Published in final edited form as:
J Adv Nurs. 2023 February ; 79(2): 593–604. doi:10.1111/jan.15498.

The identification of clusters of risk factors and their association with hospitalizations or emergency department visits in home health care

Jiyoun Song[1], **Sena Chae**[2], **Kathryn H. Bowles**[3,4], **Margaret V. McDonald**[4], **Yolanda Barrón**[4], **Kenrick Cato**[1,5], **Sarah Collins Rossetti**[1,6], **Mollie Hobensack**[1], **Sridevi Sridharan**[4], **Lauren Evans**[4], **Anahita Davoudi**[4], **Maxim Topaz**[1,4,7]

[1]Columbia University School of Nursing, New York City, New York, USA

[2]College of Nursing, The 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 City, USA

[5]Emergency Medicine, Columbia University Irving Medical Center, New York City, New York, USA

[6]Department of Biomedical Informatics, Columbia University, New York City, New York, USA

[7]Data Science Institute, Columbia University, New York City, New York, USA

Abstract

Aims: To identify clusters of risk factors in home health care and determine if the clusters are associated with hospitalizations or emergency department visits.

Design: A retrospective cohort study.

Methods: This study included 61,454 patients pertaining to 79,079 episodes receiving home health care between 2015 and 2017 from one of the largest home health care organizations in the United States. Potential risk factors were extracted from structured data and unstructured clinical notes analysed by natural language processing. A K-means cluster analysis was

---

Correspondence: Jiyoun Song, Columbia University School of Nursing, 560 West 168th Street, New York, NY, 10032, USA. js4753@cumc.columbia.edu.

AUTHOR CONTRIBUTIONS
All authors have agreed on the final version and meet at least one of the following criteria (recommended by the ICMJE): (1) substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; (2) drafting the article or revising it critically for important intellectual content. The specific contribution is as follows: Study concept and design: J Song, S Chae, K Bowles, M Margaret, Y Barron, S Rossetti, M Topaz; Acquisition of data: J Song, S Chae, M Hobensack, S Sridevi, L Evans, A Davoudi; Analysis and interpretation of data: J Song, S Chae, K Bowles, M Margaret, Y Barron, K Cato, S Rossetti, M Topaz; Drafting of the manuscript: J Song, S Chae, K Bowles, M Margaret, Y Barron, M Topaz; Critical revision of the manuscript of important intellectual content: All authors.

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

ETHICS STATEMENT
This study was approved by the VNS Health Institutional Review Boards (IRB# I20–003).

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

conducted. Kaplan–Meier analysis was conducted to identify the association between clusters and hospitalizations or emergency department visits during home health care.

Results: A total of 11.6% of home health episodes resulted in hospitalizations or emergency department visits. Risk factors formed three clusters. *Cluster 1* is characterized by a combination of risk factors related to “impaired physical comfort with pain,” defined as situations where patients may experience increased pain. *Cluster 2* is characterized by “high comorbidity burden” defined as multiple comorbidities or other risks for hospitalization (e.g., prior falls). *Cluster 3* is characterized by “impaired cognitive/psychological and skin integrity” including dementia or skin ulcer. Compared to *Cluster 1*, the risk of hospitalizations or emergency department visits increased by 1.95 times for *Cluster 2* and by 2.12 times for *Cluster 3* (all $p < .001$).

Conclusion: Risk factors were clustered into three types describing distinct characteristics for hospitalizations or emergency department visits. Different combinations of risk factors affected the likelihood of these negative outcomes.

Impact: Cluster-based risk prediction models could be integrated into early warning systems to identify patients at risk for hospitalizations or emergency department visits leading to more timely, patient-centred care, ultimately preventing these events.

Patient or Public Contribution: There was no involvement of patients in developing the research question, determining the outcome measures, or implementing the study.

Keywords

clinical deterioration; cluster analysis; home health care; natural language processing; nursing informatics; Omaha system; risk assessment

1 | INTRODUCTION/BACKGROUND

Home health care (HHC) includes skilled nursing care, occupational and physical therapy, social work service, and personal care assistance. A patient is eligible to receive care at home based on an assessment of the patient’s condition by a healthcare provider (e.g., physicians). Intermittent services are typically provided in person to patients at home with the goal of promoting recovery from illness and the prevention of deterioration (The Medicare Payment Advisory Commission, 2019). Over the past decades, the need for HHC has grown substantially in the United States (U.S.) and internationally, and the demand will likely continue to grow with an ageing population and longer life expectancies (Landers et al., 2016). The current trends of shorter hospital stays have contributed to increased clinical complexity of patients admitted to HHC (Burke et al., 2015). In the U.S., HHC services are usually provided for an “episode” which is a period up to 60 days paid for by the Centers for Medicaid and Medicare Services (Centers for Medicare & Medicaid Services, 2017) (see Footnote[1]). The patient may recertify for the continuation of the HHC during the comprehensive reassessment, or the patient may be discharged from the HHC upon reaching the clinical goals of their plan of care (The Medicare Payment Advisory Commission, 2019).

[1]In 2020, the length of CMS-reimbursable HHC episode was reduced from 60 days to 30 days (Centers for Medicare and Medicaid Services, 2019). Data for this study were collected between 1/1/2015 and 12/31/17, hence we used an episode length of 60 days.

Although continuous efforts are made to reduce negative outcomes in HHC, such as the utilization of acute care services (i.e., hospitalizations or emergency department (ED) visits), on average more than 20% of HHC patients are admitted to the hospital or visit an ED within the first 60 days after beginning HHC services (Centers for Medicare and Medicaid Services, 2019). Up to 40% of these events are preventable with timely care (National Center for Health Statistics, 2018; O'Connor et al., 2014; Solberg et al., 2018), so it is imperative to identify the patients at risk early so that healthcare providers are able to intervene (Zolnoori et al., 2021). Awareness of a patient's risk status allows healthcare providers to better monitor patients for worsening symptoms and provide early interventions when needed. These early interventions could include more frequent HHC visits, acute care interventions at home (e.g., intravenous therapies, medication adjustments), or telemonitoring.

In previous studies, risk factors associated with hospitalizations or ED visits in HHC were examined using standardized assessments, such as the Outcome and Assessment Information Set (OASIS) (Fortinsky et al., 2014; Lohman et al., 2018; Ma et al., 2017). Risk factors for hospitalization included being male, being Black, having a history of previous hospitalizations or polypharmacy, depressive symptoms, greater functional disability, dyspnea severity, and more. However, these risk factors commonly co-exist and are therefore likely to be frequently seen together in a patient's medical history (Shi & Stevens, 2005). Several studies conducted among patients with heart failure, other cardiovascular diseases, and stroke have demonstrated higher risk of adverse clinical outcomes when two or more risk factors are present at the same time (Peters et al., 2018; Son & Won, 2018). To support early identification of patients at risk, determining the combination of risk factors can potentially be more effective than identifying risk factors in isolation.

