Journal of the American Medical Informatics Association, 2023, **30(11)**, 1801–1810
https://doi.org/10.1093/jamia/ocad101
Advance access publication 20 June 2023
Research and Applications

Research and Applications

Uncovering hidden trends: identifying time trajectories in risk factors documented in clinical notes and predicting hospitalizations and emergency department visits during home health care

Jiyoun Song [1,*], Se Hee Min [1], Sena Chae [2], Kathryn H. Bowles [3,4], Margaret V. McDonald [4], Mollie Hobensack [1], Yolanda Barrón [4], Sridevi Sridharan [4], Anahita Davoudi [4], Sungho Oh [3], Lauren Evans [4], and Maxim Topaz [1,4,5]

[1] Columbia University School of Nursing, New York City, New York, USA
[2] College of Nursing, University of Iowa, Iowa City, Iowa, USA
[3] Department of Biobehavioral Health Sciences, University of Pennsylvania School of Nursing, Philadelphia, Pennsylvania, USA
[4] Center for Home Care Policy & Research, VNS Health, New York, New York, USA
[5] Data Science Institute, Columbia University, New York City, New York, USA

*Corresponding Author: Jiyoun Song, PhD, AGACNP-BC, RN, Columbia University School of Nursing, 560 West 168th Street, New York, NY 10032, USA; js4753@cumc.columbia.edu

ABSTRACT

Objective: This study aimed to identify temporal risk factor patterns documented in home health care (HHC) clinical notes and examine their association with hospitalizations or emergency department (ED) visits.

Materials and Methods: Data for 73 350 episodes of care from one large HHC organization were analyzed using dynamic time warping and hierarchical clustering analysis to identify the temporal patterns of risk factors documented in clinical notes. The Omaha System nursing terminology represented risk factors. First, clinical characteristics were compared between clusters. Next, multivariate logistic regression was used to examine the association between clusters and risk for hospitalizations or ED visits. Omaha System domains corresponding to risk factors were analyzed and described in each cluster.

Results: Six temporal clusters emerged, showing different patterns in how risk factors were documented over time. Patients with a steep increase in documented risk factors over time had a 3 times higher likelihood of hospitalization or ED visit than patients with no documented risk factors. Most risk factors belonged to the physiological domain, and only a few were in the environmental domain.

Discussion: An analysis of risk factor trajectories reflects a patient's evolving health status during a HHC episode. Using standardized nursing terminology, this study provided new insights into the complex temporal dynamics of HHC, which may lead to improved patient outcomes through better treatment and management plans.

Conclusion: Incorporating temporal patterns in documented risk factors and their clusters into early warning systems may activate interventions to prevent hospitalizations or ED visits in HHC.

Key words: home health care, dynamic time warping, natural language processing, risk assessment, clinical deterioration, nursing informatics

INTRODUCTION

Home health care (HHC) offers personalized healthcare, such as nursing and social support services, to patients in their homes.[1] In the United States, HHC is one of the fastest-growing healthcare sectors due to a rapidly aging population and the need to accommodate an individual's desire for alternatives to institutional care.[2,3] HHC aims to assist patients in recovering from illness or injury, managing chronic conditions, maintaining independence, and minimizing the need for acute care services.[4] Despite ongoing efforts to reduce negative outcomes, over 20% of patients experienced hospitalizations and emergency department (ED) visits during HHC services.[5] Up to 40% of these negative outcomes are preventable with timely care[6–8]; early identification of HHC patients at risk can lead to closer surveillance and earlier interventions to prevent hospitalizations or ED visits.[9]

Previous studies have used standardized assessments and other structured data to identify risk factors associated with hospitalizations or ED visits in HHC.[10–14] However, a significant portion of HHC risk information is not solely captured in structured data or standard assessments but is also found in unstructured data such as clinical notes.[15] Prior research has shown that utilizing natural language processing (NLP) to analyze clinical notes has allowed the extraction of additional risk factors for hospitalizations or ED visits in HHC

Received: 6 March 2023. **Revised:** 4 May 2023. **Editorial Decision:** 25 May 2023. **Accepted:** 2 June 2023

compared with studying structured data alone.[16–18] Our team found that including information from HHC clinical notes through NLP can significantly improve the ability of machine learning algorithms to predict the patients at risk for hospitalizations or ED visits.[19]

To identify risk factors in clinical notes, it is imperative to use standardized terminology which can serve as a reliable guide to risk identification in clinical notes. Standardized terminology promotes consistent and clear identification and documentation of risk factors, which in turn facilitates evidence-based treatment and prevention strategies.[20,21] The Omaha System comprises 3 key components: the Problem Classification Scheme, the Intervention Scheme, and the Problem Rating Scale for Outcomes. These components enable healthcare professionals to precisely document and monitor patient progress. Furthermore, the Problem Classification Scheme—which guides the identification of risk factors throughout this study—includes categories for health status, environmental factors, and health-related behaviors, resulting in a comprehensive understanding of the patient's condition.[22,23] The Omaha System was developed to represent aspects unique to community-based care. In our previous NLP study, we used the Omaha System to identify risk factors for hospitalization or ED visits in HHC.[16,19,24,25]

In HHC, clinicians conduct multiple home visits to patients throughout episodes of care (ie, all services provided between the patient's admission and discharge from the HHC or within 60 days of the recertifying period), enabling longitudinal information to be collected. Longitudinal data, also known as time series data include multiple measurements, observations, and clinical notes documented at different intervals during the patient's care, providing information about their clinical condition over time and allowing for analysis of trends and patterns. However, one limitation of previous HHC risk identification studies is that NLP-extracted risk factors were aggregated and analyzed at the episode level.[13,17,19,26] As a result of this aggregation, changes in risk factors over time were not examined. Clustering techniques can aid in analyzing time trajectories of risk factors by grouping similar patterns together, enabling the identification of representative patterns in time series data.[27] Despite the potential benefits of clustering for understanding and analyzing time series data, this approach has not been applied in HHC.

To address limitations in prior research, the aims of this study were to: (1) identify the clusters of temporal risk patterns documented in HHC clinical notes, (2) examine the association between the clustering in temporal risk patterns and hospitalizations and ED visits, and (3) identify the interrelationships between the risk factor temporal domains.

MATERIALS AND METHODS

We conducted a retrospective cohort study utilizing 2 data sources: (1) structured data, consisting of the Outcome and Assessment Information Set (OASIS) and other assessment items extracted from the electronic health record (EHR) and (2) unstructured data (ie, clinical notes). This study was approved by the Institutional Review Boards of the participating institutions.

