Causal Models with Tiny Data: The Case of Rural People Living with Dementia*

Ranveer Singh[1**†], Saurabh Mathur[2**], Kavimayil P. Komarasamy[1], Ameet Soni[3], Cliff Whetung[4], Wayne Warry[4], Kristen Jacklin[4], Melissa Blind[4], and Sriraam Natarajan[1]

[1] The University of Texas at Dallas, Richardson, TX 75080, USA
[2] Technische Universität Darmstadt, Darmstadt, Germany
[3] Swarthmore College, Swarthmore, PA 19081, USA
[4] University of Minnesota Medical School, Duluth Campus, Duluth, MN 55812, USA

Abstract. Causal modeling for specialized populations like non-urban dementia patients is critical for developing targeted interventions, yet reliable causal models are unavailable. Moreover, the inherent scarcity of data makes automatic causal discovery challenging. We study several approaches for building causal graphs under this data-scarce setting: expert elicitation, Large Language Model (LLM) generation, a data-driven method, and a hybrid method that refines expert and LLM-elicited graphs using data. Through direct structural comparison, our analysis reveals areas of agreement between the graphs, but also contradictions in different tendencies impacting the directionality and inclusion of socioeconomic and clinical factors. Additionally, we show that tiny data sets can be used to empirically validate conditional independencies encoded in these candidate causal graphs, concluding that causal modeling for specialized populations requires reconciling expert and LLM-generated causal discovery with tiny datasets.

Keywords: Life-Space Assessment. Causal Discovery. Data Scarce Domains. Large Language Models.

1 Causal Modeling for Life-Space Assessment

Dementia is a progressive decline in cognitive functions. It can negatively impact the medical, social, functional, and psychological well-being of a patient, ranging from memory loss to depression to social withdrawal [2]. As the average life expectancy and aging population increase, dementia diagnoses will rise as well. Therefore, it is necessary to conduct more comprehensive studies to understand dementia and improve clinical outcomes. Persons Living with Dementia (PLWD) can be studied by assessing their life-space mobility. It captures a person's physical and social environment, movement, and daily activities [1,8]. A decrease in

* We acknowledge support from NIH awards R01NS133142 and 1R21AG072566.
** Equal Contribution. † Corresponding Author: ranveer.singh@utdallas.edu

life-space is associated with various measures of declining health in older adults, including cognitive decline [9].

Life-space mobility is primarily measured through a written survey called the Life-Space Assessment (LSA), which quantifies an older adult's geographic area in environmental zones and captures the rate at which they travel throughout the day [7]. For PLWD, LSA can be a useful tool to chart a patient's progress or mental decline [9]. However, LSAs have primarily focused on urban populations, limiting the applicability of LSAs to rural or Indigenous settings, where dementia risk is higher [10]. Therefore, LSA should be adapted for these settings, for example, by redefining environmental zones for rural and Indigenous contexts.

While life-space is an effective indicator of quality of life, developing interventions to improve outcomes requires causally modeling its relationship with relevant demographic and environmental variables. Causal models are often expressed as graphs, consisting of nodes representing the variables of interest, and directed edges between nodes representing cause-effect relationships [5]. Causal graphs concisely represent qualitative causal knowledge about a domain. In the absence of such domain knowledge, they might be learned from data.

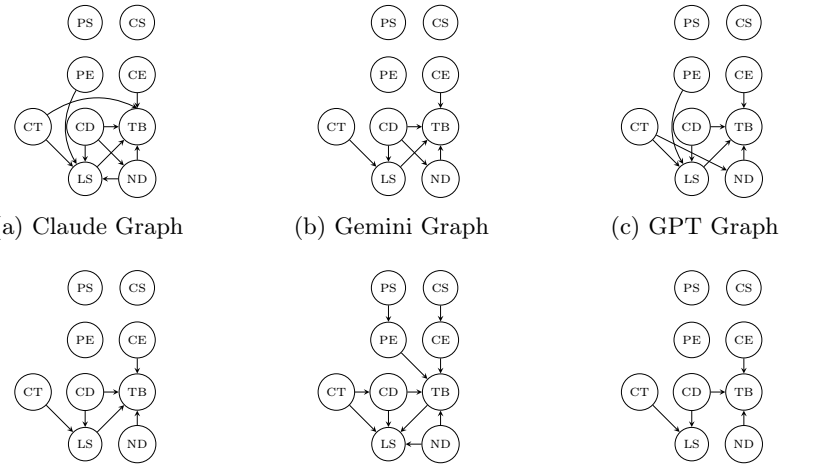
Data-driven causal discovery methods exploit patterns in large amounts of high-dimensional data, along with assumptions about the underlying causal dynamics, to induce causal graphs. However, the paucity of data about understudied populations makes data-driven causal discovery challenging [3]. In low-data settings, a hybrid approach is adopted; smaller datasets are used to refine graphs derived from incomplete domain knowledge obtained from either domain experts or from LLMs [4], owing to their ability to capture correlations over causal facts from their vast training corpora [11]. Each of these approaches depends on good initial causal graphs, which are difficult to come by in understudied domains.

