ORIGINAL RESEARCH **Open access**

Attention-Based Temporal Fusion Transformer for Forecasting Daily Census in Skilled Nursing Facilities Using Admission Patterns, Discharge Destinations, and Local COVID-19 Prevalence

Noura Al-Hinai[1*], Amal Al-Balushi[1], Saif Al-Shukaili[2]

Abstract

Skilled nursing facilities (SNFs) in the U.S. serve over 1.5 million residents and experience continuous census volatility driven by admissions, discharges, and mortality, impacting staffing, bed availability, and care quality. Existing forecasting methods rarely capture these dynamics together, leading to reactive and inefficient operational decisions. A need exists for accurate, multi-horizon, and data-integrated forecasting systems. Traditional models like ARIMA and LSTM are limited in SNF census forecasting because they produce single-point estimates, fail to model uncertainty, and cannot effectively integrate heterogeneous data such as facility characteristics, temporal utilization patterns, and external factors like COVID-19 prevalence. They also lack interpretability, reducing their usefulness for decision-making. This study introduces an attention-based Temporal Fusion Transformer (TFT) for multi-horizon SNF census forecasting (1, 7, 14, and 30 days). It integrates admissions, discharges, and COVID-19 prevalence through dedicated encoders and applies variable selection networks, LSTM layers, and multi-head attention to capture temporal dependencies and feature importance. The model outputs quantile forecasts (10th, 50th, 90th percentiles) to quantify uncertainty. The TFT enhances interpretability by identifying which past events and features most influence predictions at each horizon, enabling administrators to understand how admissions trends, discharge patterns, and COVID-19 surges affect census dynamics. The proposed framework enables proactive SNF capacity planning by combining multi-source data with interpretable, uncertainty-aware forecasting, supporting a shift from reactive staffing to anticipatory resource allocation and improved operational efficiency.

Keywords Interpretable machine learning, Temporal Fusion Transformer, Skilled nursing facility, Census forecasting, Multi-horizon prediction, COVID-19

*Correspondence:
Noura Al-Hinai
noura.hinai@outlook.com
[1] Department of Healthcare Informatics and AI, Sultan Qaboos University, Muscat, Oman
[2] Department of Intelligent Clinical Analytics, University of Nizwa, Nizwa, Oman

Introduction

Skilled nursing facilities represent a critical component of the post-acute care continuum in the United States, serving over 1.5 million residents annually through a combination of short-stay rehabilitation and long-term custodial care services. The daily census of a SNF—the number of residents present on a given day—functions as the fundamental determinant of staffing requirements, with labor costs constituting 40 to 60 percent of total operating expenses in these settings [1, 2]. When census drops

unexpectedly, facilities face the difficult choice of maintaining excess staff at significant financial cost or reducing staffing levels that may compromise care quality for remaining residents. Conversely, census surges without corresponding staffing increases can lead to missed treatments, medication errors, and increased risk of adverse events including falls and hospital readmissions [3].

The COVID-19 pandemic exposed and amplified the inherent fragility of SNF census management, creating unprecedented volatility that overwhelmed traditional manual forecasting approaches. Facilities experienced sudden admission freezes as hospitals suspended elective procedures and state health departments imposed moratoriums on new admissions during outbreaks, while simultaneously facing elevated mortality among residents who contracted the virus [4, 5]. Staff shortages driven by illness, quarantine requirements, and workforce attrition compounded these challenges, creating scenarios where facilities operated at reduced capacity despite having physical beds available [6, 7]. The pandemic demonstrated that external shocks, particularly infectious disease surges, can produce census fluctuations of a magnitude and speed that exceed the adaptive capacity of conventional operational planning [8, 9].

Daily SNF census emerges from a dynamic system governed by three interacting processes: admissions flowing from hospital discharge planners, direct community referrals, and inter-facility transfers; discharges occurring through planned returns to home, acute care hospital readmissions, and resident deaths; and external factors including seasonal illness patterns, regulatory changes, and local disease prevalence such as COVID-19 case rates [10, 11]. Each of these processes exhibits distinct temporal signatures, with hospital admissions typically peaking early in the week following weekend surgical schedules, readmissions clustering in the first week after SNF admission, and COVID-19 effects manifesting with a delay of 7 to 14 days following community case surges [12, 13]. Accurate census forecasting requires modeling these heterogeneous dynamics within a unified framework capable of representing both short-term operational fluctuations and longer-term trends driven by external conditions.