Study dataset and population

This study included patients who received HHC services between January 1, 2015 and December 31, 2017 from one of the largest not-for-profit HHC organizations in the Northeastern United States. During HHC episodes, patients involved in the study received several visits from HHC clinicians over a period of up to 60 days. To analyze trends in risk factors over time, we excluded patients who received only one HHC visit because data from a single visit are not enough to establish a trajectory. This led to the removal of 6.8% (5467) episodes for an effective study sample of 78 847 episodes. Since patients could have multiple home visits within the same episode, and multiple episodes of HHC over the study period, consequently, this study's sample included 551 681 home visits (mean = 7.5, standard deviation [SD]=4.6) during 73 350 HHC episodes conducted for 57 572 unique patients.

Structured datasets: OASIS and EHR

OASIS is a standardized Center for Medicare and Medicaid Services-mandated assessment tool for assessing patients in HHC at the beginning and end of their HHC episode. OASIS assesses over 100 patient characteristics, including sociodemographics, physiological conditions, comorbidities, medication and equipment management needs, neurocognitive and behavioral status, functional status (including activities of daily living [ADLs] and instrumental activities of daily living [IADLs]), and health service utilization.[28,29]

In addition, this study used several data elements extracted from administrative EHR data, including length of HHC episodes, HHC visit dates, and clinical note dates.

Unstructured dataset: clinical notes

For this patient group, approximately 2.3 million HHC clinical notes were extracted. Nurses documented most of these notes, while physical/occupational therapists and social workers generated the rest of the notes. In HHC, clinical notes are categorized into 2 types: (1) visit notes describe the patient's status and care provided during the HHC visit ($n = 1\ 029\ 535$), and (2) care coordination notes describe communication among healthcare clinicians and other administrative care-related activities ($n = 1\ 292\ 442$).

Utilizing the Omaha system to identify risk factors in clinical notes

In a previous study,[16] we developed a NLP system to extract 31 expert-defined hospitalization or ED visit risk factors from HHC clinical notes. Each risk factor was mapped to the standardized terminology, the Omaha System (see Supplementary Appendix S1 for a complete list of risk factors). The Omaha System has 4 domains (Environmental, Psychosocial, Physiological, and Health-related Behaviors) that are further divided into 42 problems (eg, "Income," "Abuse," "Circulation," "Medication regimen").[22] The NLP system performed well (an average F-score = 0.84) in identifying risk factors in HHC clinical notes. Further details about this work are published elsewhere.[16] In this current study, this NLP system was used to generate an indicator of whether or not the 31 risk factors (mapped to Omaha System problems) were documented during HHC visits, with a binary (present vs absent) response for each risk factor.

Applying dynamic time warping to align study's time points

We used the total number of risk factors identified in the clinical notes to create a longitudinal dataset associated with the day in the episode of care when the note was documented and the risk factor identified. This generated a time series data, consisting of a sequence of data points collected over time intervals, with varying lengths (ie, different frequencies and intervals), ranging from 1 to 60 days. Traditional analysis methods of time series (eg, Euclidean distance) assume similarity of time spans between observations (ie, equal number of HHC visits and equal length of time between the visits).[30] However, HHC visit patterns vary across patients; for example, 1 patient can have 7 HHC visits with an average of 4 days between the visits, while another patient might have only 3 HHC visits with an average of 3 days between the visits. Thus, to address nonlinear temporal patterns and accommodate time series of varying lengths, while also accounting for the lack of strict alignment between time points, we applied dynamic time warping methods.[31–33] Dynamic time warping is a method of aligning 2-time series sequences to calculate their pattern similarity. It accomplishes this by stretching or compressing one or both time series to minimize the differences between corresponding points. Figure 1 illustrates how dynamic time warping can accommodate for these variations in length in 2 different scenarios. We used the "*dtw (version 1.23-1)*" package in R.[34]

Generating temporal patterns of risk factors via unsupervised hierarchical cluster analysis

Data clustering is a data mining technique that involves grouping homogeneous data into uniform clusters.[35] In time series data analysis, clustering refers to identifying patterns or relationships within finite sequences of real numbers.[27] Hierarchical clustering groups data points into a layered structure, starting with each point as a separate cluster and then iteratively merging the closest pairs of clusters.[36] This approach can be used to identify patterns and select features to compress the dataset. In this study, dynamic time warping was used on the time series data to calculate the distance between 2 time series by allowing for local shifts and stretching of the time series data based on their similarity. Then hierarchical clustering used this distance to determine which time series are most similar to each other and should be grouped together. We identified temporal clusters of risk factors using a hierarchical clustering method using "*hclust (version 3.6.2)*" package in R.[37] Coinvestigators with expertise in HHC and machine learning (JS, SHM, KHB, and MT) visually examined the dendrogram (ie, a tree-like diagram that displays the hierarchical relationships between the clusters) to determine the optimal number of clusters that maximize clinical interpretability and usefulness.

Study outcome: hospitalizations and ED visits

We determined that a patient had been hospitalized if they had an OASIS assessment indicating the reason for assessment including transfers to inpatient facilities (OASIS item M0100) during the HHC episode, and had an ED visit if there was a record of receiving emergent care, captured in OASIS item M2300. Based on these 2 outcomes, we created a composite binary outcome variable indicating whether the patient had a hospitalization or an ED visit during their time in HHC.

Statistical analysis

First, the temporal patterns of risk factors documented in clinical notes over multiple HHC visits were identified using cluster analysis, then subsequent analysis was conducted at the HHC episode level. Differences in clinical characteristics between clusters were compared using analysis of variance (ANOVA) tests. A multivariate logistic regression analysis analyzed the association between the clusters of temporal risk patterns and hospitalizations and ED visits while adjusting for sociodemographic characteristics, comorbidities, and ADL/IADL function. We also utilized *UpSet* plots, a data visualization method that depicts intersecting set data to identify overlap between the risk factor encompassing Omaha System domains (Environmental, Psychosocial, Physiological, and

