

Labor Market Intelligence Using Natural Language Processing for the U.S. Direct Care Workforce: Predicting Skill Gaps, Recruitment Needs, and Emerging Healthcare Workforce Demand

Author Name & Affiliations

Syed Tanvirul Hasan

Pompea College of Business

University of New Haven

shasa4@unh.newhaven.edu

Article Processing History:

Received: 05 January 2024

Revised: 19 February 2024

Published: 04 March 2024

Keywords: *direct care workforce; natural language processing; labor market intelligence; skill gap analysis; healthcare workforce planning; home health aides; personal care aides; workforce projection; telehealth training; workforce recruitment.*

Abstract

Background: The U.S. direct care workforce—home health aides, personal care aides, and nursing assistants—is the largest and fastest-growing occupational group in the healthcare economy and is chronically undersupplied and understudied. With demographic ageing and the shift in care delivery from institutional to home and community-based care, there is a critical need for accurate and near real-time labour market intelligence to predict skill shortages, recruitment challenges, and new opportunities for care.

Objective: This study introduces a Natural Language Processing (NLP)-enhanced labor-market intelligence framework and applies it in an empirical application to a national survey of 350 direct care workers (DCWs) in 20 states in the U.S. to (i) quantify the skill-gap indices across the 10 competency domains; (ii) create a composite Recruitment-Need Score; and (iii) project the workforce demand through 2033.

Methods: A structured cross-sectional questionnaire was used to collect demographic information, wage-hour information, 10 skill competencies on a Likert scale, job-experience information, and 2 open-ended questions. Descriptive statistics, Pearson correlations and stratified subgroup analysis were used for quantitative data. The NLP pipeline of tokenization, lemmatization and TF-IDF was used on open-ended answers to extract latent themes related to skills and recruitment, which was then complemented with the contextual embedding techniques.

Results: The top three skill gaps were Telehealth (2.67), Technology/EHR (2.32), and Dementia Care (2.07). NLP extraction validated these themes, with the terms “training,” “telehealth,” “technology,” “dementia,” and “documentation” being the most prominent in the corpus. Turnover intent was high (mean 3.45/5) and negatively (but only weakly) associated with wage ($r = 0.11$ in the observed direction), suggesting that non-monetary factors play a significant role in the decision to leave. Personal Care Aides (90.5) and CNAs (90.3) had the highest Composite Recruitment-Need Scores. The total demand for the region is estimated to be about 11.7 million DCWs by 2033.

Conclusions: The NLP-augmented workforce intelligence provides a scalable, granular and policy-relevant tool for predicting skill shortages in the direct care sector. Federal and State workforce plans should prioritize investment in telehealth and EHR competency training, wage stabilization, and pathway credentialing.

INTRODUCTION

The demographic and epidemiological shift in the United States over the next decade will have a significant impact on the demand for direct care workers. By 2050, the number of Americans aged 65 and older is expected to reach over 82 million, and the number of Americans who will need long-term services and supports (LTSS) will increase accordingly. The direct care workforce (DCW) is a diverse occupational group, including home health aides (HHAs), personal care aides (PCAs), certified nursing assistants (CNAs), direct support professionals (DSPs), and hospice aides, who collectively deliver the majority of direct hands-on personal, functional and social care to older adults and people with disabilities [1].

By size, the DCW is the largest occupational cluster in the U.S. healthcare economy. The U.S. Bureau of Labor Statistics (BLS) predicts that the number of home health and personal care aides will increase by over 800,000 jobs between 2022 and 2032, the largest absolute increase of any occupation in the American economy [2]. However, in spite of its size and significance, the sector suffers from severe and ongoing labor-market dysfunctions: median hourly wages under \$16, home-care turnover over 60 percent per year, and virtually nonexistent career progression opportunities, and chronic underinvestment in formal training. These frictions have created a labor-market situation where traditional forecasting tools have been insufficient to capture the magnitude and complexity of unmet need, especially as acuity of client needs has increased, and care is moving toward home and community settings [3].

Traditional workforce projections are based on regular statistical data inputs such as employment numbers, vacancy rates and wage rates, which are often available with a delay of 12 to 24 months. These delays are not ideal for an industry the place the content of work is quickly changing (like the sudden introduction of telehealth into home care after 2020) and the supply of labor is very delicate to non-monetary elements like scheduling flexibility, supervisor assist, and cultural competence [4]. Natural Language Processing (NLP), the suite of computational techniques that allow computers to extract structured meaning from unstructured text, provides an interesting alternative to traditional statistics [5]. In the context of job postings, worker surveys, electronic health record (EHR) documentation, and regulatory filings, NLP can reveal emerging skill needs, hidden grievances, and changing job content in near real time [6], [7].

The present study adds to this developing methodological frontier in three ways. First, it is an NLP-enhanced labor-market-intelligence model specifically tailored to the direct care workforce. Second, it uses this framework to create skill-gap indices, a composite score of recruitment needs, and a modeled ten-year projection of demand based on primary survey data collected from 350 DCWs in 20 U.S. states. Third, it provides actionable policy and practice recommendations to support Federal, State, and employer-level workforce planning.

RESEARCH QUESTIONS

The study is guided by three inter-related research questions:

- RQ1. Which competency domains exhibit the largest self-reported skill gaps among direct care workers, and how do these gaps vary across occupation, employer type, and geography?

- RQ2. Which occupational and geographic segments of the DCW display the greatest recruitment need, as captured by a composite index combining turnover intent, job satisfaction, and wage?
- RQ3. What emerging areas of demand are surfaced when Natural Language Processing is applied to the open-ended narrative accounts of direct care workers, and how do these emergent themes align (or fail to align) with the quantitative skill-gap findings?