This article proposes a conceptual framework centered on the Temporal Fusion Transformer (TFT) architecture for multi-horizon SNF census forecasting at 1, 7, 14, and 30-day horizons, leveraging the model's capacity to integrate heterogeneous input types while providing interpretable attention weights and uncertainty estimates through quantile regression. The TFT's variable selection networks offer a principled mechanism for identifying which admission sources, discharge destinations, and external covariates drive predictions at each time step, transforming the forecasting process from an opaque numerical exercise into a transparent decision-support tool [14]. The framework prioritizes operational relevance by generating predictions at multiple horizons that align with distinct SNF decision cycles, from daily staffing assignments to weekly supply procurement and monthly budget planning [15]. The roadmap for this article proceeds through background context on SNF operations and forecasting methods, detailed articulation of the TFT framework components, and discussion of evaluation strategies and limitations.

Background

SNF census drivers

The daily census of a skilled nursing facility represents the net result of admission and discharge dynamics operating across multiple time scales with varying degrees of predictability. Admissions derive primarily from hospital discharge referrals, which account for the majority of post-acute SNF placements and are influenced by hospital occupancy, surgical volumes, and discharge planning practices that vary systematically by day of week and season [3, 10]. Community admissions, including direct entries from home and transfers from assisted living facilities, follow distinct patterns related to caregiver availability, functional decline trajectories, and holiday-period family decisions. The discharge side of the census equation comprises three primary destinations: returns to home or community settings following rehabilitation completion, acute care hospital readmissions that occur at rates of 10 to 25 percent within 30 days of SNF admission, and resident deaths that reflect the underlying clinical frailty of the SNF population [11, 12]. Length of stay distributions differ markedly between short-stay rehabilitation patients with Medicare coverage and long-stay custodial residents, creating bimodal occupancy dynamics that complicate aggregate forecasting [16].

COVID-19 Impact on SNFs

The COVID-19 pandemic subjected skilled nursing facilities to an unprecedented stress test that revealed fundamental

vulnerabilities in census management and infection control infrastructure. Facility-level outbreaks triggered cascading operational disruptions: state health departments mandated admission moratoriums that severed the primary inflow of hospital referrals, cohorting requirements forced facilities to dedicate wings or floors to COVID-positive residents thereby reducing effective bed capacity, and elevated mortality rates among infected residents accelerated census declines through the discharge-to-death pathway [4, 17, 18]. Staff shortages emerged as a binding constraint on operational capacity, with facilities unable to admit new residents despite available physical beds because insufficient nursing staff precluded meeting regulatory care hour requirements [6, 7]. Empirical analyses demonstrated that facilities with higher staffing ratios and those operated by unionized workforces experienced lower COVID-19 mortality rates, suggesting that pre-existing workforce conditions modulated pandemic impact [19, 20]. The pandemic period also revealed systematic undercounting of COVID-19 cases and deaths in federal reporting systems, with estimates suggesting that a substantial fraction of resident mortality went unreported during the early pandemic months [8, 9].

Census forecasting methods

Existing approaches to healthcare census forecasting span a methodological spectrum from classical statistical methods to modern machine learning architectures, each with characteristic strengths and limitations for SNF applications. Autoregressive integrated moving average (ARIMA) models and exponential smoothing techniques capture temporal dependencies in univariate census time series but cannot natively incorporate exogenous covariates such as local disease prevalence or hospital admission volumes [16]. Vector autoregressive models extend the ARIMA framework to multivariate settings, enabling joint modeling of admissions and discharges, though they impose linearity assumptions and struggle with the complex interaction effects that characterize healthcare operations. Recurrent neural network architectures including long short-term memory networks and gated recurrent units offer nonlinear modeling capacity and have been applied to healthcare occupancy prediction, yet they typically produce single-horizon point forecasts without uncertainty quantification and operate as black boxes that provide no insight into feature importance or prediction rationale [21]. Gradient-boosted tree methods such as XGBoost can incorporate heterogeneous feature sets and provide feature importance measures, but they treat time as a static feature rather than learning temporal dynamics, limiting their capacity to capture the sequential dependencies central to census evolution [22].