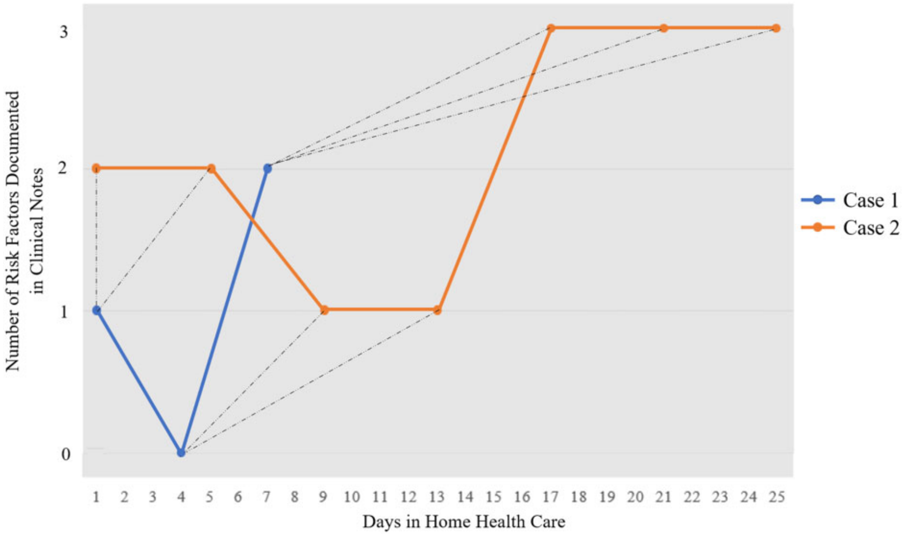

Figure 1. Illustrates an example of dynamic time warping applied to HHC episodes for 2 patients whose time series different lengths (ie, varying visit frequency and time intervals). Case 1 refers to one patient who had 3 HHC visits with an average interval of 3 days between visits, while Case 2 refers to another patient who had 7 HHC visits with an average interval of 4 days between visits. Dots indicate HHC visits. A dotted line in the diagram indicates the identification of a similar pattern (ie, similarity) between 2 HHC episodes of different lengths in the dynamic time wrapping method.

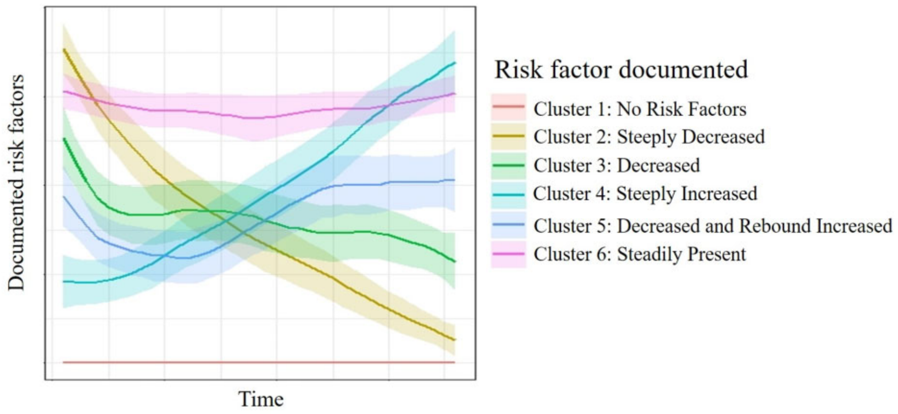

Figure 2. The temporal pattern of risk factors documented in clinical notes. The Y values were normalized to aid in visual comparison, and the color bands indicated the standard deviation.

Health-related Behaviors) and their clusters. A *P*-value <.05 (2-tailed) was considered statistically significant for all analyses. All analyses were implemented using R software version 4.2.2 (Foundation of Statistical Computing, Vienna).

RESULTS

During the study period, 8227/73 350 (11.2%) of HHC episodes resulted in a hospitalization or ED visit. The average patient's age was 79 years, and 64.3% were female. Among the sample, 63% of the patients were non-Hispanic Whites patients, followed by 17.2% of non-Hispanic Blacks patients, 13.4% of Hispanics patients, and 6.1% of others. For patients experiencing hospitalization or ED visits, the time gaps between the last note used for clustering and the hospital/ED visits varied. The minimum time gap was 0 days, showing that some hospitalizations or ED visits happened on the same day as the home visit. The median, mean, and maximum time gaps were 3, 11.1, and 59 days, respectively.

Clusters of temporal patterns in risk factors

Six clusters best represented the temporal patterns of risk factors documented in clinical notes. The clusters shown in Figure 2 can be summarized as follows: *Cluster 1* had no documented risk factor at any point in time (henceforth, labeled as "*No Risk Factors*"), *Cluster 2* had a steep decrease in the number of documented risk factors over time (henceforth, labeled *as* "*Steeply Decreased*"), *Cluster 3* had a moderate decrease in documented risk factors over time (henceforth, labeled as "*Decreased*"), *Cluster 4* had a steep increase in documented risk factors over time (henceforth, labeled as "*Steeply Increased*"), *Cluster 5* initially had a decrease in documented risk factors but risk factors number rebounded and increased over time (henceforth, labeled as "*Decreased and Rebound Increased*"), and *Cluster 6* had documented risk factors consistently present over time (henceforth, labeled as "*Steadily Present*").

A comparison of cohort characteristics by clusters

The cohort's clinical characteristics by clusters are presented in Table 1. *Cluster 3* "*Decreased*" had the highest mean age (79.3 years), while *Cluster 6* "*Steadily Present*" had the lowest mean age (77.4 years). Non-Hispanic White patients were predominant in *Cluster 3* "*Decreased*" (64.7%), while all other ethnicities were more predominant in *Cluster 1* "*No Risk Factors*." The proportion of patients who lived alone was highest in *Cluster 6* "*Steadily Present*" (40.7%) and lowest in *Cluster 1* "*No Risk Factors*" (35.0%). *Cluster 5* "*Decreased and Rebound Increased*" had a highest number of comorbidities and their severity, whereas *Cluster 3* "*Decreased*" had the highest level of ADL/IADL function.

Association between clusters and risk for hospitalization or ED visit

Table 2 shows the results of multivariate logistic regression analysis that examines the association of the clusters with the risk of hospitalization or visits to the ED. After adjusting for sociodemographic factors, chronic diseases, and ADL/IADL function, the odds of experiencing hospitalizations or ED visits were significantly higher for all clusters compared with *Cluster 1* "*No Risk Factors*" (odds ratios [OR] ranged 1.27–2.95). In further analysis, when compared with patients in *Cluster 2* "*Steeply Decreased*," patients in *Cluster 4* "*Steeply Increased*" had a 2.3 times greater likelihood of experiencing hospitalizations and ED visits (OR, 2.32 [95% CI, 2.15–2.51]) (all *P*-value <.001).

Relationships between the Omaha system domains of risk factors and clusters

Table 1 shows the distribution in clusters based on their Omaha System domains. Overall, the Physiological domain was the most common across all clusters, ranging from 88.6% to 95.7% of HHC episodes. Conversely, the Environmental domain was the least prevalent among all clusters, with prevalence ranging from 9.3% to 18.7% HHC episodes. The domains of Psychosocial, Physiological, and Health-related Behaviors follow a similar pattern, with the highest prevalence in *Cluster 3* "*Decreased*," and the lowest prevalence in *Cluster 4* "*Steeply Increased*." In contrast, the environmental domain had the highest prevalence in *Cluster 4* "*Steeply Increased*" and the lowest prevalence in *Cluster 2* "*Steeply Decreased*."