Cluster analysis is a data mining technique that groups similar observations into a number of data groups (i.e., clusters) based on measured characteristics to identify representation in specific groups (Tan et al., 2013). In recent years, cluster analysis has been used to identify phenotypes that have similar combinations of clinical factors (Sharma, 2021). Various clustering algorithms such as connectivity-based clustering (e.g., hierarchical clustering analysis), centroid models (e.g., K-means), and distribution models (e.g., expectation–maximization algorithm) have been applied to group a variety of factors including similar symptoms, patients with similar experiences, or risk factors (Li et al., 2019; Streur et al., 2018). Clustering is a method for aggregating the data, making it clinically meaningful and useful for prediction purposes (Alonso-Betanzos & Bolón-Canedo, 2018). Such data aggregation methods can reveal hidden patterns in the data, thus improving risk prediction accuracy by revealing data structure and regularities (Dalmaijer et al., 2022; Huang et al., 2019). In essence, clustering methods help to identify the complex interplay between different patient-level risk factors affecting a certain outcome rather than examining the impact of individual risk factors. Additionally, such an approach would be able to provide a more comprehensive picture of the patient's condition, along with identifying patient cohorts who need tailored treatment. As part of their efforts to develop personalized symptom management therapies, the National Institute of Nursing Research has identified the importance of managing co-occurring symptoms, which can be detected through grouping or clustering symptoms together (National Institute of Nursing Research, 2019). Despite clustering being a promising method for identifying hidden combinations of risk factors, no previous studies in HHC have examined clusters of risk factors and their association with hospitalization or ED visits over time.

2 | THE STUDY

2.1 | Aims

To address this knowledge gap, the aims of this study were to: (1) identify clusters of risk factors in HHC using unsupervised and data-driven analysis, (2) investigate the association between clusters of risk factors and hospitalizations or ED visits within 60 days considering time-to-event for each cluster, and (3) examine the associations between clusters of risk factors and the timing of hospitalization or ED visits.

2.2 | Design

This retrospective observational cohort study used the data obtained from one of the largest non-profit HHC organizations in the Northeastern U.S.

2.3 | Sample/participants

This study sample included patients who received HHC services between 1 January 2015 and 31 December 2017. An HHC "episode" refers to all services provided between the patient's admission and discharge from the HHC or 60 days, whichever occurs first. This study included 79,079 HHC episodes pertaining to 61,454 unique patients, since patients could have multiple episodes during the study period.

2.4 | Data collection

Two major data sources were retrieved: structured data (i.e., Outcome and Assessment Information Set (OASIS) and other assessment items from the electronic health record (EHR)) and unstructured data (i.e., clinical notes).

2.4.1 | Structured datasets: OASIS and EHR—The Center for Medicare and Medicaid Services mandates OASIS as a standardized outcome and assessment tool for HHC. At the time of admission and at the end of an episode of HHC, an OASIS assessment must be completed for each patient. The OASIS assessment captures over 100 patient characteristics including socio-demographics, physiologic conditions, comorbidities, medication and equipment management, cognitive and behavioural status (e.g., Activities of Daily Living (ADL)/Instrumental Activities of Daily Living (IADL)), and utilization of health care during the HHC episode (Tullai-McGuinness et al., 2009). We used both OASIS-C1 released in 2015 and OASIS-C2 released in 2017.

A dataset from the institution's EHR included features beyond OASIS, such as socioeconomic factors, insurance, county of residence, information on comorbidities, admission and discharge dates, and medications.

2.4.2 | Unstructured dataset: Clinical notes—The study cohort had about 2.3 million clinical notes generated during their episodes of care. Clinical notes were primarily written by nurses, but physical and occupational therapists and social workers also contributed. Clinical notes included (1) visit notes detailing the patient's condition and treatment during the HHC visit (total $n = 1{,}029{,}535$), and (2) care coordination notes describing the exchange of information between healthcare clinicians and other administrative duties (total $n = 1{,}292{,}442$).

In a previous study, our team developed a natural language processing algorithm (NLP-an artificial intelligence field in which computers analyse, understand, and extract meaning from human language in a text form) to extract the risk factors for hospitalizations or ED visits from HHC clinical notes (Song, Ojo, et al., 2022). Details on our previous NLP development and validation are described elsewhere (Song, Ojo, et al., 2022). In essence, based on the Omaha System—a standardized nursing terminology commonly utilized in community health (Martin, 2005)—a subset of 31 Omaha System problems, including "Circulation," "Respiration," "Healthcare supervision," etc., were identified as risk factors for hospitalizations or ED visits in HHC. Then, using the Omaha System as a tag of risk factors, the NLP algorithm was applied on all the clinical notes to identify risk factors for hospitalizations or ED visits. The NLP algorithm achieved high-risk factor identification accuracy with precision of 0.95, recall of 0.78 and an F-score of 0.84 (a harmonic means between precision and recall). A summary of the development of NLP is described in Appendix S1, and the risk factors identified for hospitalizations or ED visits are listed in Appendix S2.

2.4.3 | Outcome: Utilization of acute care services (i.e., hospitalizations or ED visits)—Hospitalizations or ED visits were identified from OASIS item M0100: "reason for completing assessment at present" (i.e., transfer to an inpatient facility including patient discharged or not discharged) and M2301 "emergent care" (i.e., utilization of the hospital ED, including hospital admission or non-admission). Time to hospitalizations or ED visits was calculated as the number of days between the date of HHC admission and the date of hospitalization or ED visit. For patients who did not have the outcome during the episode of care (censored), we defined follow-up time as the number of days between HHC admission and discharge or 60 days, whichever occurred first. All analyses were conducted at the HHC episode level.

2.5 | Ethical considerations

The study was approved by the Institutional Review Board of the participating institution (IRB# I20–003). Since our study used retrospective anonymous data, a waiver of informed consent was obtained. De-identified data were analysed. The highest safety standards were been followed with protection of study subject confidentiality as per national and international regulations for studies on human subjects included in the Declaration of Helsinki on Biomedical Research.