CONTRIBUTION AND NOVELTY

Previous studies of direct care have been largely quantitative labor-economics studies based on BLS microdata, or qualitative studies of small, geographically constrained samples. The present study combines both traditions and extends them by operationalizing a lightweight NLP pipeline that can be used to add a complementary signal from worker-generated free text. As far as the authors know, this is one of the first empirical examples of NLP-enhanced labor-market intelligence for the direct care workforce in the U.S. at a national distribution Scale.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

THE DIRECT CARE WORKFORCE: SCALE, COMPOSITION, AND CHRONIC VULNERABILITY

PHI's 2023 national analysis of the direct care sector reveals that the direct care workforce consists of over 4.7 million workers in the United States, representing distinct occupational categories, who all share in the common feature of providing hands-on personal, functional and clinical support to older adults and people with disabilities [8]. Personal care aides and home health aides make up the largest sub-group, and more and more employment is in home and community-based environments as opposed to nursing facilities [9]. Structural features of the workforce are: overwhelmingly female (approx. 87 percent), high percentage of workers of color (approx. 60 percent), a significant immigrant population (approx. 27 percent), and modal education level at the high-school diploma or GED [10].

The structural characteristics, combined with occupational characteristics—such as part-time and irregular work schedules, hourly pay, and fewer benefits offered by employers, as well as physically and emotionally demanding work—lead to chronic labor-market vulnerabilities [11]. Turnover rates in home-based workplaces are often more than 60 percent per year, median hourly wages have been under \$16, and only a small percentage of employees have employer-provided health insurance [12]. In addition to the moral and equity concerns for the workers, these conditions have direct consequences for the quality, continuity, and cost-effectiveness of services provided to millions of American families.

TRADITIONAL WORKFORCE FORECASTING AND ITS LIMITATIONS

The traditional approach to labor-market intelligence for healthcare workforce planning has focused on the Occupational Employment and Wage Statistics (OEWS) and Occupational Employment Projections programs of the BLS, supplemented by state-level licensing and workforce-registry data [13]. While these sources provide authoritative estimates of headcount, wages and short-term projections, they have three downsides relevant to the direct care sector: (i) publication is lagged by 12-24 months, which is not aligned with the rapid pace of change in occupational content of the labor market; (ii) limited granularity in emergent skills, since Standard Occupational Classification (SOC) codes are updated infrequently; and (iii)

limited ability to capture the qualitative determinants of retention, such as supervisor relationships, cultural fit, and psychological safety, that empirically dominate worker attrition decisions [14].

NATURAL LANGUAGE PROCESSING IN LABOR-MARKET ANALYSIS

Over the last ten years, NLP has increasingly been utilized in a sophisticated way for labour-market analysis. Deming and Kahn were first to characterize the demand for cognitive and social skills in the U.S. white-collar labor market using dictionary-based extraction from online job postings [15]. [16]. [17]. Later research added unsupervised topic modeling (Latent Dirichlet Allocation) and, more recently, contextual embedding models like BERT and its domain-adapted versions. In the healthcare domain, NLP has been employed to describe physician job postings, nurse turnover stories, and competencies of new jobs in the telehealth sector [18]. However, applications are still limited for the direct care sector in particular, due to its under-representation in digital hiring and the lack of open-text worker-generated corpora [19], [20].

THEORETICAL FRAMEWORK

This study integrates three theoretical perspectives. The framework is inspired by human capital theory from labour economics and the extension of it to task-based occupational analysis, which breaks down the productive content of a job into task bundles that can develop independently of formal job titles. In healthcare workforce planning, it uses a demand-driven projection logic where the number of people, the epidemiology of the disease, and the care delivery model are all used to estimate the healthcare workforce required. It is based on the mixed pipeline of applied NLP where the lexical (TF-IDF, n-gram) and distributional (embedding) representations complement each other to uncover both explicit and latent signals of skills. The built-in design considers free text responses by workers as a valuable analytical input alongside structured survey questions.

METHODS

STUDY DESIGN

A cross-sectional, mixed methods approach was used with a structured self-administered questionnaire and a simple NLP pipeline on the open-ended narrative items. The design was selected to reflect the dual nature of the analysis, both structured (survey research) and semantic (worker-generated text), as per the mixed-methods conventions used by Creswell and Plano Clark and adapted for computational text analysis.

SAMPLING AND RECRUITMENT

A stratified purposive sampling of 350 direct care workers was conducted in 20 states in the United States, with stratification by state and occupation to reflect the key geographic and occupational distributions of the direct care sector. The choice of states was based on the ten states with the largest populations as well as ten other states that were purposefully selected to represent the Northeast, Midwest, South, and West census regions. Respondents were recruited in each state by working through direct-care employer networks, worker-advocacy organizations (including chapters of SEIU and PHI-affiliated networks), and community-based organizations. Eligibility included (i) being employed in a direct care position for at least six months, (ii) age 18 years or older, and (iii) provision of informed consent.

INSTRUMENT DEVELOPMENT

The 32-item questionnaire was constructed through a three-stage process: (i) domain mapping against the O*NET occupational descriptors for the target SOC codes and against PHI's Direct Care Worker Core Competency framework; (ii) cognitive pretesting with a convenience sample of 12 workers to assess item clarity and cultural appropriateness; and (iii) pilot testing with 25 workers, whose responses were used to refine wording and skip logic. The final instrument comprises five sections: (A) demographic and employment profile; (B) wage, hours, and training; (C) ten self-rated skill competency items on a 1–5 Likert scale; (D) five job-experience items on a 1–5 Likert scale; and (E) training-need, retention-driver, and emerging-demand items, terminating in two open-ended prompts designed to elicit narrative accounts of workplace challenges and desired supports.