Temporal fusion transformer

The Temporal Fusion Transformer represents a significant advance in time series forecasting methodology, specifically designed to address the limitations of prior approaches for multi-horizon prediction with heterogeneous inputs and interpretable outputs [14]. The TFT architecture incorporates three core innovations: variable selection networks that learn which input features are relevant at each time step, gated residual network blocks that enable flexible nonlinear processing with adaptive depth, and a multi-head attention mechanism that provides interpretable weights over past time points for each prediction horizon. The model handles three categories of input variables through dedicated encoding pathways: static covariates that remain constant over the forecast period, past-observed time-varying inputs known only up to the prediction time, and future-known inputs such as day-of-week indicators available across the entire forecast horizon [2, 14]. Quantile regression at the output layer enables the TFT to produce prediction intervals at user-specified confidence levels, moving beyond point forecasts to capture the inherent uncertainty in complex time series [14, 23]. In healthcare applications, TFT models have demonstrated strong performance for intraoperative blood pressure forecasting, emergency department occupancy prediction, and population-level depression incidence forecasting [24-26].

Framework Overview

High-level architecture

The proposed framework operationalizes daily SNF census forecasting as a sequence-to-sequence prediction problem in which historical time series of admissions, discharges, and local COVID-19 prevalence are mapped to quantile forecasts at four predetermined horizons. Input data streams are organized at daily resolution, with the admission stream comprising hospital referral counts and community admission counts, the discharge stream enumerating home discharges, hospital readmissions, and deaths, and the external stream capturing county-level COVID-19 case rates per 100,000 population [5, 27]. These three streams are concatenated with temporal indicator variables including day of week and holiday markers, then

processed through the TFT architecture to generate predictions of facility census at 1, 7, 14, and 30 days into the future. The output layer produces three quantile estimates—10th percentile representing optimistic scenarios, 50th percentile representing the expected census, and 95th percentile representing pessimistic surge scenarios—for each horizon, together with attention weights that identify which historical observations most strongly influence each prediction [14].

Figure 1 presents the proposed attention-based Temporal Fusion Transformer framework for integrating SNF admission flows, discharge destinations, local COVID-19 prevalence, temporal covariates, and static facility characteristics into interpretable multi-horizon quantile census forecasts.

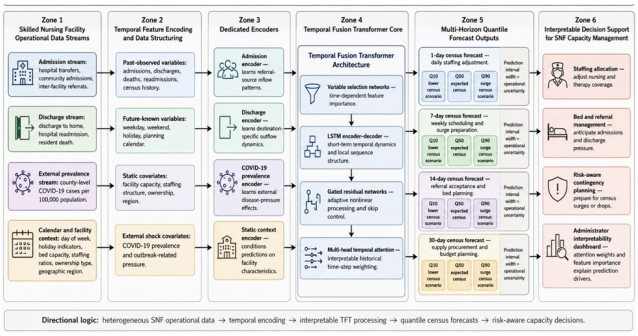

Figure 1. Attention-Based Temporal Fusion Transformer Framework for Multi-Horizon Skilled Nursing Facility Census Forecasting

Core assumptions

The framework rests on several operational and data assumptions that define the conditions under which reliable forecasting can be achieved. First, the framework assumes availability of at least two years of historical daily census data with corresponding admission and discharge logs at the individual resident level, enabling construction of daily count time series with sufficient temporal depth to capture seasonal patterns and learn the relationships between admissions, discharges, and net census change [17]. Second, the framework assumes access to contemporaneous local COVID-19 prevalence data at the county level, sourced from public health surveillance systems, which serve as the external covariate modulating admission and discharge dynamics. Third, the framework assumes that the structural relationships governing admissions and discharges remain relatively stable over time, such that patterns learned from historical data generalize to future periods barring major regulatory or reimbursement changes. These data requirements align with the information typically maintained by SNFs for regulatory compliance and billing purposes, suggesting feasibility of implementation in well-resourced facilities with electronic health record infrastructure.

Design principles

The framework design is guided by four principles that prioritize operational utility in real-world SNF settings. Multi-horizon prediction addresses the need for census forecasts at time scales corresponding to distinct operational decisions: daily staffing assignments, weekly supply orders, biweekly scheduling, and monthly budget projections. Interpretability ensures that facility administrators and nursing directors can understand not only what the model predicts but why it predicts that outcome, building trust and enabling human override when contextual factors beyond model inputs are relevant [18]. Uncertainty quantification through quantile regression enables risk-aware decision-making, allowing facilities to prepare for worst-case scenarios while optimizing resource allocation around expected values. Real-time update capability requires that the model can incorporate new observations as they become available, with prediction updates occurring daily as the previous day's actual census and admission and discharge counts are recorded, enabling a continuously refreshed operational picture [15].