2.6 | Data analysis

2.6.1 | Variable selection—A full dataset was created including all available structured data elements (e.g., sociodemographic characteristics, comorbidities, functional status) and all variables derived by applying the NLP algorithm to clinical notes. We then applied the criteria described below to guide our selection of variables for inclusion in the clustering analysis. As done in our previous work (Song, Hobensack, et al., 2022; Song, Woo, et al., 2021; Song, Zolnoori, et al., 2021), variables with missing data over 20% were excluded, and the remaining variables with missing data were replaced with the median for continuous variables and the mode for categorical variables. To avoid linear dependency issues, we excluded redundant variables with strong correlations (Pearson correlation coefficient above 0.5 or below −0.5) and retained only variables with a higher frequency. Afterwards, we conducted a bivariate analysis (Student's *t*-test or Fisher exact test) between patients with hospitalizations or ED visits and those without to identify the variables that were statistically significant ($p < .05$). Lastly, to reduce the noise caused by small samples within each variable when conducting clustering tasks, those variables in which only less than 10% of the data indicate a presence of corresponding variables were excluded (Im et al., 2020). Our variable selection process resulted in 45 variables for use in the clustering analysis: 36 of these variables were derived from the OASIS assessment or the EHR data and 9 variables were derived from the application of NLP to the clinical notes.

2.6.2 | Cluster analysis—We sought to discover clusters of clinical characteristics using the K-means cluster analysis (Likas et al., 2003)—an unsupervised machine learning technique widely used in data mining, pattern recognition, and decision support. Clustering by K-means is the process of grouping N observations into groups of K. To classify observations into groups, the degree of similarity/dissimilarity or distance between observation pairs was calculated using the Euclidean distance (Singh et al., 2013). The centre of the cluster (i.e., centroid) represents the average of all observations assigned to the cluster. Then, each object is assigned to its closest centroid based on the distance between the observations and the centroid. Ultimately, each observation belongs to the group with the closest cluster mean or centroid (Likas et al., 2003). An elbow method was used to determine the optimal number of clusters, in which the sum of squares for each K within a cluster is plotted over a curve, and the point where the curve appears sharpest indicates the optimal number of clusters (Syakur et al., 2018).

2.6.3 | Statistical analysis—Following K-means cluster analysis, differences in clinical characteristics between patients with hospitalizations or ED visits versus patients without those outcomes were compared using analysis of variance (ANOVA) test. Then, a Kaplan–Meier analysis was used to estimate the survival rate from hospitalizations or ED visits in HHC within 60 days between clusters (i.e., a combination of risk factors), and a log-rank test was used to compare differences between clusters. The hazard ratio (HR) and 95% confidence interval (CI) were presented to estimate the association of clusters on time-to-event outcomes (i.e., hospitalizations or ED visits). Lastly, we performed a post-hoc ANOVA on episodes that included hospitalizations or ED visits to examine the association between the timing of hospitalizations or ED visits and clusters. For all analyses, a *p*-value <.05 (two-tailed) was considered statistically significant. All analyses were implemented using R software version 4.1.0 (Foundation of Statistical Computing, Vienna).

2.7 | Validity, reliability and rigour

Through K-means cluster analysis as an unsupervised machine learning technique, large numbers of observations can be categorized into groups with similar properties. The credibility of K-means cluster analysis has already been recognized and used (Likas et al., 2003). A discussion was held within the research group concerning the data analysis to guarantee methodological coherence, adequate sampling, and responsiveness. The lead author of this paper conducted the analyses independently, but the other authors critically reviewed the findings, which led to a consensus on the themes and labelling of clusters based on their characteristics.

3 | RESULTS

During the study period, 11.6% (9182/79,079) of HHC episodes resulted in utilization of acute care services (i.e., hospitalizations or ED visits).

3.1 | Cohort demographics and clinical characteristics

The average patient age was 78.8 years, and 64% of patients were female. Hypertension, diabetes, and arthritis were the most common diagnoses (65%, 30%, and 24%, respectively). Approximately 24% of patients experienced multiple hospitalizations within the 6 months before receiving HHC services; a history of prior hospitalization was more common in patients who experienced hospitalization or ED utilization during their HHC episode compared to those who did not experience these outcomes (40% vs. 22%, respectively). The most frequently documented risk factors in clinical notes were “Pain,” followed by “Neuromusculoskeletal function,” “Circulation,” and “Mental health” issues (48%, 46%, 35%, and 31%, respectively). The following problems were more frequently documented in the clinical notes of patients with hospitalizations or ED visits than those without: “Circulation (42 vs. 33%),” “Cognition (20 vs. 15%),” “Mental health (39 vs 30%),” “Pain (50 vs. 46%),” and “Skin (28% vs. 18%).” Additional details are presented in Table 1.

3.2 | Cluster analysis of risk factors

Using elbow methods, the optimal number of clusters was determined to be three.

Table 2 presents the distinct clinical characteristics of clusters associated with each of the clusters. *Cluster 1* is characterized by a combination of risk factors for “impaired physical comfort with pain”. For this cluster, there was predominant documentation of pain in clinical notes and in OASIS, as well as clinical situations that could potentially increase pain, such as the history of arthritis or surgical wounds. *Cluster 2* is labelled “high comorbidity burden” defined as multiple comorbidities (such as diabetes or cardiovascular disease) and multiple other risks for hospitalization (such as prior falls and multiple prior hospitalizations) compiled from OASIS, as well as circulatory or respiratory problems documented in clinical notes. *Cluster 3* is characterized by “impaired cognitive/psychological and skin integrity”. Patients in this cluster had: (a) significant cognitive/psychological issues such as dementia, confusion, or anxiety noted in the OASIS, or mental health or cognition issues documented in clinical notes, and (b) significant skin issues, such as an open wound or skin ulcer, noted in the OASIS, or a skin condition documented in clinical notes.

3.3 | Association of the Clusters with risk for hospitalization or ED visit

Compared with the *Cluster 1* group of risk factors, those with *Cluster 2* and *Cluster 3* risk factors were at higher risk of being hospitalized or visiting the ED within 60 days of admission to HHC. The risk of hospitalizations or ED visits was 1.95 times higher for *Cluster 2* (hazard ratio (HR), 1.95 [95% CI, 1.86–2.04]) and 2.12 times higher for *Cluster 3* (HR, 2.12 [95% CI, 1.99–2.26]) compared with *Cluster 1* (all $p < .001$).

In the post-hoc analysis that included only patients with hospitalizations or ED visits, the time to event was 38 days (standard deviation [*SD*] = 18.1) in *Cluster 1*, 41.7 days (*SD* = 17.5) in *Cluster 2*, and 38.7 days (*SD* = 18.2) in *Cluster 3*. Thus, among patients who were hospitalized or used the ED, those with *Cluster 1* symptoms (impaired physical comfort with pain) had the shortest time to event, which was slightly shorter than *Cluster 3*, and significantly shorter than *Cluster 2* ($p < .001$).

4 | DISCUSSION

Our study is the first to our knowledge to evaluate the clusters of risk factors in HHC and their association with risk for hospitalizations or ED visits. Using data mining-based unsupervised cluster analysis, hidden patterns and combinations of risk factors were identified in a large sample of patients receiving HHC service. A heterogeneity of the combination of risk factors was observed, with distinct characteristics in each cluster: *Cluster 1*—Impaired physical comfort with pain; *Cluster 2*—High comorbidity burden; *Cluster 3*—Impaired cognitive/psychological and skin integrity.