Figure 3 shows the *UpSet* plots that display clusters, each containing a combination of different Omaha System domains, along the top 7 most frequent combinations of risk factor domains. In most clusters, we selected the top 7 because they include at least 1% of the population by cluster. The following 4 top-ranking domains are consistent in most clusters: Physiological, Physiological with Psychosocial domain,

Table 1. Clinical characteristics and Omaha System risk factors in the cohort by cluster

Temporal pattern	Cluster 1: no risk factors	Cluster 2: steeply decreased	Cluster 3: decreased	Cluster 4: steeply increased	Cluster 5: decreased and rebound increased	Cluster 6: steadily present
Frequency [*n*, (%)]	5413 (7.38%)	35 193 (48%)	14 537 (19.8%)	5608 (7.65%)	7747 (10.56%)	4852 (6.61%)
Hospitalizations and ED visits [*n*, (%)]	394 (7.28%)	3033 (8.62%)	1854 (12.8%)	1021 (18.2%)	1258 (16.2%)	667 (13.7%)
1. Socio-demographic factors						
Age, years [mean, (SD)]	79.0 (12.0)	79.2 (11.5)	79.3 (11.5)	78.1 (12.1)	79.0 (11.7)	77.4 (12.0)
Female gender	3398 (62.8%)	22 821 (64.8%)	9481 (65.2%)	3498 (62.4%)	4985 (64.3%)	2976 (61.3%)
Ethnicity [*n*, (%)]						
Non-Hispanic White patients	3199 (59.1%)	22 414 (63.7%)	9406 (64.7%)	3526 (62.9%)	4896 (63.2%)	3011 (62.1%)
Non-Hispanic Black patients	1092 (20.2%)	5842 (16.6%)	2428 (16.7%)	999 (17.8%)	1357 (17.5%)	888 (18.3%)
Hispanic patients	746 (13.8%)	4687 (13.3%)	1902 (13.1%)	754 (13.5%)	1049 (13.5%)	659 (13.6%)
Other	376 (6.95%)	2250 (6.4%)	801 (5.51%)	329 (5.87%)	445 (5.74%)	294 (6.1%)
Type of insurance [*n*, (%)]						
Dual eligibility	332 (6.13%)	2081 (5.91%)	924 (6.36%)	325 (5.8%)	497 (6.42%)	310 (6.39%)
Medicare/Medicaid Fee-for-service only	4957 (91.6%)	32 228 (91.6%)	13 266 (91.3%)	5111 (91.14%)	7071 (91.27%)	4388 (90.44%)
Any managed care only	25 (0.5%)	192 (0.55%)	78 (0.54%)	37 (0.66%)	42 (0.54%)	41 (0.85%)
Other (eg, private)	102 (1.88%)	692 (1.97%)	269 (1.85%)	135 (2.41%)	137 (1.77%)	113 (2.33%)
Living arrangements						
Patient lives alone	1897 (35.0%)	13 958 (39.7%)	5619 (38.7%)	2023 (36.1%)	2895 (37.4%)	1973 (40.7%)
2. Comorbidity [mean, (SD)]						
Number of Comorbidities[a]	1.7 (2.5)	1.58 (2.5)	1.75 (2.6)	1.78 (2.56)	1.84 (2.6)	1.52 (2.5)
Severity of Comorbidities[b]	3.29 (5.12)	2.9 (4.9)	3.3 (5.1)	3.4 (5.1)	3.5 (5.3)	2.8 (4.8)
3. ADLs/IADLs [mean, (SD)]						
ADL Needed[c]	7.99 (1.59)	8.06 (1.45)	8.17 (1.37)	8.01 (1.55)	8.15 (1.38)	7.76 (1.77)
ADL Severity[d]	15.5 (7.06)	15.3 (6.40)	16.0 (6.67)	15.3 (6.73)	16.0 (6.76)	14.1 (6.34)
4. Risk factor documented in clinical notes (Omaha System problem) [mean, (SD)]						
Total number of risk factors	–	3.36 (1.93)	3.89 (2.13)	2.97 (2.04)	3.68 (2.14)	3.14 (1.67)
5. Domains of risk factors documented in clinical notes (Omaha System Domain) [*n*, (%)]						
Environmental Domain	–	3273 (9.30%)	1979 (13.6%)	1048 (18.7%)	1312 (16.9%)	694 (14.3%)
Psychosocial Domain	–	16 786 (47.7%)	7990 (55.0%)	2383 (42.5%)	4003 (51.7%)	2172 (44.8%)
Physiological Domain	–	33 262 (94.5%)	13 905 (95.7%)	4971 (88.6%)	7309 (94.3%)	4619 (95.2%)
Health-related behaviors Domain	–	7168 (20.4%)	3699 (25.4%)	1129 (20.1%)	1855 (23.9%)	992 (20.4%)

Note: Using the analysis of variance (ANOVA) test, clinical characteristics between clusters were compared, and variables with *P*-values <.05 were listed in the table.

[a] "Number of Comorbidities" which was defined as the summed binary ASPE Diagnosis indicators created using the OASIS Diagnosis items (eg, cancer, cardiac disease, stroke).[38]

[b] "Severity of Comorbidities" was calculated by totaling the response categories of the severity level in comorbidity items (total ranged from 0 to 24).

[c] "ADL Needed" which was defined as the summed binary ADL/IADL items (ranging from 0 to 9) derived from ADL items such as grooming, dressing upper and lower, bathing, toileting, transferring, ambulating, and eating, as well as IADL items such as meal preparation. Binary indicator 0 was given if response 0 was given (no issue); otherwise, 1 was given (moderate or significant issue).

[d] "ADLs Severity" was calculated by totaling the response categories of the dependency level in ADL/IADL items (total ranged from 0 to 38).

Physiological with Psychosocial and Health-related Behavioral domain, and Physiological with the Health-related Behavioral domain. A Health-related Behavioral domain emerged after a combination of Physiological and/or Psychosocial domains, which were consistent patterns across all clusters.

DISCUSSION

The study was the first to have identified the temporal patterns of risk factors based on the Omaha System documented in HHC clinical notes and examined the relationship between the temporal patterns and the risk for hospitalizations and ED visits. The analysis revealed 6 clusters that exhibited distinct patterns in how risk factors were documented over time within the Omaha System Problem Classification Scheme. Further, different patterns of time trajectory of risk factors showed varying effects on hospitalizations and ED visits.