Temporal Fusion Transformer Architecture

Variable selection networks

Variable selection networks constitute a foundational architectural component of the TFT that addresses the challenge of identifying which input features are relevant for prediction at each time step [14]. In the SNF census context, these networks receive as input the full set of covariates at each historical time point—including hospital admissions, community admissions, home discharges, hospital readmissions, deaths, COVID-19 case rates, and temporal indicators—and learn to assign importance weights that determine how much each feature contributes to the downstream processing layers. The variable selection mechanism operates through a gated residual network that processes each input feature independently, generating a feature-specific representation that is then

weighted by learned importance coefficients before being passed to the LSTM encoder-decoder [6]. This architecture provides two forms of interpretability: global feature importance aggregated across all time steps, revealing which inputs consistently drive predictions, and local feature importance at each time step, exposing how the predictive relevance of different variables shifts in response to changing conditions such as the onset of a COVID-19 surge.

LSTM encoder-decoder

The LSTM encoder-decoder component of the TFT processes the locally weighted input representations to capture short-term temporal dynamics that are essential for accurate census forecasting [7, 14]. The encoder processes a lookback window of historical observations, typically spanning 30 to 90 days, applying LSTM cells that maintain a hidden state summarizing the sequence of past inputs with emphasis on recent time steps due to the recurrent architecture's inherent recency bias. This local processing enables the model to learn short-term patterns such as the typical lag between a spike in hospital admissions and the resulting census increase, or the deceleration of discharges during weekend periods when therapy services are reduced. The decoder generates predictions for each horizon by unrolling the LSTM from the final encoder state, producing horizon-specific representations that capture the distinct temporal dynamics operating at each forecast distance. The gated skip connections and residual layers surrounding the LSTM blocks allow the network to adaptively control the degree of nonlinear processing applied to the input sequence, preventing overfitting when simpler linear relationships adequately characterize the data.

Multi-head attention

The multi-head attention mechanism provides the TFT's capacity to learn long-range temporal dependencies and generate interpretable attention weights that reveal which historical time points most strongly influence predictions at each horizon [8, 14]. Unlike the LSTM component that emphasizes recent observations through its recurrent structure, the attention mechanism can directly attend to any past time step, enabling the model to identify predictive patterns at specific lags such as weekly periodicity or holiday effects. In the SNF census application, attention weights may reveal that admissions occurring seven days prior to the prediction date carry substantial weight for the 7-day forecast, as those residents will have completed their typical short-stay rehabilitation and be approaching discharge. For the 30-day horizon, attention may concentrate on persistent features such as the baseline occupancy level and seasonal trends rather than daily fluctuations. The multi-head architecture runs multiple attention operations in parallel, each learning different temporal relationships, and the resulting attention patterns can be visualized and interpreted by facility administrators to validate that the model's predictions align with operational understanding of census dynamics [15].

Input Feature Encoding

Static covariates

Static covariates encode time-invariant characteristics of the skilled nursing facility that modulate the relationships between admissions, discharges, and census without themselves changing over the forecast period [9]. Bed capacity defines the physical upper bound on census and influences the rate at which admissions translate to occupancy changes, with facilities operating near capacity exhibiting different admission-to-census dynamics than those with substantial vacancy buffers. Staffing ratios, including nursing hours per resident day and therapist full-time equivalents, constrain the facility's ability to admit new residents and influence discharge timing, as higher staffing levels are associated with reduced hospital readmission rates and accelerated rehabilitation progress [20]. Facility ownership type—whether for-profit, non-profit, or government-operated—captures systematic differences in admission practices and length of stay that have been documented in the health services literature [4]. Geographic region and urban-rural designation encode differences in hospital referral networks and community disease prevalence patterns. These static covariates are processed through a separate encoder that generates context vectors used to modulate the dynamic processing of time-varying inputs throughout the TFT architecture.