Although the themes of clusters (i.e., pain, comorbidities, cognitive impairment and poor integumentary status) can be mapped to established risk factors for hospitalization and ED visits (Fortinsky et al., 2014; Lohman et al., 2018; Ma et al., 2017; Shang et al., 2020; Song, Woo, et al., 2021), our attempt to identify combinations of risk factors in HHC patients is novel. Our results also support previous studies showing that clinical characteristics should not be considered as isolated factors, since they tend to cluster together (Murphy et al., 2019). Characterizing these groups with the different combinations of risk factors can provide a basis for tailoring treatment for patients with these risk factors. Based on the results of this study, HHC health care providers should identify patients at risk of pain, comorbidities, cognitive impairment, and poor integumentary status early across HHC treatments to plan the most effective interventions and follow-up during HHC trajectories. For example, an early pain management strategy including postoperative pain control, mobilization with therapy, early referral for interdisciplinary pain management may help patients with risk factors in *Cluster 1* avoid hospitalization or ED visits (Wells et al., 2008). Patients in the *Cluster 2* group may benefit from assistance with planning self-management strategies to deal with the burden of chronic disease or from more frequent monitoring for medication adjustments and/or to ensure adherence to medication for chronic disease (Grady & Gough, 2014). Lastly, patients in the *Cluster 3* group might benefit from cognitive function stimulation or counselling strategies (Silva et al., 2021), or wound management (Karadağ & Çakar, 2022).

Notably, associations between clinical characteristics were identified within each cluster. For example, in *Cluster 3*, older age was clustered together with the risk of cognitive impairments (e.g., dementia) and the risk of having sensory impairments (e.g., difficulty seeing, hearing, and speaking) (Loughrey et al., 2018). In addition, elderly patients with cognitive decline are more prone to having pressure ulcers and requiring greater assistance with ADLs/IADLs due to their vulnerability to poor self-care and decreased mobility (Edwards et al., 2020; Jaul et al., 2018). This cluster also included a higher proportion of Hispanic patients; these patients are indeed more likely to be diagnosed with dementia compared to White patients (Chen & Zissimopoulos, 2018). Thus, cluster analysis could reveal hidden patterns by incorporating the clinical characteristics that are potentially associated with a cluster.

Our findings also showed that a certain combination of risk factors (i.e., clusters) was associated with the time to or incidence of hospitalization or ED visits. In a previous study in which individual risk factors for hospitalizations or ED visits in HHC were examined, chronic comorbidities (e.g., diabetes), mental illness, or psychological issues were not identified as statistically significant in a multivariate logistic regression analysis (Song, Woo, et al., 2021). In contrast, the current study showed that clusters of risk factors related to high comorbidities burden or cognitive/psychological or skin issues increases the risk of hospitalizations or ED visits. In addition, certain combinations of risk factors were associated with earlier hospitalizations or ED visits. Perhaps, acute pain, a demonstrable cognitive/psychological impairment, or integumentary issues demand immediate attention, therefore, patients with those conditions may have been hospitalized earlier than those with chronic conditions that are comparatively not as urgent unless there are exacerbations (Green et al., 2018). Further research is needed to determine whether trajectories of risk factors are associated with earlier hospitalizations or ED visits.

This study is also innovative because it leveraged various types of data streams, such as structured data and unstructured data (e.g., clinical notes) to perform cluster analysis and leveraged the problem and symptom terms within a standardized nursing terminology to facilitate the NLP (Martin, 2005). Our findings indicated that data retrieved from clinical notes and structured assessments have homogeneity in terms of content. For example, pain recorded in structured data was captured in the clinical notes as well. Thus, information extracted from these convergent data sources can be leveraged as valid indicators to determine the risk of hospitalization or ED visit, increasing the possibility of capturing the hidden combinations of risk factors and identifying patients' risk profiles. A comprehensive set of symptoms that were documented in the unstructured clinical note could be identified by using the Omaha System problem which included broad signs and symptoms (e.g., 'does not follow recommended dosage/schedule' under the problem of 'Medication regimen', or 'fails to obtain routine/preventive health care' under the problem of 'Health care supervision'). Structured data, on the other hand, has not been available for such data. From this perspective, along with highlighting the utilization of clinical notes in identifying risk factors for hospitalizations or ED visits, future efforts may include adding signs/symptoms that could be used to populate a structured symptom checklist in the HHC EHR to reduce the burden of narrative documentation.

4.1 | Future clinical implications

Considering these results, developing cluster-based risk prediction models may be feasible in HHC. These models could be incorporated into early warning systems for identifying HHC patients at risk for hospitalizations or ED visits. Ultimately, the integration of such early warning systems into HHC clinical workflows would alert nurses about patients at risk, enabling them to intervene to reduce risks and improve outcomes. Although it has been demonstrated that early warning systems are effective in improving clinical outcomes in hospital settings (Gerry et al., 2020), little is known about their effectiveness in HHC. Therefore, further research is needed to develop such early warning systems and evaluate their effectiveness in HHC settings to improve patient outcomes, such as reducing hospitalizations and ED visits.

4.2 | Limitations

There are several notable limitations to this study. First, this investigation was conducted at a single HHC organization located in an urban area in the northeastern U.S. This limits its generalizability to other geographic locations, which require external validation. Since data collected from 2015 to 2017 were utilized in the analysis, results should be replicated in more recent patient cohorts. Also, several clinical characteristics, even though they were associated with hospitalizations or ED visits, were not included in the cluster analysis because they were not selected in the initial variable selection stage due to their low prevalence; some information might have been lost as a result. Given we used unstructured clinical notes based on the English language, the current developed NLP approach is not available to the international nursing community, but structured data might be useful without language restrictions. Future work should also examine whether using structured data alone (e.g., OASIS) can produce similar clustering results. Moreover, the present study was based on retrospective data which limits our ability to infer causal relationships. Lastly, survival analysis in this study has a limitation in that information about hospitalization or ED visits was not available after discharge from HHC, leading us to underestimate these outcomes in some cases.

5 | CONCLUSIONS

This study identified three distinct clusters of risk factors associated with hospitalizations or ED visits. Our findings demonstrate the heterogeneity of the combination of risk factors and clearly show that every cluster had its own characteristics. The different combinations of risk factors showed different effects on the likelihood of hospitalizations or ED visits, and the timing of such visits. Our findings suggest that patients who experience ‘impaired cognitive/psychological and skin integrity,’ more frequently be hospitalized or visit the ED, have many unmet risk management needs, and may require the highest level of supportive care need and intervention during HHC. Future studies should explore the use of risk cluster-based early warning systems to prevent hospitalizations or ED visits in HHC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

FUNDING INFORMATION

This study was funded by Agency for Healthcare Research and Quality [AHRQ] (R01 HS027742), "Building risk models for preventable hospitalizations and emergency department 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. Ms Hobensack is supported by the National Institute for Nursing Research training grant Reducing Health Disparities through Informatics (RHeaDI) (T32NR007969) as a predoctoral trainee and the Jonas Scholarship.