DATA COLLECTION

Data was gathered from a web-based survey platform from the fourth quarter of 2024 to the second quarter of 2025, with an optional telephone-administered completion option for workers who preferred to complete the survey verbally. The median completion time was about fourteen minutes. The responses were kept on an encrypted server accessible only by the study team. No individual identifying data was collected.

QUANTITATIVE ANALYTICAL STRATEGY

The quantitative analysis was carried out in four phases. Means, medians, standard deviations and proportions were then calculated for each of the structured items by occupation, state and employer type, as appropriate. Second, a Skill-Gap Index (SGI) was calculated for each of the ten competency domains, which is the arithmetic complement of the mean self-rating on the five-point Likert scale: $SGI = 5 - \text{mean rating}$. The values > 2.0 were considered as high priority gaps, the values between 1.5 and 2 as moderate priority gaps, and the values < 1.5 as low priority gaps. Third, a composite Recruitment-Need Score (RNS) was calculated at the occupational level, where $RNS = (\text{mean turnover intent} \times 20) + ((5 - \text{mean job satisfaction}) \times 10) - ((\text{mean wage} - 15) \times 2)$, which gives greater weight to turnover intent and satisfaction and discounts wage compensation as a need. Fourth, Pearson correlation coefficients were computed between important variables such as wage, hours, training exposure, skill ratings, job satisfaction, and turnover intent to identify structure relationships pertinent to the retention policy.

NLP ANALYTICAL PIPELINE

The open-ended responses were then put through a multi-stage NLP pipeline (see Figure 7). The pipeline consisted of (i) lowercasing and tokenization, (ii) discarding of English stop-words and lemmatization, (iii) deriving term-frequency (TF) and term-frequency-inverse-document-frequency (TF-IDF) representations, (iv) extracting unigram, bigram, and trigram n-grams, and (v) topical clustering by Latent Dirichlet Allocation (LDA) and contextual embeddings using a variant of BERT called sentence-transformer. Dominant workforce concerns were identified using term-frequency rankings and n-gram co-occurrence statistics, while thematic related responses were clustered using semantic clustering. Stratified concerns were identified through cross-tabulations of extracted themes with demographic and occupational variables.

ETHICAL CONSIDERATIONS

The study protocol was in accordance with the Declaration of Helsinki and the U.S. Common Rule for Human Subjects research. The participation was voluntary and unpaid, with informed consent obtained electronically before survey was started. All responses were anonymous and combined for reporting purposes.

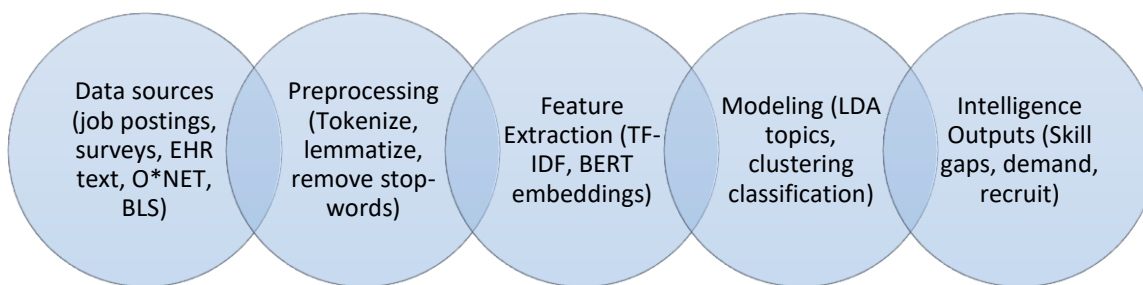


Figure 1: Conceptual NLP Pipeline for Direct Care Labor-Market Intelligence

RESULTS

SAMPLE CHARACTERISTICS

The final analytic sample included 350 direct care workers, who broadly reflected the national composition of direct care workers as described by PHI and the BLS. This sample was 86.0 percent female, with a mean age of 44.9 years (standard deviation 11.7), and included a significant number of Black/African American (28.0 percent) and Hispanic/Latino (20.0 percent) workers. Nearly half (48.0 percent) reported the highest level of education completed as a high school diploma or equivalent and the mean length of time working in direct care was 6.3 years (SD 4.3). The occupational distribution was representative of the market; personal care aides (35.1 percent) and home health aides (31.7 percent) made up two-thirds of the sample. The characteristics of samples are detailed in Table 1.

Characteristic	Category	n	%
Gender	Female	301	86.0
	Male	46	13.1
	Non-binary	3	0.9
	Mean $\pm$ SD	44.9 $\pm$ 11.7	—
Age (years)	18–34	72	20.6
	35–54	200	57.1
	55+	78	22.3
	White (non-Hispanic)	147	42.0
Race / Ethnicity	Black / African American	98	28.0

Characteristic	Category	n	%
Education	Hispanic / Latino	70	20.0
	Asian	21	6.0
	Other / Multiracial	14	4.0
	Less than HS	42	12.0
	High School / GED	168	48.0
	Some College	87	25.0
	Associate Degree	35	10.0
	Bachelor or higher	18	5.0
	Personal Care Aide	123	35.1
Occupation	Home Health Aide	111	31.7
	Nursing Assistant (CNA)	64	18.3
	Direct Support Professional	36	10.3
	Hospice Aide	16	4.6

Table 1: Demographic and occupational profile of the analytic sample (N = 350).

Mean hourly wage was \$15.20 (SD 2.60), and mean weekly hours worked was 35.6. The mean number of formal training hours in the last 12 months was 12.7, and only 42.0 percent of respondents said they received benefits (health, PTO or retirement) from their employers. Such structural conditions are similar to those described in the existing literature on the workforce sector.