We aim to analyze the plausibility and mutual compatibility of the graphs obtained by these methods in terms of the empirical validity of their inferences. Specifically, we empirically evaluate the statistical independence statements inferred from each graph. Alongside qualitatively analyzing the similarities and differences across the graphs, the analysis furthers the goal of creating more comprehensive causal graphs by combining the expert knowledge, approximate knowledge from LLMs, and tiny data.

2 Qualitative and Quantitative Evaluation

We answer the following research questions:
(Q1) Is there any consensus in causal relations identified by experts and LLMs?
(Q2) Are the graphs obtained from experts and LLMs compatible with the statistical patterns present in the tiny dataset?

Dataset. To answer these questions, we constructed a dataset from LSAs of **20** patients conducted in **rural** and **Indigenous** communities in northern Minnesota, USA, and Ontario, Canada. We extracted nine boolean features from the LSA, demographic data for patients and caregivers, and the daily diaries completed by the caregiver over the course of a month. The features include the



(a) Claude Graph (b) Gemini Graph (c) GPT Graph

(d) LLMs Consensus (e) Expert Graph (f) Overall Consensus

Fig. 1: The causal graphs for the Life-space Domain. (**a-c**) shows the causal graphs provided by the different LLMs, (**d**) shows the consensus (common causal relationships) across the different LLMs, (**e**) shows the expert causal graph, and (**f**) shows the consensus between the LLMs and the experts.

life-space score (LS), sex of the patient (PS) and caregiver (CS), education of the patient (PE) and caregiver (CE), and their community type (CT). We also consider the number of non-routine days (ND) and challenging days (CD) for the patient, as well as the total burden on the caregiver (TB). Continuous variables were binarized using expert-provided or sample mean-based thresholds[1]

Methods. For these variables, we consider 4 types of causal graph construction methods: Expert elicitation, LLM-generation, data-driven discovery, and Hybrid subtractive refinement. We obtain candidate causal graphs from multiple experts' consensus and 3 LLMs – Claude (4.5 Sonnet), Gemini (3 Thinking), and ChatGPT (GPT 5.2). To construct the graphs using the LLMs, they were prompted five times each with the variables and their description, and asked to come up with causal relationships as well as their corresponding weight[1]. The graphs were then pooled to form a single graph by adding causal edges in decreasing order of their average weight, excluding any edges introducing a cycle or having an average weight below a threshold. Additionally, we considered a hybrid subtractive refinement method [4] that refines the above graphs by deleting edges to optimize the Minimal Description Length score. Finally, we implemented the purely data-based baseline using the Fast Causal Inference (FCI) algorithm [6].

Results. To answer **(Q1)**, we compare the graphs structurally using consensus sets (relations common to all graphs) and conflict sets (disagreements between candidate graphs). Figure 1 shows that the **LLMs agree on six direct causal relationships with each other**: Non-routine days (ND), community type (CT), education of caregiver (CE), and life-space score (LS) are direct

[1] The thresholds, prompts, and their responses are provided in the supplementary.

Model	Total	Precision	FPR
Expert	113	0.73	0.53
GPT	45	0.67	0.48
Claude	45	0.73	0.39
Gemini	23	0.70	0.33
LLM Consesus	31	0.68	0.48
Overall Consesus	39	0.72	0.52

Table 1: **Empirically validating candidate graphs.** For each model, we summarize (i) Total: number of conditional independencies (CI) implied by graph structure (via d-separation), (ii) Precision: fraction of CI statements entailed by the model that were compatible with the data, and (iii) False positive rate (FPR): fraction of CI statements incompatible with the data that were entailed by the model. CIs were empirically tested using the G-test at a significance threshold of $\alpha = 0.05$ and a conditioning set of size at most one.

causes of total burden (TB); challenging days (CD), and community type (CT) are direct causes of life-space score (LS). While experts agree with the LLMs on five of these six relationships, *they invert the final edge*, placing the causal direction from TB to LS. This disparity suggests a fundamental difference in modeling perspectives: *the experts appear to be more interested in the patient's outcomes*, treating the life-space score (LS) as the final sink node, while the LLMs consistently identify the caregiver's total burden (TB) as the ultimate sink variable. This disparity might also point to the bidirectional nature of the relationship, requiring a disaggregation across time. Further, the LLMs exclude socio-economic factors such as sex and education, which the experts identify as significant causal drivers.