Funding information

Agency for Healthcare Research and Quality, Grant/Award Number: R01 HS027742; National Institute of Nursing Research, Grant/Award Number: T32NR007969; Jonas Scholarship

DATA AVAILABILITY STATEMENT

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

Alonso-Betanzos A, & Bolón-Canedo V (2018). Big-data analysis, cluster analysis, and machine-learning approaches. In Kerkhof PLM & Miller VM (Eds.), Sex-specific analysis of cardiovascular function (pp. 607–626). Springer International Publishing. 10.1007/978-3-319-77932-4_37

Burke RE, Juarez-Colunga E, Levy C, Prochazka AV, Coleman EA, & Ginde AA (2015). Patient and hospitalization characteristics associated with increased postacute care facility discharges from US hospitals. Medical Care, 53(6), 492–500. 10.1097/mlr.0000000000000359 [PubMed: 25906015]

Centers for Medicare & Medicaid Services. (2017). Medicare benefit policy manual. Chapter 7, home health services. https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/bp102c07.pdf

Centers for Medicare and Medicaid Services. (2019). Home health compare. CMS IOM Publication 100–02. https://www.medicare.gov/homehealthcompare/search.html

Chen C, & Zissimopoulos JM (2018). Racial and ethnic differences in trends in dementia prevalence and risk factors in the United States. Alzheimer's & Dementia, 4, 510–520. 10.1016/j.trci.2018.08.009

Dalmaijer ES, Nord CL, & Astle DE (2022). Statistical power for cluster analysis. BMC Bioinformatics, 23(1), 205. 10.1186/s12859-022-04675-1 [PubMed: 35641905]

Edwards RD, Brenowitz WD, Portacolone E, Covinsky KE, Bindman A, Glymour MM, & Torres JM (2020). Difficulty and help with activities of daily living among older adults living alone with cognitive impairment. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 16(8), 1125–1133. 10.1002/alz.12102

Fortinsky RH, Madigan EA, Sheehan TJ, Tullai-McGuinness S, & Kleppinger A (2014). Risk factors for hospitalization in a national sample of medicare home health care patients. Journal of Applied Gerontology, 33(4), 474–493. 10.1177/0733464812454007 [PubMed: 24781967]

Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, & Collins GS (2020). Early warning scores for detecting deterioration in adult hospital patients: Systematic review and critical appraisal of methodology. BMJ, 369, m1501. 10.1136/bmj.m1501 [PubMed: 32434791]

Grady PA, & Gough LL (2014). Self-management: A comprehensive approach to management of chronic conditions. American Journal of Public Health, 104(8), e25–e31. 10.2105/AJPH.2014.302041 [PubMed: 24922170]

Green LA, Chang H-C, Markovitz AR, & Paustian ML (2018). The reduction in ed and hospital admissions in medical home practices is specific to primary care–sensitive chronic conditions. Health Services Research, 53(2), 1163–1179. 10.1111/1475-6773.12674 [PubMed: 28255992]

Huang L, Shea AL, Qian H, Masurkar A, Deng H, & Liu D (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. Journal of Biomedical Informatics, 99, 103291. 10.1016/j.jbi.2019.103291 [PubMed: 31560949]

Im S, Qaem MM, Moseley B, Sun X, & Zhou R (2020). Fast noise removal for k-means clustering. International Conference on Artificial Intelligence and Statistics.

Jaul E, Barron J, Rosenzweig JP, & Menczel J (2018). An overview of co-morbidities and the development of pressure ulcers among older adults. BMC Geriatrics, 18(1), 305. 10.1186/s12877-018-0997-7 [PubMed: 30537947]

Karadağ A, & Çakar V (2022). Evidence-based prevention and management of pressure injuries in home care: A scoping review. Advances in Skin & Wound Care, 35(3), 172–179. 10.1097/01.ASW.0000815484.50141.5d [PubMed: 35188484]

Landers S, Madigan E, Leff B, Rosati RJ, McCann BA, Hornbake R, MacMillan R, Jones K, Bowles K, Dowding D, Lee T, Moorhead T, Rodriguez S, & Breese E (2016). The future of home health care: A strategic framework for optimizing value. Home Health Care Management & Practice, 28(4), 262–278. 10.1177/1084822316666368 [PubMed: 27746670]

Li H, Ji M, Scott P, & Dunbar-Jacob JM (2019). The effect of symptom clusters on quality of life among patients with type 2 diabetes. The Diabetes Educator, 45(3), 287–294. 10.1177/0145721719837902 [PubMed: 30873908]

Likas A, Vlassis N, & Verbeek JJ (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461.

Lohman MC, Scherer EA, Whiteman KL, Greenberg RL, & Bruce ML (2018). Factors associated with accelerated hospitalization and re-hospitalization among medicare home health patients. The Journals of Gerontology: Series A, 73(9), 1280–1286. 10.1093/gerona/glw335

Loughrey DG, Kelly ME, Kelley GA, Brennan S, & Lawlor BA (2018). Association of age-related hearing loss with cognitive function, cognitive impairment, and dementia: A systematic review and meta-analysis. JAMA Otolaryngology. Head & Neck Surgery, 144(2), 115–126. 10.1001/jamaoto.2017.2513 [PubMed: 29222544]

Ma C, Shang J, Miner S, Lennox L, & Squires A (2017). The prevalence, reasons, and risk factors for hospital readmissions among home health care patients: A systematic review. Home Health Care Management & Practice, 30(2), 83–92. 10.1177/1084822317741622

Martin KS (2005). The Omaha system: A key to practice, documentation, and information management. Elsevier Saunders. https://books.google.com/books?id=89Rjdd-ulIwC

Mikolov T, Sutskever I, Chen K, Corrado G, & Dean J (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26. https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Murphy JJ, MacDonncha C, Murphy MH, Murphy N, Timperio A, Leech RM, & Woods CB (2019). Identification of health-related behavioural clusters and their association with demographic characteristics in Irish university students. BMC Public Health, 19(1), 121. 10.1186/s12889-019-6453-6 [PubMed: 30691428]

National Center for Health Statistics. (2018). Health, United States 2018 Chartbook. https://www.cdc.gov/nchs/data/hus/hus18.pdf

National Institute of Nursing Research. (2019). Symptom science center: A resource for precision health. https://www.ninr.nih.gov/newsandinformation/events/sscevent