Time-varying inputs

Time-varying inputs constitute the primary temporal signals from which the TFT learns census dynamics and are divided into past-observed variables known only up to the current time and future-known variables available across the entire prediction horizon. Past-observed inputs include daily counts of hospital admissions, community admissions, discharges to home, hospital readmissions, and resident

deaths, each extracted from facility electronic health record or administrative data systems [11, 13]. Local COVID-19 prevalence, measured as new cases per 100,000 population at the county level, enters as an external covariate reflecting community transmission intensity that influences both hospital referral volumes through acute care capacity strain and internal facility dynamics through staff absenteeism and cohorting requirements [5, 18]. Future-known inputs include day-of-week indicators that capture the strong weekly periodicity in healthcare operations, with hospital admissions peaking on weekdays and discharges declining on weekends, and holiday indicators that mark periods of reduced hospital surgical activity and altered family caregiving patterns [3]. This categorization into past-observed and future-known variables enables the TFT to appropriately handle inputs whose future values are known with certainty, improving forecast accuracy by incorporating calendar structure into the prediction process.

Attention Mechanisms

Interpretable attention weights

The multi-head attention mechanism within the Temporal Fusion Transformer generates interpretable attention weights that explicitly quantify the influence of each past observation on predictions at each forecast horizon, transforming the model from a black-box predictor into a transparent decision-support tool [14]. For the 7-day census forecast, attention patterns may reveal that hospital admissions occurring three to five days prior receive the highest weights, reflecting the typical post-acute length of stay for short-stay rehabilitation patients who constitute a substantial fraction of daily census. At the 30-day horizon, attention weights shift toward longer-term features including the moving average of community admissions over the preceding two weeks and the baseline occupancy level, as daily fluctuations are smoothed out over this extended prediction window. Visual inspection of attention weight distributions across time steps provides SNF administrators with a mechanism for validating model behavior against operational knowledge, such as confirming that the model appropriately reduces attention on periods when the facility was under an admission moratorium and therefore admissions data carry limited predictive signal [15]. This interpretability feature addresses a critical barrier to adoption of machine learning methods in healthcare operations, where clinical and administrative leaders require understanding of prediction rationale before incorporating algorithmic outputs into decision-making workflows.

Feature importance over time

The variable selection networks embedded within the TFT architecture produce time-dependent feature importance scores that reveal how the predictive relevance of different input variables evolves across the forecasting period and in response to changing external conditions [14]. During periods of stable community COVID-19 prevalence, the variable selection mechanism typically assigns highest importance to hospital admission counts and day-of-week indicators, reflecting the dominant role of post-acute referrals and weekly periodicity in driving census fluctuations under routine operating conditions. When local COVID-19 case rates surge, the variable importance profile undergoes a marked shift, with the COVID-19 prevalence encoder output rising to become the dominant predictor as outbreak-related admission restrictions, cohorting requirements, and elevated mortality reshape census dynamics [23]. The model may further reveal that the relative importance of discharge-to-death increases during pandemic surges, capturing the elevated mortality pathway through which COVID-19 reduces census, while the importance of discharge-to-home diminishes as rehabilitation completions are delayed or deferred. These time-varying importance profiles provide SNF leadership with a dynamic understanding of which operational levers and external factors are most impactful at any given time, enabling more targeted resource allocation and contingency planning.

Table 1 consolidates the interpretability functions of the Temporal Fusion Transformer and explains how each component transforms census prediction from a black-box forecast into transparent operational intelligence for SNF administrators.

Table 1. Interpretability Functions of the Temporal Fusion Transformer in SNF Census Forecasting

TFT interpretability component	What it explains	SNF-specific analytical contribution	Examp operatic interpret
Variable selection networks	Which input features matter at	Distinguishes whether admissions,	Risin COVID- prevale

	each time step	discharges, deaths, COVID-19 prevalence, or calendar effects dominate prediction	becom more important routin weekd admiss patter during a surge
Static covariate conditioning	How facility characteristics shape prediction behavior	Accounts for bed capacity, staffing ratios, ownership type, region, and facility context	A facility capaci respon different new hos referrals a facility high vac
LSTM encoder–decoder	How recent sequential patterns affect near-term forecasts	Captures short-term lags between admissions, discharges, readmissions, and census change	A spike hospit transfers the last t days increases 7-day expect censu
Multi-head temporal attention	Which historical days influence each forecast horizon	Reveals horizon-specific reliance on recent admissions, prior discharge waves, holiday effects, or baseline occupancy	The 30- foreca attends r strongly baseli census seasor patterns to yester admissi
Quantile output layer	How uncertain the census forecast is	Converts predicted census into lower, median, and upper planning scenarios	A wide Q Q90 inte signal unstab operati conditio durin outbrea