To answer **(Q2)**, we examine the results of the refinement procedure. The number of edges deleted during this process serves as a measure of compatibility between the causal models and the dataset. Refinement eliminated nearly all edges, leaving fewer than three in the final graphs. Further, while the data-driven baseline did not identify any causal edges, it did identify 3 undirected associations: caregiver's sex with patient's sex, caregiver's education with community type, and challenging days with total burden. These results indicate high incompatibility between the tiny dataset and the LLMs' and the experts' graphs.

Next, we analyze the nature of this incompatibility by empirically evaluating the conditional independence (CI) relations entailed by each graph. We treat each network as a model that predicts whether variables X and Y are independent given a variable Z. We evaluate these predictions against data-driven labels using the G-test. Since incorrectly assuming independence is more detrimental to causal inference than failing to identify it, we quantify each graph's divergence from data-driven CIs using False Positive Rate (FPR) and Precision metrics. Table 1 summarizes the results. Clearly, the expert graph entails more CIs than the LLM graphs. As a result, it has a higher or equal precision, but a higher FPR than the LLM graphs. On the other hand, the LLM graphs entail fewer independencies overall, but still entail independencies incompatible with data.

Of these, Claude's graph has a lower FPR and the same precision as the expert graph, but entails less than half the number of independencies.

Neither the LLM-generated graphs nor the expert-constructed graphs are compatible with the data. However, they are incompatible in different ways, illustrating the differences between expert and LLM-based causal models.

3 Discussion

We considered the problem of causally modeling the life-space mobility of people living with dementia in non-urban environments. The domain's understudied nature and the resulting data paucity make this a challenging problem. We considered two additional sources of causal knowledge: domain experts and LLMs. While data-driven causal discovery and refinement of expert and LLM-elicited causal graphs proved unsuccessful, the tiny dataset allowed us to compare the two types of graphs. The graph structures reveal distinct underlying assumptions. Experts prioritized patient-centric outcomes and integrated socio-economic factors like sex and education into the causal structure. In contrast, the LLMs focused on the total burden and ignored the socio-economic factors. Our analysis of the differences among sources of causal knowledge – domain experts, LLMs, and tiny datasets – provides a foundation for hybrid causal models. Understanding the interplay between expert-led and LLM-driven discovery remains essential for developing robust clinical decision-support systems.

References

1. Baker, P.S., Bodner, E.V., Allman, R.M.: Measuring life-space mobility in community-dwelling older adults. J Am Geriatr Soc. **51**(11), 1610–1614 (2003)
2. Geldmacher, D.S., Whitehouse, P.J.: Evaluation of dementia. N Engl J Med. **335**(5), 330–336 (1996)
3. Hasan, U., Hossain, E., Gani, M.O.: A survey on causal discovery methods for iid and time series data. TMLR (2023)
4. Mathur, S., Singh, R., et al.: Llm-guided causal bayesian network construction for pediatric patients on ecmo. In: AIME. pp. 255–260. Springer (2025)
5. Pearl, J.: Causality. Cambridge university press (2009)
6. Spirtes, P.: An anytime algorithm for causal inference. In: AISTATS. pp. 278–285. PMLR (2001)
7. Stalvey, B.T., Owsley, C., et al.: The life space questionnaire: a measure of the extent of mobility of older adults. J Appl Gerontol **18**(4) (1999)
8. Taylor, J.K., Buchan, I.E., Van Der Veer, S.N.: Assessing life-space mobility for a more holistic view on wellbeing in geriatric research and clinical practice. Aging Clin Exp Res **31**(4), 439–445 (2019)
9. Ullrich, P., Eckert, T., et al.: Life-space mobility in older persons with cognitive impairment after discharge from geriatric rehabilitation. Arch Gerontol Geriatr **81**, 192–200 (2019)
10. Xie, Z., Hu, J., et al.: Rural-urban differences in modifiable dementia risk factors among us populations aged 45 years or older. J Alzheimers Dis Rep **9** (2025)
11. Zečević, M., Willig, M., Dhami, D.S., Kersting, K.: Causal parrots: Large language models may talk causality but are not causal. TMLR (2023)