O'Connor M, Hanlon A, & Bowles KH (2014). Impact of frontloading of skilled nursing visits on the incidence of 30-day hospital readmission. Geriatric Nursing, 35(2 Suppl), S37–S44. 10.1016/j.gerinurse.2014.02.018 [PubMed: 24702719]

Peters SAE, Wang X, Lam T-H, Kim HC, Ho S, Ninomiya T, Knuiman M, Vaartjes I, Bots ML, Woodward M, & Asia Pacific Cohort Studies Collaboration. (2018). Clustering of risk factors and the risk of incident cardiovascular disease in Asian and Caucasian populations: Results from the Asia Pacific cohort Studies collaboration. BMJ Open, 8(3), e019335. 10.1136/bmjopen-2017-019335

Shang J, Russell D, Dowding D, McDonald MV, Murtaugh C, Liu J, Larson EL, Sridharan S, & Brickner C (2020). A predictive risk model for infection-related hospitalization among home healthcare patients. Journal for Healthcare Quality, 42(3), 136–147. 10.1097/jhq.0000000000000214 [PubMed: 32371832]

Sharma S (2021). The bioinformatics: Detailed review of various applications of cluster analysis. Global Journal on Application of Data Science, 5(1–2021), 1–14

Shi L, & Stevens GD (2005). Vulnerability and unmet health care needs. Journal of General Internal Medicine, 20(2), 148–154. 10.1111/j.1525-1497.2005.40136.x [PubMed: 15836548]

Silva R, Bobrowicz-Campos E, Santos-Costa P, Cruz AR, & Apóstolo J (2021). A home-based individual cognitive stimulation program for older adults with cognitive impairment: A randomized controlled trial. Frontiers in Psychology, 12, 741955. 10.3389/fpsyg.2021.741955 [PubMed: 34880809]

Singh A, Yadav A, & Rana A (2013). K-means with three different distance metrics. International Journal of Computer Applications, 67(10), 13–17.

Solberg LI, Ohnsorg KA, Parker ED, Ferguson R, Magnan S, Whitebird RR, Neely C, Brandenfels E, Williams MD, Dreskin M, Hinnenkamp T, & Ziegenfuss JY (2018). Potentially preventable hospital and emergency department events: Lessons from a large innovation project. The Permanente Journal, 22, 17–102. 10.7812/TPP/17-102

Son YJ, & Won MH (2018). Symptom clusters and their impacts on hospital readmission in patients with heart failure: A cross-sectional study. Research and Theory for Nursing Practice, 32(3), 311–327. 10.1891/1541-6577.32.3.311 [PubMed: 30567841]

Song J, Hobensack M, Bowles KH, McDonald MV, Cato K, Rossetti S, Chae S, Kennedy E, Barrón Y, Sridharan S, & Topaz M (2022). Clinical notes: An untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care. Journal of Biomedical Informatics, 128, 104039. 10.1016/j.jbi.2022.104039 [PubMed: 35231649]

Song J, Ojo M, Bowles KH, McDonald MV, Cato K, Rossetti S, Adams V, Chae S, Hobensack M, Kennedy E, Tark A, Kang M-J, Woo K, Barrón Y, Sridharan S, & Topaz M (2022). Detecting language associated with home health care patient's risk for hospitalization and emergency department visit. Nursing Research, 71, 285–294. 10.1097/NNR.0000000000000586 [PubMed: 35171126]

Song J, Woo K, Shang J, Ojo M, & Topaz M (2021). Predictive risk models for wound infection-related hospitalization or ed visits in home health care using machine-learning algorithms. Advances in Skin & Wound Care, 34(8), 1–12. 10.1097/01.Asw.0000755928.30524.22

Song J, Zolnoori M, McDonald MV, Barrón Y, Cato K, Sockolow P, Sridharan S, Onorato N, Bowles KH, & Topaz M (2021). Factors associated with timing of the start-of-care nursing visits in home health care. Journal of the American Medical Directors Association, 22, 2358–2365.e3. 10.1016/j.jamda.2021.03.005 [PubMed: 33844990]

Streur M, Ratcliffe SJ, Callans D, Shoemaker MB, & Riegel B (2018). Atrial fibrillation symptom clusters and associated clinical characteristics and outcomes: A cross-sectional secondary data analysis. European Journal of Cardiovascular Nursing, 17(8), 707–716. 10.1177/1474515118778445 [PubMed: 29786450]

Syakur M, Khotimah B, Rochman E, & Satoto BD (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering, 336(1), 012017.

Tan P-N, Steinbach M, & Kumar V (2013). Data mining cluster analysis: Basic concepts and algorithms. Introduction to Data Mining, 487, 533.

The Medicare Payment Advisory Commission. (2019). Report to the congress-Medicare payment policy: Home health care services. http://www.medpac.gov/docs/default-source/reports/mar19_medpac_entirereport_sec.pdf

Topaz M, Murga L, Bar-Bachar O, McDonald M, & Bowles K (2019). NimbleMiner: An open-source nursing-sensitive natural language processing system based on word embedding. Computers, Informatics, Nursing, 37(11), 583–590. 10.1097/cin.0000000000000557

Tullai-McGuinness S, Madigan EA, & Fortinsky RH (2009). Validity testing the outcomes and assessment information set (OASIS). Home Health Care Services Quarterly, 28(1), 45–57. 10.1080/01621420802716206 [PubMed: 19266370]

Wells N, Pasero C, & McCaffery M (2008). Improving the quality of care through pain assessment and management. In Patient safety quality: An evidence-based handbook for nurses. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008 Apr. Chapter 17.

Zolnoori M, McDonald MV, Barrón Y, Cato K, Sockolow P, Sridharan S, Onorato N, Bowles K, & Topaz M (2021). Improving patient prioritization during hospital-homecare transition: Protocol for a mixed methods study of a clinical decision support tool implementation. JMIR Research Protocols, 10(1), e20184. 10.2196/20184 [PubMed: 33480855]

Table 1.

Patient Characteristics and Information Extracted from Clinical Notes Between Patients with Hospitalization/ED Visit and those without. For bivariate analysis, student t-tests or Fisher exact tests were used, as appropriate (all *p*-value < 0.05).