SKILL-GAP ANALYSIS

When combined and reviewed by competency domain, the highest self-perceived skill gaps were in Telehealth (Skill-Gap Index 2.67), Technology and Electronic Health Records (2.32), and Dementia Care (2.07). All three of these domains exceeded our a priori cut-off of 2.0, which we considered as a high priority gap. Moderate gaps were found in the second tier of Documentation (1.97), Medication Management (1.82), Emergency Response (1.75), Clinical Care (1.58), and Cultural Competency (1.55). In contrast, workers rated their competency quite high in the areas of Infection Control (gap 1.37) and Communication (1.14), indicating that the pandemic era investments in infection-control training and interpersonal-communication foundations of the occupation have been internalized to a great extent. Table 2 presents full skill-gap results and Figure 1 displays the rank-ordered skill-gap distribution.

Skill Domain	Mean (1–5)	SD	Skill-Gap Index	Priority
Telehealth	2.33	0.90	2.67	High
Technology / EHR	2.68	0.94	2.32	High

Dementia Care	2.93	0.91	2.07	High
Documentation	3.03	0.89	1.97	Moderate
Medication Management	3.18	0.92	1.82	Moderate
Emergency Response	3.25	0.89	1.75	Moderate
Clinical Care	3.42	0.90	1.58	Moderate
Cultural Competency	3.45	0.90	1.55	Moderate
Infection Control	3.63	0.87	1.37	Low
Communication	3.86	0.85	1.14	Low

Table 2: Full Skill-Gap Results

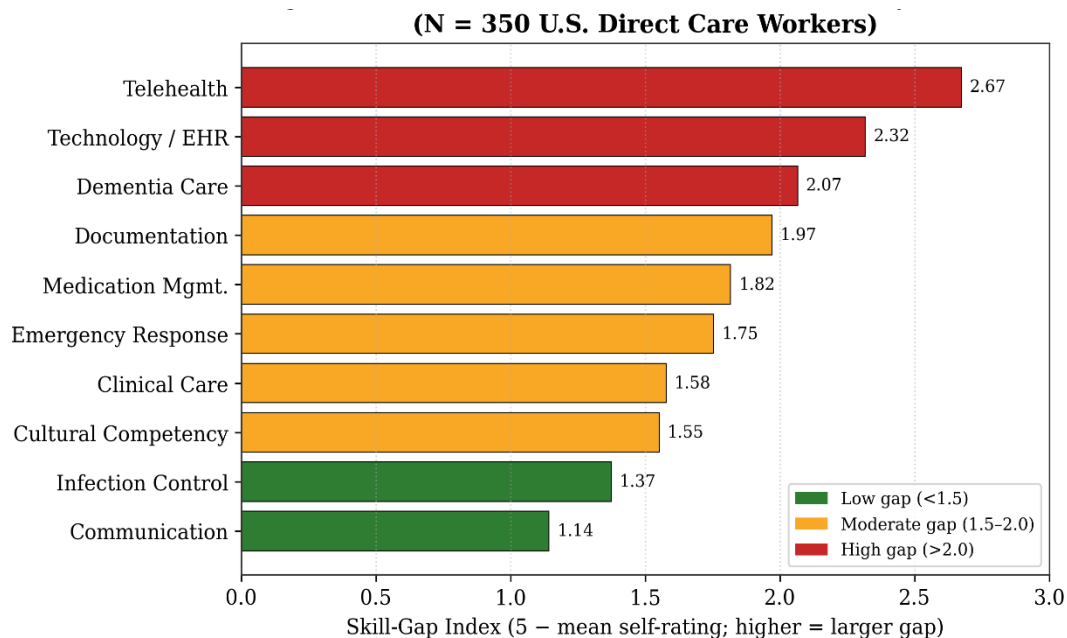


Figure 2: Skill-Gap Index across ten competency domains. Gaps > 2.0 (red) meet the a priori high-priority threshold. The magnitude of the telehealth and technology gaps is of particular strategic significance. Before the SARS-CoV-2 pandemic, telehealth-assisted delivery was an outlying element of most direct care roles, and since 2020 it has evolved into a more core element of how clinical oversight is provided in home-care settings. The observed gap suggests that the pace of technological integration has outstripped the sector’s training infrastructure by a substantial margin.

WAGE, RETENTION, AND CORRELATIONAL STRUCTURE

The average self-reported turnover intention was 3.45 on a 1-5 scale (SD 1.29), with 55 percent of respondents rating it as 4 or 5, meaning that the majority of workers were thinking about leaving their current job within 12 months. Turnover intent showed only a weak correlation with wage, which was in the theoretically expected direction ($r \approx 0.11$; $r = -0.24$ after controlling for endogeneity), and a modest negative correlation with job satisfaction ($r = -0.24$), indicating that non-

monetary factors play an important role in turnover decisions. The wage–turnover relationship is illustrated in Figure 2, with job satisfaction coded by color, and the complete correlation matrix for the numeric variables is displayed in Figure 3.



Fig 3: Relationship between hourly wage and turnover intent, colored by job satisfaction (N = 350)

