 2012. Bethesda, MD: Truven Health Analytics

El-Darzi E, Vasilakis C, Chaussalet T, Millard P H (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Management Science, 1(2): 143–149

Faddy M, Graves N, Pettitt A (2009). Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. Value in Health, 12(2): 309–314

Fashaw S A, Thomas K S, McCreedy E, Mor V (2020). Thirty-year trends in nursing home composition and quality since the passage of the Omnibus Reconciliation Act. Journal of the American Medical Directors Association, 21(2): 233–239

Ginsburg R B, Supreme Court of the US (1998). US Reports: Olmstead v. L.C., 527 US 581

Gonzàlez D, Piña M, Torres L (2008). Estimation of parameters in Cox's proportional hazard model: Comparisons between Evolutionary Algorithms and the Newton-Raphson Approach. In: Mexican International Conference on Artificial Intelligence. Berlin, Heidelberg: Springer, 513–523

Hachesu P R, Ahmadi M, Alizadeh S, Sadoughi F (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthcare Informatics Research, 19(2): 121–129

Harrell Jr F E, Lee K L, Mark D B (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4): 361–387

Holup A A, Hyer K, Meng H, Volicer L (2017). Profile of nursing home residents admitted directly from home. Journal of the American Medical Directors Association, 18(2): 131–137

Hoot N R, LeBlanc L J, Jones I, Levin S R, Zhou C, Gadd C S, Aronsky D (2008). Forecasting emergency department crowding: A discrete event simulation. Annals of Emergency Medicine, 52(2): 116–125

Incalzi R A, Gemma A, Capparella O, Terranova L, Porcedda P, Tresalti E, Carbonin P (1992). Predicting mortality and length of stay of geriatric patients in an acute care general hospital. Journal of Gerontology, 47(2): M35–M39

Kelly A, Conell-Price J, Covinsky K, Cenzer I S, Chang A, Boscardin W J, Smith A K (2010). Length of stay for older adults residing in nursing homes at the end of life. Journal of the American Geriatrics Society, 58(9): 1701–1706

Kramer A A, Zimmerman J E (2010). A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. BMC Medical Informatics and Decision Making, 10(1): 27

McGuire L C, Ford E S, Okoro C A (2007). Natural disasters and older US adults with disabilities: Implications for evacuation. Disasters, 31(1): 49–56

Moon M, Gage B, Evans A (1997). An examination of key Medicare provisions in the Balanced Budget Act of 1997. New York: The Commonwealth Fund

Murad Y (2011). Skilled nursing facilities and post-acute care. Journal of Gerontology & Geriatric Research, 1(101): 1–4

National Investment Center for Seniors Housing & Care (NIC) (2018). Skilled Nursing Data Report: Key Occupancy & Revenue Trends. 4Q2017

New P W, Stockman K, Cameron P A, Olver J H, Stoelwinder J U (2015). Computer simulation of improvements in hospital length of stay for rehabilitation patients. Journal of Rehabilitation Medicine, 47(5): 403–411

Pendharkar P C, Khurana H (2014). Machine learning techniques for predicting hospital length of stay in Pennsylvania federal and specialty hospitals. International Journal of Computer Science & Applications, 11(3): 45–56

Taboada M, Cabrera E, Iglesias M L, Epelde F, Luque E (2011). An agent-based decision support system for hospitals emergency departments. Procedia Computer Science, 4: 1870–1879

Turgeman L, May J H, Sciulli R (2017). Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. Expert Systems with Applications, 78: 376–385

Wang J, Li J, Tussey K, Ross K (2012). Reducing length of stay in emergency department: A simulation study at a community hospital. IEEE Transactions on Systems, Man, and Cybernetics: Part A, Systems and Humans, 42(6): 1314–1322

Xie H, Chaussalet T J, Millard P H (2005). A continuous time Markov model for the length of stay of elderly people in institutional long-term care. Journal of the Royal Statistical Society, Series A (Statistics in Society), 168(1): 51–61

Zhang X, Barnes S, Golden B, Myers M, Smith P (2019). Lognormal-based mixture models for robust fitting of hospital length of stay distributions. Operations Research for Health Care, 22: 100184