	Patients without hospitalizations/ED visits (N = 69,897)	**Patients with hospitalizations/ED visits (N = 9,182)**
Length of episode (mean: days, SD)	30.6 (14.1)	39.9 (17.9)
Structured data (OASIS Item): Socio-demographic factor		
Age (mean: years, SD)	77.9 (11.6)	78.8 (12.7)
Gender: Female [n, (%)]	44,913 (64.3%)	5,675 (61.8%)
Race/Ethnicity [n, (%)]		
Asian	3,925 (5.62%)	411 (4.48%)
Black	11,708 (16.8%)	1,989 (21.7%)
Hispanic	9,007 (12.9%)	1,498 (16.3%)
White	44,936 (64.3%)	5,246 (57.1%)
Type of insurance [n, (%)]		
Dual eligibility	4,009 (5.7%)	753 (8.2%)
Medicare/Medicaid fee-for-service only	61,362 (87.8%)	7,659 (83.4%)
Any managed care	3,056 (4.4%)	595 (6.5%)
Other (e.g., private)	1,446 (2.1%)	175 (1.9%)
Living Condition: Living alone [n, (%)]	26,979 (38.5%)	3,599 (39.2%)
Structured data (OASIS Item): Medical conditions - Active diagnoses [n, (%)]		
Acute myocardial infarction	12,618 (18.1%)	1,957 (21.3%)
Arthritis	17,880 (25.6%)	1,361 (14.8%)
Cardiac dysrhythmias	10,505 (15.0%)	1,716 (18.7%)
Cancer	1,000 (1.43%)	395 (4.30%)
Diabetes	19,325 (27.6%)	3,421 (37.3%)
Dementia	8,894 (12.7%)	1,336 (14.6%)
Heart failure	9,019 (12.9%)	2,198 (23.9%)
Hypertension	45,574 (65.2%)	6,129 (66.8%)
Pulmonary disease	10,262 (14.7%)	1,813 (19.7%)
Renal failure	2,277 (3.26%)	729 (7.94%)
Skin ulcer	6,681 (9.56%)	1,737 (18.9%)
Structured data (OASIS Item): Risk for Hospitalization [n, (%)]		
History of falls in the past 12 months	14,904 (21.3%)	2,087 (22.7%)
Multiple hospitalizations in the past 6 months	15,654 (22.4%)	3,658 (39.8%)
Currently taking 5 or more medications	55,108 (78.8%)	7,698 (83.8%)
Decline in mental, emotional, or behavioral status in the past 3 months	9,850 (14.1%)	1,686 (18.4%)

	Patients without hospitalizations/ED visits (N = 69,897)	Patients with hospitalizations/ED visits (N = 9,182)
Structured data (OASIS Item): Sensory Status [n, (%)]		
Vision impaired	13,306 (19%)	2,141 (23.3%)
Hearing impaired	16,195 (23.2%)	2,417 (26.3%)
Difficulty in understanding verbal content	19,729 (28.3%)	3,058 (33.3%)
Difficulty in verbal expression	20,724 (29.7%)	3,424 (37.3%)
Having Pain	53,891 (77.1%)	6,813 (74.2%)
Structured data (OASIS Item): Integumentary Status [n, (%)]		
Having a risk of developing pressure ulcers	26,833 (38.4%)	4,486 (48.9%)
Having at least one Unhealed Pressure Ulcer at Stage II or Higher	4,169 (6%)	1,013 (11%)
Having stasis wound	1,255 (1.8%)	318 (3.46%)
Having surgical wounds	20,176 (28.9%)	1,670 (18.2%)
Having skin lesion or open wound	13,015 (18.6%)	2,416 (26.3%)
Structured data (OASIS Item): Elimination [n, (%)]		
Urinary Tract Infection in the past 14 days	4,457 (6.4%)	875 (9.5%)
Structured data (OASIS Item): Neuro, Emotional, and Behavioral Status [n, (%)]		
Cognitive functioning (i.e., required prompting, assistance or totally dependent)	8,370 (12%)	1,490 (16.2%)
Structured data (OASIS Item): Overall Status [n, (%)]		
Stable	4,849 (6.94%)	542 (5.90%)
Likely to be stable	55,260 (79.1%)	6,726 (73.3%)
Fragile	9,439 (13.5%)	1,820 (19.8%)
Serious	349 (0.50%)	94 (1.02%)
Structured data (OASIS Item): ADLs / IADLs		
ADL Needed [mean, (SD)][†]	8.05 (1.52)	8.24 (1.36)
ADL Severity [mean, (SD)][‡]	15.5 (6.74)	17.2 (7.51)
Unstructured Clinical Notes: Using the Omaha System as a risk factor (Identified through NLP approaches) [n, (%)]		
Abuse	1,375 (1.97%)	322 (3.51%)
Bowel function	2,759 (3.95%)	831 (9.05%)
Circulation	23,108 (33.1%)	3,842 (41.8%)
Cognition	10,654 (15.2%)	1,828 (19.9%)
Infectious condition	15,422 (22.1%)	3,070 (33.4%)
Consciousness	1,492 (2.13%)	554 (6.03%)
Digestion/hydration	4,798 (6.86%)	1,097 (11.9%)
Genitourinary function	1,684 (2.41%)	460 (5.01%)
Health care supervision	6,017 (8.61%)	864 (9.41%)
Medication regimen	3,311 (4.74%)	582 (6.34%)
Mental health	21,029 (30.1%)	3,585 (39.0%)

	Patients without hospitalizations/ED visits (N = 69,897)	Patients with hospitalizations/ED visits (N = 9,182)
Neglect	2,153 (3.08%)	471 (5.13%)
Nutrition	4,674 (6.69%)	1,092 (11.9%)
Neuro musculoskeletal function	32,143 (46.0%)	4,112 (44.8%)
Pain	32,414 (46.4%)	4,595 (50.0%)
Respiration	13,455 (19.2%)	2,990 (32.6%)
Skin	12,804 (18.3%)	2,535 (27.6%)
Social contact	12,828 (18.4%)	1,573 (17.1%)
Speech and language	2,730 (3.91%)	472 (5.14%)
Substance use	376 (0.54%)	71 (0.77%)

Note: SD = standard deviation; OASIS = Outcome and Assessment Information Set; ADLs / IADLs; Activities of Daily Livings/Instrumental Activities of Daily Livings; NLP = natural language processing.

†:"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).

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

Table 2.

Clinical Characteristics by Clusters. Consequently, *Cluster 1* is characterized by the combination of risk factors associated with “impaired physical comfort with pain,” *Cluster 2* by the combination of risk factors associated with “high comorbidity burden,” and *Cluster 3* by the combination of risk factors associated with “impaired cognitive/psychological and skin integrity.” We conducted analysis of variance (ANOVA) tests to examine the differences in clinical characteristics between each cluster (all p-values < 0.05).