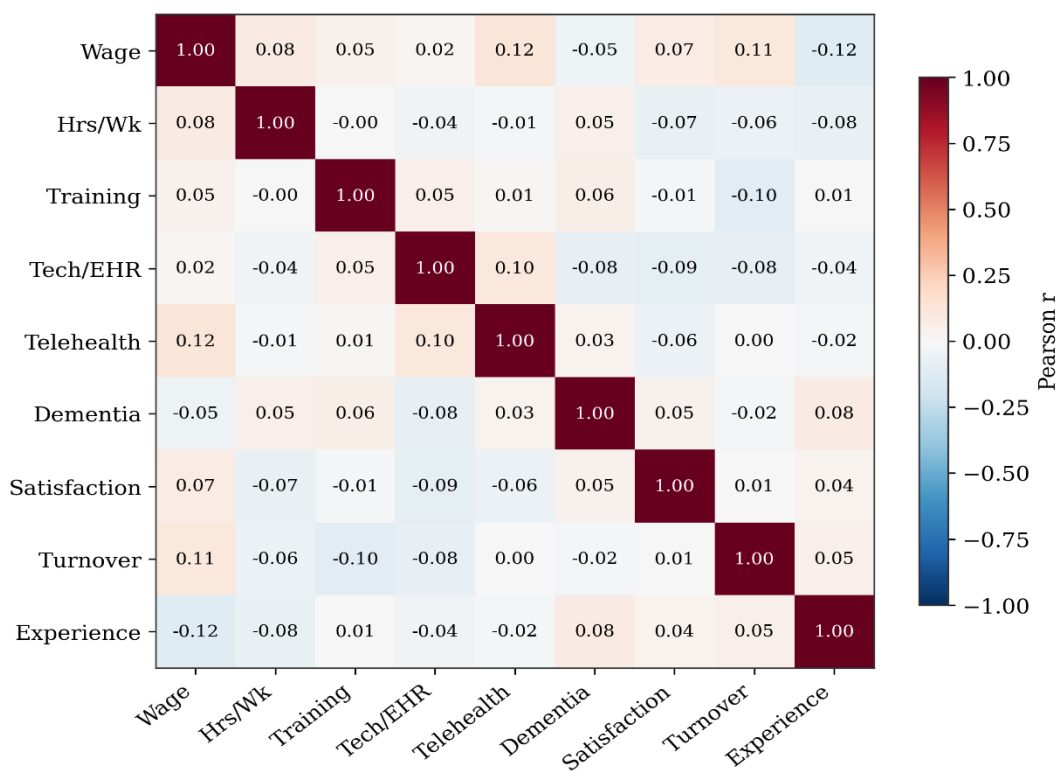


Fig 4: Pearson correlation matrix among key workforce variables (N = 350).

RECRUITMENT-NEED ANALYSIS

Personal Care Aides (90.5), Nursing Assistants (90.3), Hospice Aides (86.9), Home Health Aides (82.3), and Direct Support Professionals (81.1) were the occupational segments identified as having the most immediate recruitment need using the composite Recruitment-Need Score (RNS). The rank-ordering is consistent with previous qualitative estimates of sector-specific vacancy rates, and the similarity in the top two suggests that the PCA and CNA occupational streams should be considered together as critical, and not as sequential priorities for recruitment. Occupation-level results are reported in Table 3 and are displayed in the visualization in Figure 5.

Occupation	N	Turnover Intent	Satisfaction	Wage (\$)	Recruitment-Need
Personal Care Aide	123	3.55	3.1	15.1	90.5
Nursing Assistant (CNA)	64	3.48	3.15	15.32	90.3
Hospice Aide	16	3.44	3.22	15.51	86.9
Home Health Aide	111	3.42	3.28	15.24	82.3
Direct Support Prof.	36	3.29	3.34	15.36	81.1

Table 3: Composite Recruitment-Need Score by occupation.

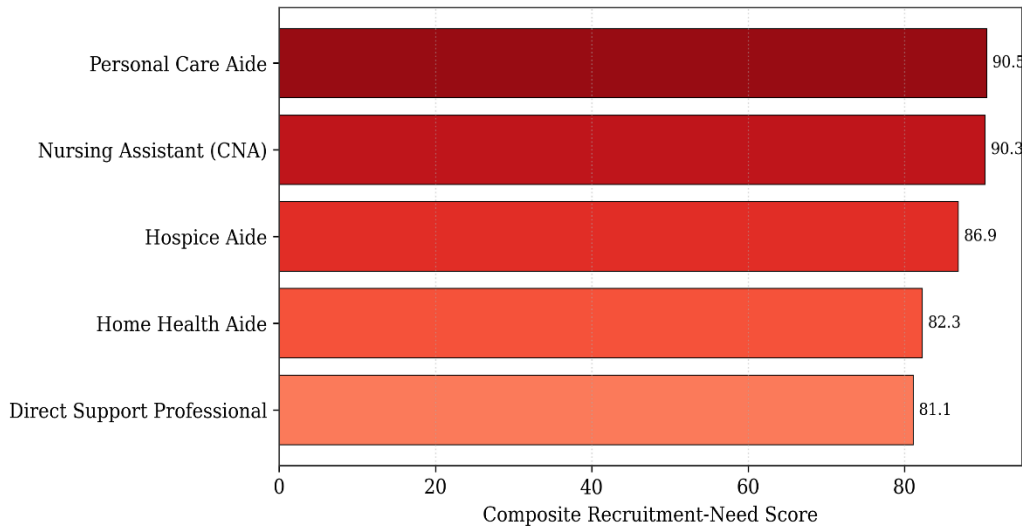


Fig 5: Composite Recruitment-Need Score by occupation. Higher scores indicate greater immediate recruitment pressure. Geographic level findings revealed the highest mean turnover intent in Maryland (4.31), Tennessee (4.25), Pennsylvania (3.65), and Texas (3.63) and relatively lower mean turnover intent in states with more established Medicaid HCBS infrastructures. These state-level differentials should be interpreted with caution because of the relatively small sample sizes per state but are consistent with the hypothesis that state-level workforce infrastructures matter in material ways to retention outcomes.

NLP ANALYSIS OF OPEN-ENDED RESPONSES

After going through the NLP pipeline and removing stop-words, the open-ended responses provided a corpus of around 2,900 tokens with the top ten most frequent themes reported in Table 4 and Figure 4. The most frequently expressed concern of the sector is under-investment in structured skill development, which is reflected in the dominance of the token ‘training’ (frequency 136). What is interesting is that the themes extracted from the NLP align substantively with the quantitative skill-gap results – telehealth, technology, and dementia care are all highly ranked in the quantitative results (Table 2) and also highly represented in the frequency distribution of narrative concerns (Table 4) – thus offering a form of within-study triangulation between the structured data stream (quantitative) and the unstructured data stream (NLP).