Given the nature of HHC services, there are challenges integrating time-related factors into clinical risk assessment analyses due to the variations in visit frequency, length of stay, and different reasons for early discharges, such as recovery from the disease or deterioration that requires hospitalization or an

Table 2. A multivariate logistic regression analysis to examine the association of the clusters with the risk of hospitalization or visits to the ED

Predictors	Adjusted odds ratio (95% CI)
Cluster	
Cluster 1: No Risk Factors	Reference
Cluster 2: Steeply Decreased	1.27 (1.14–1.42)**
Cluster 3: Decreased	1.88 (1.68–2.11)**
Cluster 4: Steeply Increased	2.95 (2.60–3.34)**
Cluster 5: Decreased and Rebound Increased	2.47 (2.19–2.79)**
Cluster 6: Steadily Present	2.27 (1.98–2.59)**
Age	0.99 (0.99–1.00)
Gender	
Male	Reference
Female	0.84 (0.8–0.88)**
Ethnicity	
Non-Hispanic White patients	Reference
Non-Hispanic Black patients	1.41 (1.32–1.50)**
Hispanic patients	1.36 (1.27–1.45)**
Other (eg, Asian/Pacific islander)	0.86 (0.77–0.96)*
Type of insurance	
Dual eligibility	Reference
Medicare/Medicaid fee-for-service only	0.86 (0.78–0.94)**
Any managed care	0.89 (0.65–1.22)
Other (eg, private)	0.77 (0.63–0.93)*
Living arrangements	
Living with others	Reference
Living alone	1.08 (1.03–1.13)*
Comorbidity	
Number of Comorbidities[a]	1.07 (1.05–1.10)**
Severity of Comorbidities[b]	1.03 (1.02–1.04)**
ADLs/IADLs (activities of daily livings/instrumental activities of daily livings)	
ADL Needed[c]	0.98 (0.96–1)
ADL Severity[d]	1.04 (1.01–1.05)**

[a] "Number of Comorbidities" which was defined as the summed binary ASPE Diagnosis indicators created using the OASIS Diagnosis items (eg, cancer, cardiac disease, stroke).[38]

[b] "Severity of Comorbidities" was calculated by totaling the response categories of the severity level in comorbidity items (total ranged from 0 to 24).

[c] "ADL Needed" which was defined as the summed binary ADL/IADL items (ranging from 0 to 9) derived from ADL items such as grooming, dressing upper and lower, bathing, toileting, transferring, ambulating, and eating, as well as IADL items such as meal preparation. Binary indicator 0 was given if response 0 was given (no issue); otherwise, 1 was given (moderate or significant issue).

[d] "ADLs Severity" was calculated by totaling the response categories of the dependency level in ADL/IADL items (total ranged from 0 to 38).

* *P*-value <.05.

** *P*-value <.001.

ED visit. Because of these complexities, traditional methods (eg, latent class analysis) and general time series analysis are insufficient for analyzing HHC visit-level data because these methods assume equal number of data points (ie, visit frequency) and/or equal time intervals among all patients.[39–41] To address these challenges, this study implemented a unique and innovative dynamic time warping method that can handle nonlinear temporal patterns and time series with variations with minimal loss of time-related information.[32] We also used unsupervised clustering methods to identify 6 temporal trends patterns in risk factors documented in clinical notes over time. These patterns exhibited positive or negative directional trends and a trend line with varying degrees of the slope or fluctuations. As opposed to linear or logistic regression,[42,43] cluster analysis captured slope fluctuations, for example *Cluster 5*, which initially had a decrease in documented risk factors, but the number of risk factors rebounded. This innovative method, which employs temporal cluster features, offers a valuable strategy for analyzing not only HHC visit data but also measurement data in hospital EHR (eg, vital signs). This is particularly useful as measurements are often taken at varying times for each patient with unequal time intervals, which can make analysis challenging.

In this study, we examined the clinical characteristics of different clusters in the cohort. While *Cluster 3* "*Decreased*," which has a majority of non-Hispanic White patients, showed a decrease in documented risk factors over time (64.7%), all other ethnicities were more prevalent in *Cluster 1*"*No Risk Factors*." These results indicate that ethnicity may play a role in disparities in healthcare, as certain ethnic groups may have distinct health challenges or higher rates of specific health conditions,[44–47] which may be attributed to earlier or more aggressive strategies in addressing the risk factors specific to certain ethnicities. An alternative explanation might be that patients other than White patients are less likely to have their risk factors documented, despite reporting more problems during verbal communication with healthcare providers.[48] Our results also showed high prevalence of patients living alone in *Cluster 6* "*Steadily Present*" compared with *Cluster 1* "*No Risk Factors*." This suggests that caregiver support is important for maintaining symptom control and preventing negative outcomes.[49]

While there were statistically significant differences in other demographic and clinical factors extracted from structured data including OASIS, these differences between the clusters were relatively small; thus, their impact was not considered clinically relevant. The nature of the data may explain these relatively small differences: assessment data for OASIS start-of-care documentation are captured primarily during the first visit, while risk factors extracted from clinical notes are reported over time in multiple visits. Consequently, differences may exist between the characteristics of the cohort observed in the time trajectories of risk factors derived from clinical notes, which include the time aspect, and those assessed through the OASIS system. As a result, standard assessments at one time period cannot provide a complete representation of the changes that occur during a patient's time in HHC. Therefore, further research is required to explore ways to capture changes in patient characteristics, along with temporal clustering of structured and unstructured data.

Our study examined the association between clusters and risk for hospitalizations or ED visits. Notably, compared with *Cluster 1* "*No Risk Factors*," the risk of hospitalization or ED visits was higher in the other clusters, due to the documentation of risk factors regardless of the directional trends or degree of slope *Cluster 4* "*Steeply Increased*," showed the highest risk of hospitalization and ED visits (OR, 2.95), followed by *Cluster 5* "*Decreased and Rebound Increased*" (OR, 2.47). As seen in *Cluster 5* "*Decreased and Rebound Increased*," if new risk factors emerged or if patients experienced a cycle of illness, recovery, and then a recurrence of symptoms, the risk of hospitalization or ED visits may be higher compared with *Cluster 6* "*Steadily Present*." Furthermore, compared with *Cluster 2* "*Steeply Decreased*," *Cluster 4* "*Steeply Increased*" demonstrated higher odds of hospitalization or ED visits. Based on these findings, it is reasonable to suggest that time-related variables are significantly associated with risk of hospitalization or ED visits. However, the method used in this study (ie, a dynamic time-warping

