

Detecting The Invisible: AI-Based Identification of Dementia Caregiver Burden



#10416

Nancy A. Hodgson, PhD¹, Ari Klein, PhD², Emily Summerhayes¹, Li Shen, PhD²
¹Hodgson Research, Biobehavioral Health Sciences Department, School of Nursing, University of Pennsylvania
²Shen lab, Biostatistics, Epidemiology and Informatics Department, School of Medicine, University of Pennsylvania

Effective implementation of an AI-assisted burden caregiver detection platform requires thoughtful human-AI collaboration, which if done successfully and ethically can shape how the field approaches automated psychosocial assessment in dementia.

INTRODUCTION	METHODS	RESULTS	CONCLUSIONS	REFERENCES							
55 million people worldwide live with dementia [1]. Each year, US family caregivers (CG) provide over 19 billion hours of unpaid care [1]. Dementia CGs experience twice the emotional, physical, & financial burden & an increased risk of depression compared to non-AD/ABRD caregivers [2]. CG burden exacerbates situations leading to • <i>Poorer health outcomes,</i> • <i>Earlier nursing home placement,</i> [3] • <i>\$600+ billion in annual costs</i> [4]. Early burden detection is challenging due to its multidimensional, subjective nature. This study compared 3 machine learning (ML) approaches for detecting CG burden in clinical conversations to inform development of a real-time screening platform for interdisciplinary team (IDT) meetings. This is a secondary research study using data collected in a clinical trial (NCT04165213).	We used 126 CG transcripts to develop and evaluate 3 ML approaches to burden detection. <table><tr><th>Approaches</th><th>Findings</th></tr><tr><td>1. Semi-Supervised ML learning BERT models trained via transcripts classified by CG Zarit Burden Index (ZBI) scores [5].</td><td>F1-score: 0.58 (precision = 0.48, recall = 0.77) <i>Minimally exceeded the baseline (F1=0.57) from classifying all caregivers as burdened.</i></td></tr><tr><td>2. Iterative Large Language Model Prompting Trained HIPAA-compliant GPT system with burden definition and study purpose.</td><td>F1-score: 0.60 (precision = 0.43, recall = 1) <i>Marginally higher precision than method #1.</i></td></tr><tr><td>3. Fully supervised ML learning Via transcripts coded at sentence level (n=126) • High burden transcripts (n=57) • “Low” burden transcripts (n=69) </td><td>F1-score: 0.25 (precision = 0.19, recall = 0.35) <i>Similar performance despite expert coding. Coding reveals dataset class imbalance.</i> • 1.1% of sentences coded as burden </td></tr></table> Literal statements were detected but missed contextual subtext clinicians would identify. Linda Chen (DOB: 02/18/1945) - 90-Day Care Plan Review Primary Caregiver: Grace Chen (Daughter) Next of Kin Diagnosis: Alzheimer's Disease, Stage 4 Care Level: Moderate Assistance Download PDF IDT IN SESSION Medical Review Functional Status Caregiver Support Care Plan Updates Discharge Planning Burden Risk Trend Analysis Last 6 Months Jun 15, 2025 Moderate Risk Voice analysis complete Aug 22, 2025 High Risk Transcription analyzed Today - Dec 24, 2025 High Risk Current Assessment Caregiver Mental Health Concern HIGH PRIORITY Offcode analysis indicates elevated burden risk (89% Confidence Score) in primary caregiver Grace Chen across recent visits. Team intervention recommended. IDT Members Present SW Maria Rodriguez MD Dr. James Wilson RN Sarah Thompson CH Rev. Michael Park PT Lisa Chang Meeting Notes Burden minimization Interpretation	Approaches	Findings	1. Semi-Supervised ML learning BERT models trained via transcripts classified by CG Zarit Burden Index (ZBI) scores [5].	F1-score: 0.58 (precision = 0.48, recall = 0.77) <i>Minimally exceeded the baseline (F1=0.57) from classifying all caregivers as burdened.</i>	2. Iterative Large Language Model Prompting Trained HIPAA-compliant GPT system with burden definition and study purpose.	F1-score: 0.60 (precision = 0.43, recall = 1) <i>Marginally higher precision than method #1.</i>	3. Fully supervised ML learning Via transcripts coded at sentence level (n=126) • High burden transcripts (n=57) • “Low” burden transcripts (n=69)	F1-score: 0.25 (precision = 0.19, recall = 0.35) <i>Similar performance despite expert coding. Coding reveals dataset class imbalance.</i> • 1.1% of sentences coded as burden	• Current AI alone cannot distinguish caregiver burnout from burden, but the development of a clinical decision support platform for family care consults is needed. • We are designing a system that flags potential burden indicators during IDT assessments requiring human clinical review – not just a diagnosis. • This approach combines AI screening sensitivity & valuable clinician contextual judgment. • This work establishes feasibility benchmarks for AI-assisted burden screening. • Future research must pursue multimodal integration of conversational patterns, validated instruments, temporal trends, and relationship dynamics. • Next steps include, testing more loss functions and computing balanced accuracy	1. 2026 Alzheimer's disease facts and figures. (2026). <i>Alzheimer's & Dementia</i>, 22(4), e71345. https://doi.org/10.1002/alz.71345 2. Collins, R. N., & Kishita, N. (2020). Prevalence of depression and burden among informal caregivers of people with dementia: a meta-analysis. <i>Ageing & Society</i>, 40(11), 2355-2392. 3. AARP, National Alliance for Caregiving. Caregiving in the United States 2020. Washington (DC): AARP; 2020. 4. Houser, A., Caldera, S., Flinn, B., & Choula, R. (2026). <i>Valuing the Invaluable 2026</i>. AARP Public Policy Institute. https://doi.org/10.26419/ppi.00402.001 5. Zarit SH, Reever KE, Bach-Peterson J. Relatives of the impaired elderly: correlates of feelings of burden. <i>Gerontologist</i> 1980;20:649-55.
Approaches	Findings										
1. Semi-Supervised ML learning BERT models trained via transcripts classified by CG Zarit Burden Index (ZBI) scores [5].	F1-score: 0.58 (precision = 0.48, recall = 0.77) <i>Minimally exceeded the baseline (F1=0.57) from classifying all caregivers as burdened.</i>										
2. Iterative Large Language Model Prompting Trained HIPAA-compliant GPT system with burden definition and study purpose.	F1-score: 0.60 (precision = 0.43, recall = 1) <i>Marginally higher precision than method #1.</i>										
3. Fully supervised ML learning Via transcripts coded at sentence level (n=126) • High burden transcripts (n=57) • “Low” burden transcripts (n=69)	F1-score: 0.25 (precision = 0.19, recall = 0.35) <i>Similar performance despite expert coding. Coding reveals dataset class imbalance.</i> • 1.1% of sentences coded as burden										
			CONTACT Nancy A. Hodgson hodgsonn@nursing.upenn.edu	ACKNOWLEDGEMENTS Transcripts were coded for burden by experts Hannah Cho, PhD, Freya Yang, Hanna Ramsburg, PhD and Emma Max and machine learning methods were tested by Mengyuan Kan, PhD, Shu Yang, PhD, Weiqing He, BS, MS, Boning Tong, BS, ME and Hongzhuo Chen, BS. Research reported in this publication/ presentation was supported by the National Institute On Aging of the National Institutes of Health under Award Number P30AG073105. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.							