			referra disrupti

Multi-Horizon Forecasting

Quantile outputs

The TFT framework produces quantile forecasts at the 10th, 50th, and 90th percentiles for each prediction horizon, generating prediction intervals that quantify the uncertainty inherent in census forecasting and enabling risk-stratified operational planning [14]. Quantile regression is implemented through a pinball loss function during model training that asymmetrically penalizes over-prediction and under-prediction depending on the target quantile, with the 10th percentile loss function heavily penalizing over-prediction to ensure the model captures the lower bound of plausible census values and the 90th percentile loss function penalizing under-prediction to capture surge scenarios. The spread between lower and upper quantiles naturally widens at longer forecast horizons, reflecting the increasing uncertainty as predictions extend further into the future, and this widening provides a quantitative measure of how forecasting confidence degrades with horizon length [19]. During periods of stable operations with predictable admission and discharge patterns, the prediction intervals remain relatively narrow, whereas pandemic surge periods characterized by volatile hospital referral patterns and outbreak-related disruptions produce substantially wider intervals that honestly communicate elevated uncertainty to decision-makers rather than providing spuriously precise point estimates.

Decision support

Multi-horizon quantile forecasts translate directly into operational decision support by mapping prediction outputs to specific capacity management actions at each temporal scale, enabling SNF administrators to move from reactive staffing adjustments to anticipatory resource allocation [20]. The 90th percentile census forecast at the 7-day horizon provides a pessimistic scenario that can trigger proactive staff scheduling adjustments, ensuring adequate nursing coverage if the surge scenario materializes, while the 10th percentile forecast supports identification of opportunities for staff reassignment or bed closure without compromising care quality. At the 14-day horizon, the prediction interval width informs decisions about accepting new hospital referrals, with narrow intervals supporting confident

admission planning and wide intervals warranting conservative approaches that preserve bed availability for unexpected surges [22]. The 30-day forecast quantile range guides longer-term resource planning including supply procurement, maintenance scheduling, and capital budgeting, areas where the financial consequences of over- or under-estimating census compound over time [28]. By providing uncertainty estimates alongside expected values, the framework encourages risk-aware decision-making that appropriately balances the costs of excess capacity against the safety risks of inadequate staffing.

Table 2 clarifies how each forecast horizon corresponds to a distinct SNF decision cycle and shows how quantile outputs convert census uncertainty into operationally actionable planning ranges.

Table 2. Operational Decision Alignment of Multi-Horizon SNF Census Forecasts

Forecast horizon	Primary operational decision cycle	Quantile interpretation	Administrative action enabled
1 day	Next-day staffing and shift coverage	Q10 indicates likely low census; Q50 indicates expected resident load; Q90 indicates possible immediate surge	Adjust nursing assignments, therapy coverage, float staff, and agency staffing requests
7 days	Weekly scheduling and supply coordination	Q10 supports conservative staff deployment; Q90 identifies short-term surge risk	Schedule staff, coordinate admissions, anticipate therapy demand, prepare medication and dietary resources
14 days	Referral acceptance and bed-capacity planning	Wider quantile intervals signal uncertainty in mid-range admission and discharge balance	Decide whether to accept hospital referrals, hold beds, or delay elective maintenance
30 days	Budgeting, procurement, and strategic capacity planning	Long-horizon interval width reflects accumulated uncertainty from admissions, discharges, and external prevalence	Plan labor budgets, supply purchasing, census targets, and contingency reserves

Evaluation Strategy

Forecasting metrics

Evaluation of the TFT census forecasting framework employs metrics assessing both point prediction accuracy and uncertainty calibration [14]. The pinball loss function serves as the primary metric, quantifying quantile prediction accuracy with lower values indicating well-calibrated intervals, while mean absolute error on median forecasts provides interpretable error magnitudes in resident counts [22]. Prediction interval coverage assesses whether actual census values fall within the 10th–90th percentile range at the nominal 80 percent rate.

Baseline comparisons

Systematic comparison against established forecasting methods quantifies the marginal value of the TFT architecture's innovations [16]. Baseline models include ARIMA for univariate temporal dependencies, vector autoregressive models incorporating admission and discharge counts, LSTM networks offering nonlinear capacity without attention mechanisms, Prophet for automated seasonal decomposition, and XGBoost with lagged features capturing interactions without temporal dynamics learning [28]. All baselines are evaluated at identical horizons using consistent data splits.