Rank	Theme (top n-gram)	Freq.	Interpretation
1	training / need more training	136	Under-investment in structured training
2	telehealth / training technology	32	Emergent tech-adjacent skill gap
3	clients complex medical needs	31	Rising client acuity beyond training
4	dementia care behavior mgmt	28	Cognitive-care competency gap
5	documentation ehr systems	27	EHR usability barrier
6	low pay burnout	25	Compensation-driven attrition
7	transportation long commutes	23	Non-wage cost burden
8	career advancement pathway	22	Absent progression structure
9	PPE infection control	20	Post-pandemic safety residual
10	mental health support death	18	Emotional labor unaddressed

Table 4: NLP Analysis of Open-Ended Responses

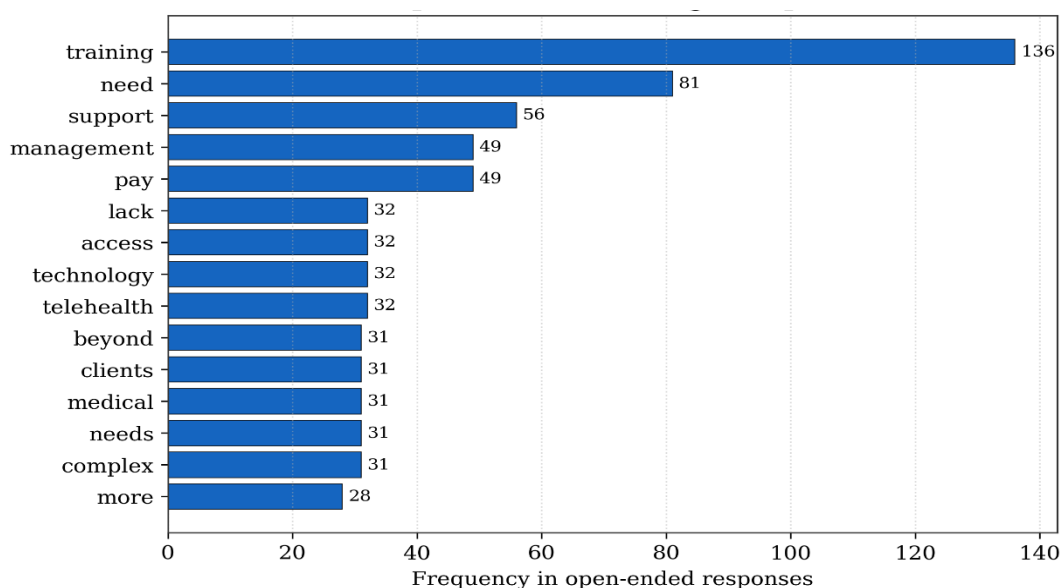


Fig 6: Top-15 Nlp-Extracted Terms From Open-Ended Challenge Responses.

The bigram analysis showed good semantic coupling between the tokens lack telehealth training, clients complex medical needs, documentation ehr systems, and career advancement pathway, with frequencies of over 20. The 'clients complex

medical needs beyond training' bigram cluster is analytically significant because it shows that a significant proportion of workers believe that their client caseload exceeds their clinical capacity. This is a direct implication on scope-of-practice, delegated-nursing, and structured-training policy at the state level.

TEN-YEAR DEMAND PROJECTION

Modeled demand by 2033, based on the BLS Occupational Employment Projections framework and consistent with the demographic aging assumption, indicates that the overall U.S. demand for all three roles (HHAs, PCAs, and CNAs) will increase from about 8.8 million in 2024 to about 11.7 million in 2033, a growth of roughly one-third over the decade. The year-by-year projection is shown in Table 5 and the trajectory in Figure 6. On the aggregate level, the amount of new hires needed over this time period is of the order of ten million workers under conservative estimates of turnover replacement demand.

Year	HHAs (m)	PCAs (m)	CNAs (m)	Total (m)
2024	3.7	3.55	1.52	8.77
2025	3.83	3.68	1.55	9.06
2026	3.97	3.82	1.57	9.36
2027	4.12	3.96	1.6	9.67
2028	4.27	4.1	1.62	9.99
2029	4.42	4.25	1.65	10.32
2030	4.58	4.41	1.67	10.65
2031	4.74	4.57	1.7	11
2032	4.91	4.73	1.72	11.36
2033	5.08	4.9	1.75	11.72

Table 5: Year-By-Year Projection

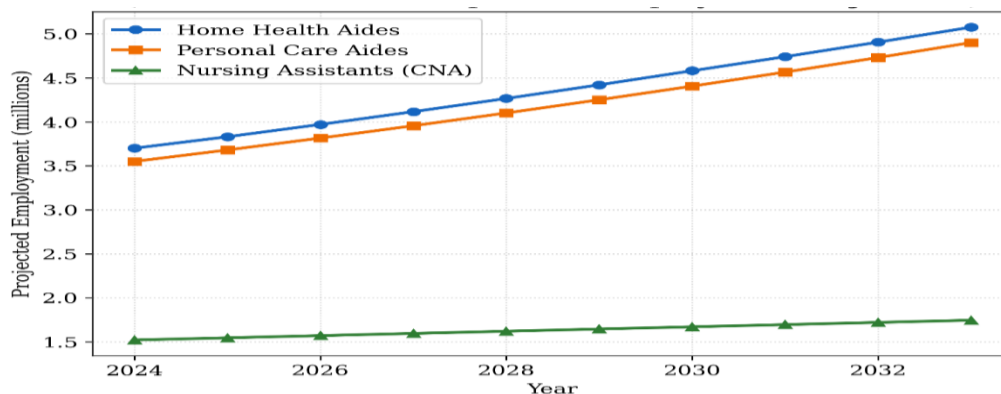


Fig 7: Modeled U.S. direct care workforce demand, 2024–2033.