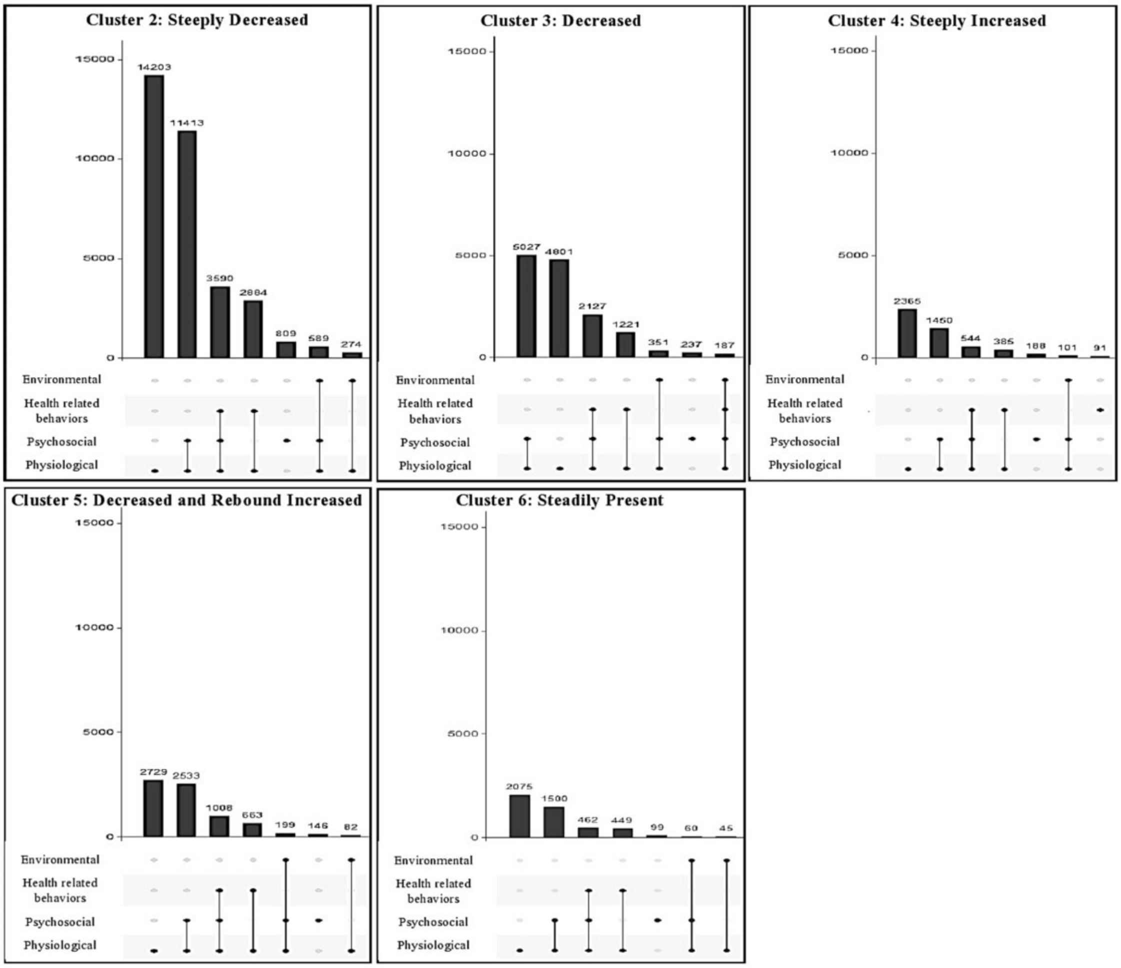

Figure 3. *UpSet* plots: The combination of risk factors based on the Omaha System Domain by clusters. Since *Cluster 1* is the group without any documented risk factors over time, the frequency of each subgroup by cluster was not presented.

algorithm to identify patterns based on similarities and an unsupervised hierarchical clustering) may not be able to distinguish between the differences in the number of risk factors, which could result in some patients with different numbers of risk factors being clustered together. Therefore, further research is needed to address this limitation and better understand the associations between risk factors and hospitalization or ED visits. These findings also suggest that the risk factors for hospitalizations or ED visits identified using the Omaha System provide a comprehensive and precise picture of the patient's situation, therefore, reliable indicators of risk prediction. By mapping the Omaha system through NLP, we could identify risk factors documented in clinical notes, which can save time and effort in extracting information from a large number of notes.

The Environmental domain encompassing social determinants of health was infrequently documented across all clusters, consistent with previous studies.[50,51] In contrast, the Physiological domain was the most frequently documented. While this could be because there were fewer identified risk factors in the Environmental domain compared with the Physiological domain, it is still important to identify social determinants of health, as these factors may be hidden and associated with negative health outcomes.[52,53] The Health-related Behaviors domain such as "Health care supervision" or "Medication regimen" was most frequent in *Cluster 2* "*Decreased*." Healthcare providers interact with patients during healthcare service and this interaction provides an opportunity for education related to self-management. Home visits over time can have a positive effect on self-care as evidenced by improvements in medication adherence, such as following the recommended dosage and schedule, which falls under the Health-related Behaviors domain. Future work might explore what nursing interventions occur with certain clusters to detect targeted teaching or care management activities.

In *Cluster 4* "*Steeply Increased*," the Physiological domain was more prevalent than other domain combinations. In other words, physical symptoms were more prominent and noticeable than other risk factors.[54] Physiological risk factors (eg, abnormal blood pressure, unable to breathe independently, and fever) in patients in HHC can change abruptly over time due to various reasons such as disease progression, medication changes, new medical conditions, or acute illness or injury.[10,55] It is important for HHC providers to monitor patients in *Cluster 4* "*Steeply Increased*," regularly and be vigilant for any changes in their condition that may require immediate medical attention. In *Cluster 2* "*Decreased*" and *Cluster 3* "*Decreased and Rebound Increased*," the Environmental domain ranked higher than other domains. This suggests that providers may prioritize having urgent symptoms controlled prior to addressing environmental factors addressed. This would be important to investigate in order to develop targeted interventions that address the underlying causes of the problem and improve health outcomes for individuals within the specific domain by examining the intercorrelations between each domain to gain a comprehensive understanding of how they affect each other. By identifying and addressing these specific risk factors, healthcare providers can better tailor their interventions to the unique needs and circumstances of their patients, ultimately leading to more effective and efficient healthcare delivery.

Lastly, the time difference between the last note used in clustering and the hospital/ED visit varied among patients, with some patients having a hospital/ED visit on the same day as the last note. This suggests that a nurse may identify a patient as high risk or in an emergency during home visits and refer them to 911, leading to an ED visit or hospitalization. On an average, there was a 3-day gap between the last note and the hospital/ED visit. It is essential to note that these time gaps represent the window of opportunity during which a clinical intervention to reduce negative outcomes could be implemented. It is crucial to monitor patients continuously and provide timely interventions to prevent hospitalization.