COVID-19 model performance

Dedicated evaluation during the pandemic period assesses performance across three phases: pre-COVID baseline, initial surge with admission freezes and elevated mortality, and recovery with modified protocols [5, 8]. This stratified analysis reveals whether incorporating COVID-19 prevalence as an explicit input maintains accuracy when univariate models relying on historical patterns would fail. Generalizability testing through pre-pandemic training with pandemic-period evaluation probes the stability of learned relationships under structural change.

Limitations

Technical limitations

The framework requires approximately two years of stable operational data, limiting applicability to new facilities or those with major service changes, while the COVID-19 prevalence encoder's specificity to a particular pathogen raises questions about generalization to future shocks with different transmission dynamics. The daily resolution cannot capture intra-day fluctuations relevant for shift-level staffing, and the architecture still demands substantial machine learning expertise for configuration, tuning, and maintenance.

Operational limitations

Implementation faces heterogeneous electronic health record systems requiring facility-specific data pipeline customization, with common challenges including missing timestamps and inconsistent discharge coding [11]. Many SNFs operate batch processing cycles introducing multi-day lags, conflicting with the framework's near-real-time data assumption. Staff training must extend to nursing directors interpreting quantile forecasts, and organizational resistance to algorithmic decision support must be addressed through interpretability features that position the tool as augmenting human judgment.

Conclusion

This conceptual framework leverages the Temporal Fusion Transformer to integrate admissions, discharges, and COVID-19 prevalence for multi-horizon SNF census forecasting, moving beyond univariate methods to model the generative processes underlying occupancy dynamics. Dedicated encoders for each input category align architecture with operational patient flow, promoting both accuracy and interpretability.

The framework's key advantages center on multi-horizon quantile forecasting aligned with decision cycles from daily staffing to monthly budgeting, interpretable attention mechanisms revealing which historical observations drive predictions, and variable selection networks providing time-dependent feature importance that supports dynamic resource allocation.

Significant limitations include two-year data requirements restricting applicability, unresolved questions about generalization beyond COVID-19 to future shocks, and implementation barriers encompassing data heterogeneity, quality challenges, and organizational readiness that require dedicated implementation science inquiry.

Future work should pursue validation using CMS resident assessment and claims databases, state regulatory datasets, and multi-facility chain electronic health records. Retrospective validation across pandemic phases would quantify accuracy under stable and disrupted conditions, while prospective pilots would evaluate operational feasibility. Extensions could incorporate influenza prevalence, extreme weather, and hospital occupancy, and explore transfer learning for facilities with limited historical data.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Dec 2024 Revised: 17 Jan 2025 Accepted: 16 Feb 2025
Published online: 20 July 2025

References

Toth M, Palmer L, Marino ME, Smith A, Schwartz C, Deutsch A, et al. Validation of the standardized function data elements among Medicare skilled nursing facility residents. J Am Med Dir Assoc. 2023;24(3):307-13.

White HK, et al. Post-acute care: current state and future directions. J Am Med Dir Assoc. 2019;20(4):392-5.

Gardner RL, Pelland K, Youssef R, Morphis B, Calandra K, Hollands L, et al. Reducing hospital readmissions through a skilled nursing facility discharge intervention: a pragmatic trial. J Am Med Dir Assoc. 2020;21(4):508-12.

Brown KA, Jones A, Daneman N, Chan AK, Schwartz KL, Garber GE, et al. Association between nursing home crowding and COVID-19 infection and mortality in Ontario, Canada. JAMA Intern Med. 2021;181(2):229-36.

Abrams HR, Loomer L, Gandhi A, Grabowski DC, et al. Characteristics of US nursing homes with COVID-19 cases. J Am Geriatr Soc. 2020;68(8):1653-6.

McGarry BE, Grabowski DC, Barnett ML, et al. Severe staffing and personal protective equipment shortages faced by nursing homes during the COVID-19 pandemic. Health Aff (Millwood). 2020;39(10):1812-21.

McGarry BE, Barnett ML, Grabowski DC, Gandhi AD, et al. Nursing home staff vaccination and COVID-19 outcomes. N Engl J Med. 2022;386(4):397-8.

Shen K, Loomer L, Abrams H, Grabowski DC, Gandhi A, et al. Estimates of COVID-19 cases and deaths among nursing home residents not reported in federal data. JAMA Netw Open. 2021;4(9):e2122885.

Zhu X, Lee H, Sang H, Muller J, Yang H, Lee C, et al. Nursing home design and COVID-19: implications for guidelines and regulation. J Am Med Dir Assoc. 2022;23(2):272-9.