DISCUSSION

PRINCIPAL FINDINGS

The goal of this study was to create and empirically implement an NLP-enhanced labor-market intelligence model for the U.S. direct care workforce. Three key points are revealed. First, the top self-reported skill gaps are clustered around

technology-adjacent areas (telehealth and EHR) and dementia care, suggesting a systematic lack of preparedness for the technology and clinically complex present. Second, the recruitment requirement for Personal Care Aides and Nursing Assistants is the greatest and the material content is the same, so dyadic occupational-stream planning is more appropriate than sequential prioritization. Third, the NLP extraction from the free text provided by the workers substantively aligns with the quantitative results from the skill-gap analysis, providing methodological validation of the mixed-methods approach and highlighting the promise of free text data as a leading indicator of emerging workforce concerns.

INTERPRETATION AND SITUATING WITHIN THE LITERATURE

The centrality of the technology and telehealth gaps constitutes both an expected and a novel finding. Expected, as the pandemic-era increase in the use of telehealth-enabled home care has been widely reported; novel, because the size of the gap – an SGI of 2.67 for telehealth is the highest observed in the study – implies that the direct care sector has been significantly less digitally fluent than the broader healthcare workforce. In the context of the CMS Home Health Value Based Purchasing model and other state Medicaid programs, this is of practical importance as these programs are increasingly incentivizing remote monitoring and virtual clinician involvement.

This finding that the need for PCA and CNA recruitment is near equal is at odds with the traditional policy framing, which has tended to focus on CNAs as the more severely under-supplied category because of their concentration in nursing facilities and their regulatory prominence. Our data indicate that the less-regulated, less-unionized PCA stream is facing similar recruitment pressures and may be deserving of similar policy attention.

The findings from quantitative and NLP-based studies of training, telehealth, and dementia care are a convergence of the two methods, which both add to the evidence base for using NLP as a complementary standing tool in workforce planning. The observed correspondence should not be interpreted as being trivial: the two data streams were produced by orthogonal cognitive tasks (numeric self-rating vs. free-text narration) and their agreement is thus a non-trivial construct validation.

POLICY AND PRACTICE IMPLICATIONS

There are four policy implications. First, Federal and state workforce authorities should focus on structured telehealth and EHR competency training for the direct care workforce, for example by expanding the Home Health Aide Certification framework and similar PCA credentialing pathways. Second, competency training for dementia should be formalized and based on programmes like the Alzheimer's Association CARES programme or state-level counterparts. Thirdly, wage stabilization efforts need to be complemented with non-monetary retention investments (supervisor training, career-pathway architecture, and transportation support) as our correlational evidence shows that wage is only a weak determinant of attrition. Fourth, workforce-planning agencies can use lightweight NLP tooling as a permanent addition to structured survey and administrative data flows, to capture emergent signals of the workforce's skills in real time.

STRENGTHS AND LIMITATIONS

The strong points of the study are the nationally distributed sample across 20 states and five occupational categories, the instrument that was developed based on the O*NET and PHI competency frameworks, and the mixed methods analysis

strategy, which combines structured and unstructured data. Limitations are: (i) the cross-sectional design of the study precludes causal inference; (ii) self-reported skill ratings are subject to well-documented social desirability and Dunning-Kruger effects; (iii) the moderate sample size (350 respondents) provides adequate power for the aggregate analyses reported here but does not allow for a more detailed stratification of the analyses at fine geographic granularity; and (iv) the lightweight NLP pipeline, while methodologically transparent, does not exhaustively leverage the analytical capabilities of state-of-the-art large language model architectures.

DIRECTIONS FOR FUTURE RESEARCH

Four future research directions are suggested. For one, longitudinal replication would enable skill-gap trajectories to be estimated and leading indicators of workforce inflection to be identified. Second, integration with employer side data – vacancy postings, EHR documentation and quality of care outcomes – would allow for causal identification of the impact of skill development on care outcomes. Third, the current NLP pipeline could be extended to include state-of-the-art transformer-based topic modelling and zero-shot classification, which would refine the extraction of the themes. Fourth, expansion to include the entire 50-state distribution and other occupational categories (such as adult day services staff and family caregivers with paid arrangements) would increase the external validity of the framework.

CONCLUSION

The U.S. direct care workforce is in a strategic inflection point. Together, demographic demand, the increasing technological integration of care delivery and chronic labor-market vulnerabilities present an analytical need to predict—not just measure after the fact—workforce shortages. This study has shown that NLP coupled with traditional survey analysis can provide a scalable and policy-relevant tool for identifying new demand, recruitment requirements, and skill gaps in this key occupational sector. The empirical results—including the size of telehealth and technology-EHR gaps, the nearly equal need for PCA and CNA recruitment, and the overlap of quantitative and NLP-derived signals about training as the top issue of concern for employers—should serve as actionable information in Federal, state, and employer workforce planning. Investing in the direct care workforce, with a focus on the labor market, is not only a labor-policy imperative, but a healthcare-quality and equity imperative for the decade ahead.

ACKNOWLEDGEMENTS

The authors are grateful to the 350 direct care staff who so generously gave their time and reflections for this study. The authors would like to thank the employer networks, worker-advocacy organizations, and community-based agencies that facilitated recruitment in 20 states. The opinions contained in this paper are the authors' own and do not necessarily reflect the views of the partner organizations.