Clinical implications

Our study suggests that clinical notes contain important clues that clinicians gather from objective assessments and subjective symptoms reported by the patient. Temporal analysis of risk factors can accurately reflect a patient's overall health status. Analysis of time trends may provide new insights into the complex temporal dynamics of HHC services and could lead to the development of more effective interventions for improving patient outcomes. In this regard, we propose to identify a temporal clustering pattern within both structured data (eg, standardized assessments) and unstructured clinical notes, and our results could be used to develop temporal cluster-based risk prediction models. Such risk models can be integrated into early warning systems to identify HHC patients at risk of hospitalization and ED visits. Based on the findings of this study, we propose an early warning system that involves tracking the number of risk factors over time for each patient with at least one prior clinical note. Starting from the first home visit, trends of the 6 cluster groups could be compared to determine the most similar cluster group. If patients are in a higher-risk cluster group, their case should be flagged for early warning related to an increased risk for hospitalization or ED visits. Incorporating early warning systems into HHC clinical workflows could effectively alert nurses about at-risk patients, enabling them to intervene and improve patient outcomes by reducing risks.

Limitations

The investigation was carried out at a single HHC organization located in an urban area in the Northeastern United States, which limits its generalizability to other geographical locations and requires external validation. In addition, since this study was based on retrospective data, we cannot conclude causal relationships. Also, to identify the clusters of temporal risk patterns, this study only considered risk factors documented in the clinical notes and did not include any risk factors that might have been included in structured data. In addition, the dynamic time-warping algorithm to identify patterns based on similarities and an unsupervised hierarchical clustering may not be able to distinguish between the differences in the number of risk factors. This can result in a situation where patients with few and numerous risk factors are clustered together based on their temporal trends. Further research is needed to address this limitation and better understand the associations between risk factors and hospitalization or ED visits. In addition, further studies might explicitly add the number of risk factors at the beginning of HHC episode as a variable to the model, allowing the model to be adjusted for differences in the number of risk factors between patients. Lastly, the study only examined the average pattern in changes of the documented risk factors by clusters; this approach might overlook other, more detailed changes that occur over time at the individual level.

CONCLUSIONS

This study applied the terms of a standardized nursing language, the Omaha System, using NLP to identify 6 clusters of temporal patterns in documented risk factors that had varying degrees of association with hospitalizations or ED visits. By analyzing for different domains of risk factors, we identified the unique clinical characteristics of each cluster. Future studies can apply our findings to develop cluster-based early warning systems to prevent hospitalizations or ED visits.

FUNDING

This work was supported by the Agency for Healthcare Research and Quality (AHRQ) (R01 HS027742), "Building risk models for preventable hospitalizations and ED visits in homecare (Homecare-CONCERN)." The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. SHM is supported by the National Institute of Nursing Research (NINR) training grant Reducing Health Disparities through Informatics (RHeaDI) (T32NR007969) as a postdoctoral trainee. MH is supported by the NINR training grant Reducing Health Disparities through Informatics (RHeaDI) (T32NR007969) as a predoctoral trainee and the Jonas Scholarship. SO is supported by the NINR training grant Individualized Care for At Risk Older Adults (T32NR009356) as a postdoctoral trainee.

AUTHOR CONTRIBUTIONS

Study concept and design: JS, SHM, SC, KHB, MVM, and MT. Acquisition of data: MVM, YB, SS, and LE. Analysis and interpretation of data: JS, SHM, KHB, MVM, MH, YB, SO, and MT. Drafting of the manuscript: JS, SHM, SC, KHB, MVM, and MT. Critical revision of the manuscript of important intellectual content: all authors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

All authors report no conflicts of interest relevant to this article.

DATA AVAILABILITY

Data available on request due to privacy/ethical restrictions: the data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