Weerahandi H, Li L, Bao H, Herrin J, Dharmarajan K, Ross JS, et al. Risk of readmission after discharge from skilled nursing facilities following heart failure hospitalization: a retrospective cohort study. J Am Med Dir Assoc. 2019;20(4):432-7.

Burke RE, Whitfield EA, Hittle D, Min SJ, Levy C, Prochazka AV, et al. Hospital readmission from post-acute care facilities: risk factors, timing, and outcomes. J Am Med Dir Assoc. 2016;17(3):249-55.

Carnahan JL, Kaehr EW, Wagle KC, et al. Opportunities for collaboration: refining postoperative readmission risk for skilled nursing facility patients. J Am Med Dir Assoc. 2019;20(9):1060-2.

Dykgraaf SH, Matenge S, Desborough J, Sturgiss E, Dut G, Roberts L, et al. Protecting nursing homes and long-term care facilities from COVID-19: a rapid review of international evidence. J Am Med Dir Assoc. 2021;22(10):1969-88.

Lim B, Arık SÖ, Loeff N, Pfister T, et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast. 2021;37(4):1748-64.

White EM, Wetle TF, Reddy A, Baier RR, et al. Front-line nursing home staff experiences during the COVID-19 pandemic. J Am Med Dir Assoc. 2021;22(1):199-203.

Riester MR, Bosco E, Silva JB, Bardenheier BH, Goyal P, O'Neil ET, et al. Causes and timing of 30-day rehospitalization from skilled nursing facilities after a hospital admission for pneumonia or sepsis. PLoS One. 2022;17(1):e0260664.

He M, Li Y, Fang F, et al. Is there a link between nursing home reported quality and COVID-19 cases? Evidence from California skilled nursing facilities. J Am Med Dir Assoc. 2020;21(7):905-8.

Gorges RJ, Konetzka RT, et al. Staffing levels and COVID-19 cases and outbreaks in US nursing homes. J Am Geriatr Soc. 2020;68(11):2462-6.

Dean A, Venkataramani A, Kimmel S, et al. Mortality rates from COVID-19 are lower in unionized nursing homes. Health Aff (Millwood). 2020;39(11):1993-2001.

Weiss M, Normand SL, Grabowski DC, Blacker D, Newhouse JP, Hsu J, et al. All-cause nursing home mortality rates have

remained above pre-pandemic levels after accounting for decline in occupancy. Health Aff Sch. 2024;2(11):qxae126.

Ayele R, Manges KA, Leonard C, Lee M, Galenbeck E, Molla M, et al. How context influences hospital readmissions from skilled nursing facilities: a rapid ethnographic study. J Am Med Dir Assoc. 2021;22(6):1248-54.

Burke RE, Xu Y, Rose L, et al. Skilled nursing facility performance and readmission rates under value-based purchasing. JAMA Netw Open. 2022;5(2):e220721.

Li Y, Temkin-Greener H, Shan G, Cai X, et al. COVID-19 infections and deaths among Connecticut nursing home residents: facility correlates. J Am Geriatr Soc. 2020;68(9):1899-906.

Kapral L, Dibiasi C, Jeremic N, Bartos S, Behrens S, Bilir A, et al. Development and external validation of temporal fusion transformer models for continuous intraoperative blood pressure forecasting. EClinicalMedicine. 2024;75.

Tuominen J, Pulkkinen E, Peltonen J, Kanniainen J, Oksala N, Palomäki A, et al. Forecasting emergency department occupancy with advanced machine learning models and multivariable input. Int J Forecast. 2024;40(4):1410-20.

Yang D, Tang Y, Chan VKY, Fang Q, Chan SSM, Luo H, et al. Population-wide depression incidence forecasting comparing autoregressive integrated moving average and vector autoregressive integrated moving average to temporal fusion transformers: longitudinal observational study. J Med Internet Res. 2025;27:e67156.

Yun D, Yang HL, Kim SG, Kim K, Kim DK, Oh KH, et al. Real-time dual prediction of intradialytic hypotension and hypertension using an explainable deep learning model. Sci Rep. 2023;13(1):18054.

Kandel B, Field C, Kaur J, Slawson D, Ouslander JG, et al. Development of a predictive hospitalization model for skilled nursing facility patients. J Am Med Dir Assoc. 2025;26(1):105288.