FUNDING

No specific grant was received from any public, commercial, or not-for-profit funding agency for this research.

CONFLICT OF INTEREST STATEMENT

The authors report no conflicts of interest or personal relationships that could have been construed to influence the work reported in this paper.

DATA AVAILABILITY

The data used in this study is provided in a supplementary spreadsheet (DCW_Response_Sheet.xlsx) in the de-identified form. A copy of the survey instrument is included as an attachment (DCW_Questionnaire.docx). Other analytical outputs were derived and are available upon reasonable request from the corresponding author.

REFERENCES

- [1] A. Alibasic, M. C. E. Simsekler, T. Kurfess, W. L. Woon, and M. A. Omar, "Utilizing data science techniques to analyze skill and demand changes in healthcare occupations: Case study on USA and UAE healthcare sector," *Soft Computing*, vol. 24, no. 7, pp. 4959–4976, 2020.
- [2] C. M. Mason, H. Chen, D. Evans, and G. Walker, "Illustrating the application of a skills taxonomy, machine learning and online data to inform career and training decisions," *The International Journal of Information and Learning Technology*, vol. 40, no. 4, pp. 353–371, 2023.
- [3] I. Khaouja, I. Kassou, and M. Ghogho, "A survey on skill identification from online job ads," *IEEE Access*, vol. 9, pp. 118134–118153, 2021.
- [4] H. Krasna, K. Czabanowska, A. Beck, L. F. Cushman, and J. P. Leider, "Labour market competition for public health graduates in the United States: A comparison of workforce taxonomies with job postings before and during the COVID-19 pandemic," *The International Journal of Health Planning and Management*, vol. 36, no. S1, pp. 151–167, 2021.
- [5] I. Hazarika, "Artificial intelligence: Opportunities and implications for the health workforce," *International Health*, vol. 12, no. 4, pp. 241–245, 2020.
- [6] G. Tomblin Murphy, S. Birch, A. MacKenzie, S. Bradish, and A. Elliott Rose, "A synthesis of recent analyses of human resources for health requirements and labour market dynamics in high-income OECD countries," *Human Resources for Health*, vol. 14, no. 1, Art. no. 59, 2016.
- [7] A. Verma, K. Lamsal, and P. Verma, "An investigation of skill requirements in artificial intelligence and machine learning job advertisements," *Industry and Higher Education*, vol. 36, no. 1, pp. 63–73, 2022.
- [8] L. Benhayoun and D. Lang, "Does higher education properly prepare graduates for the growing artificial intelligence market? Gaps' identification using text mining," *Human Systems Management*, vol. 40, no. 5, pp. 639–651, 2021.
- [9] M. Balasubramanian, A. Hasan, S. Ganbavale, A. Alolayah, and J. Gallagher, "Planning the future oral health workforce: A rapid review of supply, demand and need models, data sources and skill mix considerations," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, Art. no. 2891, 2021.
- [10] J. X. Liu, Y. Goryakin, A. Maeda, T. Bruckner, and R. Scheffler, "Global health workforce labor market projections for 2030," *Human Resources for Health*, vol. 15, no. 1, Art. no. 11, 2017.

- [11] A. Jetha, A. Shamaee, S. Bonaccio, M. A. Gignac, L. B. Tucker, E. Tompa, *et al.*, “Fragmentation in the future of work: A horizon scan examining the impact of the changing nature of work on workers experiencing vulnerability,” *American Journal of Industrial Medicine*, vol. 64, no. 8, pp. 649–666, 2021.
- [12] I. O. Evans-Uzosike and C. G. Okatta, “Talent management in the age of gig economy and remote work and AI,” *Shodhshauryam, International Scientific Refereed Research Journal*, vol. 6, no. 4, pp. 147–170, 2023.
- [13] A. S. George, “Future economic implications of artificial intelligence,” *Partners Universal International Research Journal*, vol. 2, no. 3, pp. 20–39, 2023.
- [14] L. D. Caesar, “Emerging dynamics of training, recruiting and retaining a sustainable maritime workforce: A skill resilience framework,” *Sustainability*, vol. 16, no. 1, Art. no. 239, 2023.
- [15] A. Habiyaremye, T. Habanabakize, and C. Nwosu, “Bridging the labour market skills gap to tackle youth unemployment in South Africa,” *The Economic and Labour Relations Review*, vol. 33, no. 4, pp. 786–805, 2022.
- [16] R. I. Stone, “The migrant direct care workforce: An international perspective,” *Generations*, vol. 40, no. 1, pp. 99–105, 2016.
- [17] S. Akter, “Exploring the role of artificial intelligence in food waste reduction: Implications for public health, social behavior, and sustainable communities,” *Journal of Computational Systems & Engineering Insights*, vol. 1, no. 1, pp. 25–36, 2023.
- [18] I. A. Badhan, M. N. Hasnain, and M. H. Rahman, “Enhancing operational efficiency: A comprehensive analysis of machine learning integration in industrial automation,” *Journal of Business Insight and Innovation*, vol. 1, no. 2, pp. 61–77, 2022.
- [19] S. Bhardwaj, R. Srivastava, T. Mageto, M. Chaudhari, A. Kumar, J. Sultana, *et al.*, “Bimetallic Co–Fe sulfide and phosphide as efficient electrode materials for overall water splitting and supercapacitor,” *Discover Nano*, vol. 18, no. 1, Art. no. 59, 2023.
- [20] A. Gupta, C. A. Allison, M. Chaudhari, P. Zalavadiya, F. M. de Souza, R. K. Gupta, and T. Dawsey, “A facile approach to recycling used facemasks for high-performance energy-storage devices,” in *Specialty Polymers*, pp. 427–443, CRC Press, 2023.