1. The Medicare Payment Advisory Commission. Report to the congress—Medicare payment policy: home health care services. Secondary Report to the congress—Medicare payment policy: home health care services. 2019. http://www.medpac.gov/docs/default-source/reports/mar19_medpac_entirereport_sec.pdf. Accessed December 7, 2020.
2. Mitzner TL, Beer JM, McBride SE, *et al.* Older adults' needs for home health care and the potential for human factors interventions. *Proc Hum Factors Ergon Soc Annu Meet* 2009; 53 (1): 718–22.
3. Landers S, Madigan E, Leff B, *et al.* The future of home health care: a strategic framework for optimizing value. *Home Health Care Manag Pract* 2016; 28 (4): 262–78.
4. Centers for Medicare & Medicaid Services. Medicare benefit policy manual. Chapter 7, home health services. Secondary Medicare benefit policy manual. Chapter 7, home health services. 2017. https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/bp102c07.pdf. Accessed March 11, 2022.
5. Centers for Medicare and Medicaid Services. Home health compare. Secondary Home health compare. 2019. https://www.medicare.gov/homehealthcompare/search.html. Accessed November 19, 2020.
6. Solberg LI, Ohnsorg KA, Parker ED, *et al.* Potentially preventable hospital and emergency department events: lessons from a large innovation project. *Perm J* 2018; 22: 17–102.
7. Morganti KG, Bauhoff S, Blanchard JC, *et al.* The evolving role of emergency departments in the United States. *Rand Health Q* 2013; 3 (2): 3.
8. National Center for Health Statistics. Health, United States 2018 chartbook. Secondary Health, United States 2018 chartbook. 2018. https://www.cdc.gov/nchs/data/hus/hus18.pdf. Accessed March 31, 2022.
9. Zolnoori M, McDonald MV, Barrón Y, *et al.* Improving patient prioritization during hospital-homecare transition: protocol for a mixed methods study of a clinical decision support tool implementation. *JMIR Res Protoc* 2021; 10 (1): e20184.
10. Fortinsky RH, Madigan EA, Sheehan TJ, *et al.* Risk factors for hospitalization in a national sample of Medicare home health care patients. *J Appl Gerontol* 2014; 33 (4): 474–93.
11. Kang Y, Stoddard G, Horne B. Risk score for rehospitalization among home health care patients with heart failure. *J Card Fail* 2020; 26 (10): S135.
12. Lohman MC, Scherer EA, Whiteman KL, *et al.* Factors associated with accelerated hospitalization and re-hospitalization among Medicare home health patients. *J Gerontol A* 2018; 73 (9): 1280–6.
13. Shang J, Russell D, Dowding D, *et al.* A predictive risk model for infection-related hospitalization among home healthcare patients. *J Healthc Qual* 2020; 42 (3): 136–47.
14. Ma C, Shang J, Miner S, *et al.* The prevalence, reasons, and risk factors for hospital readmissions among home health care patients: a systematic review. *Home Health Care Manag Pract* 2018; 30 (2): 83–92.
15. Khan N, Yaqoob I, Hashem IAT, *et al.* Big data: survey, technologies, opportunities, and challenges. *Sci World J* 2014; 2014: 1.
16. Song J, Ojo M, Bowles KH, *et al.* Detecting language associated with home health care patient's risk for hospitalization and emergency department visit. *Nurs Res* 2022; 71 (4): 285–94.
17. Song J, Woo K, Shang J, *et al.* Predictive risk models for wound infection-related hospitalization or ED visits in home health care using machine-learning algorithms. *Adv Skin Wound Care* 2021; 34 (8): 1–12.
18. Topaz M, Woo K, Ryvicker M, *et al.* Home healthcare clinical notes predict patient hospitalization and emergency department visits. *Nurs Res* 2020; 69 (6): 448–54.
19. Song J, Hobensack M, Bowles KH, *et al.* Clinical notes: an untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care. *J Biomed Inform* 2022; 128: 104039.
20. Clancy TR, Delaney CW, Morrison B, *et al.* The benefits of standardized nursing languages in complex adaptive systems such as hospitals. *J Nurs Adm* 2006; 36 (9): 426–34.
21. Rutherford M. Standardized nursing language: what does it mean for nursing practice. *Online J Issues Nurs* 2008; 13 (1): 1–9.
22. Monsen AK. The Omaha system as an ontology and meta-model for nursing and healthcare in an era of big data. *Kontakt* 2018; 20 (2): e109–10.
23. Martin KS. *The Omaha System: A Key to Practice, Documentation, and Information Management*. Philadelphia, PA: Elsevier Saunders; 2005.
24. Hobensack M, Ojo M, Barrón Y, *et al.* Documentation of hospitalization risk factors in electronic health records (EHRs): a qualitative study with home healthcare clinicians. *J Am Med Inform Assoc* 2022; 29 (5): 805–12.
25. Topaz M, Golfenshtein N, Bowles KH. The Omaha system: a systematic review of the recent literature. *J Am Med Inform Assoc* 2014; 21 (1): 163–70.
26. Enguidanos S, Hoang T, Hillary KA, *et al.* Predictors of hospitalization among home health managed care patients. *Home Health Care Manag Pract* 2011; 23 (5): 363–72.
27. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering—a decade review. *Inform Syst* 2015; 53: 16–38.
28. Tullai-McGuinness S, Madigan EA, Fortinsky RH. Validity testing the outcomes and assessment information set (OASIS). *Home Health Care Serv Q* 2009; 28 (1): 45–57.
29. Shang J, Larson E, Liu J, *et al.* Infection in home health care: results from national outcome and assessment information set data. *Am J Infect Control* 2015; 43 (5): 454–9.
30. Elmore KL, Richman MB. Euclidean distance as a similarity metric for principal component analysis. *Mon Weather Rev* 2001; 129 (3): 540–9.
31. Berndt D, Clifford J. *Using Dynamic Time Warping to Find Patterns in Time Series*. Seattle, WA: KDD Workshop; 1994.
32. Müller M. Dynamic time warping. In: Müller M, ed. *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007: 69–84.
33. Stübinger J, Walter D. Using multi-dimensional dynamic time warping to identify time-varying lead-lag relationships. *Sensors* 2022; 22 (18): 6884.
34. Giorgino T. Computing and visualizing dynamic time warping alignments in R: the *dtw* package. *J Stat Softw* 2009; 31 (7): 1–24.
35. Dalmaijer ES, Nord CL, Astle DE. Statistical power for cluster analysis. *BMC Bioinformatics* 2022; 23 (1): 205.
36. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining Knowl Discov* 2012; 2 (1): 86–97.
37. R Documentation. Hierarchical clustering. Secondary Hierarchical clustering. 2013. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html. Accessed December 20, 2022.
38. The Centers for Medicare & Medicaid Services. Outcome and assessment information set OASIS—C2 guidance manual. Secondary Outcome and assessment information set OASIS—C2 guidance manual. 2018. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/Downloads/OASIS-C2-Guidance-Manual-Effective_1_1_18.pdf. Accessed October 30, 2020.
39. Neumann M, Wirtz M, Ernstmann N, *et al.* Identifying and predicting subgroups of information needs among cancer patients: an initial study using latent class analysis. *Support Care Cancer* 2011; 19 (8): 1197–209.

40. Singer JD, Willett JB. Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychol Bull* 1991; 110 (2): 268–90.
41. Gasparrini A. The case time series design. *Epidemiology* 2021; 32 (6): 829–37.
42. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons; 2013.
43. Weisberg S. *Applied Linear Regression*. Hoboken, NJ: John Wiley & Sons; 2005.
44. Davitt JK, Bourjolly J, Frasso R, *et al.* Understanding racial and ethnic disparities in home health care: practice and policy factors. *Innov Aging* 2017; 1 (Suppl_1): 956.
45. Groves PS, Bunch JL, Sabin JA. Nurse bias and nursing care disparities related to patient characteristics: a scoping review of the quantitative and qualitative evidence. *J Clin Nurs* 2021; 30 (23–24): 3385–97.
46. Bailey ZD, Krieger N, Agénor M, *et al.* Structural racism and health inequities in the USA: evidence and interventions. *Lancet* 2017; 389 (10077): 1453–63.
47. Narayan MC, Scafide KN. Systematic review of racial/ethnic outcome disparities in home health care. *J Transcult Nurs* 2017; 28 (6): 598–607.
48. Song J, Zolnoori M, Scharp D, *et al.* Do nurses document all discussions of patient problems and nursing interventions in the electronic health record? A pilot study in home healthcare. *JAMIA Open* 2022; 5 (2): ooac034.
49. Burgdorf JG, Arbaje AI, Stuart EA, *et al.* Unmet family caregiver training needs associated with acute care utilization during home health care. *J Am Geriatr Soc* 2021; 69 (7): 1887–95.
50. Hobensack M, Song J, Scharp D, *et al.* Machine learning applied to electronic health record data in home healthcare: a scoping review. *Int J Med Inform* 2023; 170: 104978.
51. Vest JR, Grannis SJ, Haut DP, *et al.* Using structured and unstructured data to identify patients' need for services that address the social determinants of health. *Int J Med Inform* 2017; 107: 101–6.
52. Patra BG, Sharma MM, Vekaria V, *et al.* Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc* 2021; 28 (12): 2716–27.
53. Hatef E, Weiner JP, Kharrazi H. A public health perspective on using electronic health records to address social determinants of health: the potential for a national system of local community health records in the United States. *Int J Med Inform* 2019; 124: 86–9.
54. Bender MS, Janson SL, Franck LS, *et al.* Theory of symptom management. In: Smith MJ, Liehr PR, eds. *Middle Range Theory for Nursing*. New York, NY: Springer; 2018: 147–78.
55. Ellenbecker CH, Samia L, Cushman MJ, *et al.* Patient safety and quality in home health care. In: Hughes RG, ed. *Patient Safety Quality: An Evidence-Based Handbook for Nurses*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2008.