	Cluster 1 (N = 37,678)	*Cluster 2* (N = 30,999)	*Cluster 3* (N = 10,402)
Hospitalization or ED visits [n, (%)]	3,032 (8%)	4,521 (14.6%)	1,629 (15.7%)
Length of episode [mean: days, (SD)]	28.3 (14.3)	35.1 (14.6)	33.8 (15.2)
Structured data (OASIS Item): Socio-demographic factor			
Age [mean: years, (SD)]	73.9 (11.2)	82.9 (9.91)	**85.1 (10.8)** ***
Gender: Female [n, (%)]	21,706 (57.6%)	**21,869 (70.5%)** **	7,013 (67.4%)
Race/Ethnicity: Black [n, (%)]	6,455 (17.1%)	**5,595 (18.0%)** **	1,647 (15.8%)
Race/Ethnicity: Hispanic [n, (%)]	4,519 (12.0%)	4,251 (13.7%)	**1,735 (16.7%)** ***
Race/Ethnicity: White [n, (%)]	**24,777 (65.8%)** *	19,185 (61.9%)	6,220 (59.8%)
Living Condition: Living alone [n, (%)]	14,467 (38.4%)	**13,606 (43.9%)** **	2,505 (24.1%)
Structured data (OASIS Item): Medical conditions - Active diagnoses [n, (%)]			
Acute myocardial infarction	6,049 (16.1%)	**7,018 (22.6%)** **	1,508 (14.5%)
Arthritis	**12,102 (32.1%)** *	5,660 (18.3%)	1,479 (14.2%)
Cardiac dysrhythmias	4,293 (11.4%)	**6,550 (21.1%)** **	1,378 (13.2%)
Diabetes	9,463 (25.1%)	**10,511 (33.9%)** **	2,772 (26.6%)
Dementia	1,070 (2.8%)	3,448 (11.1%)	**5,712 (54.9%)** ***
Heart failure	2,642 (7.01%)	**7,271 (23.5%)** **	1,304 (12.5%)
Pulmonary disease	4,795 (12.7%)	**6,342 (20.5%)** **	938 (9.02%)
Skin ulcer	1,688 (4.48%)	3,886 (12.5%)	**2,844 (27.3%)** ***
Structured data (OASIS Item): Prior medical conditions in the past 14 days [n, (%)]			
Urinary incontinent	4,818 (12.8%)	23,523 (75.9%)	**8,898 (85.5%)** ***
Urinary tract infection	1,329 (3.5%)	11,821 (38.1%)	**5,135 (49.4%)** ***

	Cluster 1 (N = 37,678)	*Cluster 2* (N = 30,999)	*Cluster 3* (N = 10,402)
Impaired decision-making	654 (1.74%)	1,914 (6.17%)	**4,582 (44.0%)** ***
Structured data (OASIS Item): Risk factors that may affect current health status [n, (%)]			
Obesity	4,687 (12.4%)	**4,948 (16%)** **	790 (7.6%)
Structured data (OASIS Item): Risk for Hospitalization [n, (%)]			
History of falls in the past 12 months	5,013 (13.3%)	**9,252 (29.8%)** **	2,726 (26.2%)
Multiple hospitalizations in the past 6 months	7,013 (18.6%)	**9,686 (31.2%)** **	2,613 (25.1%)
Currently taking 5 or more medications	28,684 (76.1%)	**26,020 (83.9%)** **	8,102 (77.9%)
Decline in mental, emotional, or behavioral status in the past 3 months	1,837 (4.9%)	4,994 (16.1%)	**4,705 (45.2%)** ***
Structured data (OASIS Item): Sensory Status [n, (%)]			
Vision impaired	3,415 (9.1%)	8,307 (26.8%)	**3,725 (35.8%)** ***
Hearing impaired	2,523 (6.7%)	11,318 (36.5%)	**4,771 (45.9%)** ***
Difficulty in understanding verbal content	2,832 (7.5%)	12,647 (40.8%)	**7,308 (70.3%)** ***
Difficulty in verbal expression	3,451 (9.2%)	11,360 (36.7%)	**9,337 (89.8%)** ***
Having Pain	**30,274 (80.4%)** *	23,976 (77.3%)	6,454 (62%)
Structured data (OASIS Item): Integumentary Status [n, (%)]			
Having a risk of developing pressure ulcers	5,263 (14.0%)	17,933 (57.9%)	**8,123 (78.1%)** ***
Having surgical wounds	**18,573 (49.3%)** *	2,718 (8.8%)	555 (5.3%)
Having skin lesion or open wound	6,565 (17.4%)	6,511 (21%)	**2,355 (22.6%)** ***
Structured data (OASIS Item): Neuro, Emotional, and Behavioral Status [n, (%)]			
Cognitive functioning (i.e., required prompting, assistance or totally dependent)	303 (0.8%)	474 (1.5%)	**9,083 (87.3%)** ***
Confusion	9,122 (24.2%)	19,224 (62.1%)	**9,896 (95.1%)** ***
Anxiety	9,243 (24.5%)	11,207 (36.2%)	**4,131 (39.7%)** ***
Structured data (OASIS Item): Overall Status [n, (%)]			
Stable	3,586 (9.5%)	1,304 (4.2%)	501 (4.8%)
Likely to be stable	31,631 (84%)	23,731 (76.6%)	6,624 (63.7%)
Fragile	2,414 (6.4%)	5,789 (18.7%)	**3,056 (29.4%)** ***

	Cluster 1 (N = 37,678)	*Cluster 2* (N = 30,999)	*Cluster 3* (N = 10,402)
Serious	47 (0.1%)	175 (0.6%)	**221 (2.1%)** ***
Structured data (OASIS Item): ADLs / IADLs			
ADL Needed [mean (SD)]†	7.4 (1.8)	8.5 (0.9)	**8.9 (0.6)** ***
ADL Severity [mean (SD)]‡	12 (4.0)	16.4 (5.5)	**24.6 (7.6)** ***
Unstructured Clinical Notes: Using the Omaha System as a risk factor (Identified through NLP approaches) [n, (%)]			
Circulation	11,750 (31.2%)	**12,423 (40.1%)** **	2,777 (26.7%)
Cognition	1,863 (4.94%)	5,303 (17.1%)	**5,316 (51.1%)** ***
Infectious condition	**9,566 (25.4%)** *	6,545 (21.1%)	2,381 (22.9%)
Neuro musculoskeletal function	16,721 (44.4%)	**15,710 (50.7%)** **	3,824 (36.7%)
Mental health	9,065 (24.1%)	11,035 (35.6%)	**3,973 (38.2%)** ***
Pain	**19,373 (51.4%)** *	14,500 (46.8%)	3,136 (30.1%)
Respiration	5,448 (14.5%)	**9,248 (29.8%)** **	1,749 (16.8%)
Skin	6,274 (16.7%)	5,806 (18.7%)	**3,259 (31.3%)** ***
Social contact	6,293 (16.7%)	**7,192 (23.2%)** **	9,16 (8.81%)

Note:

* = denotes the representative risk factors for grouping in Cluster 1;

** = denotes the representative risk factors for grouping in Cluster 2;

*** = denotes the representative risk factors for grouping in Cluster 3;

SD = standard deviation; OASIS = Outcome and Assessment Information Set; ADLs / IADLs; Activities of Daily Livings/Instrumental Activities of Daily Livings; NLP = natural language processing.

†: "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).

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