arXiv:2602.02371v1 [cs.LG] 2 Feb 2026

C-kNN-LSH：一种用于序贯反事实推断的最近邻算法

Jing Wang [1] **Jie Shen** [2] **Qiaomin Xie** [3] **Jeremy C Weiss** [1]

摘要

从纵向轨迹中估计因果效应，对于理解复杂疾病的进展并优化临床决策至关重要，例如合并症和长新冠康复。我们提出*C-kNN–LSH*，这是一种用于序列因果推断的最近邻框架，专为应对此类高维、存在混杂的情形而设计。通过利用局部敏感哈希，我们能够高效识别协变量历史相似的“临床双胞胎”，从而在不断变化的疾病状态下局部估计条件治疗效应。为减轻不规则采样和患者康复特征变化带来的偏差，我们将邻域估计量与双重稳健校正相结合。理论分析保证了该估计量的一致性，并证明其对干扰误差具有二阶稳健性。在包含 13,511 名参与者的真实世界长新冠队列上进行评估后，*C-kNN-LSH* 在捕捉康复异质性和估计策略价值方面，表现优于现有基线方法。

1. 引言

长新冠，或称SARS-CoV-2感染急性后遗症（PASC），是一种慢性疾病，其特征是急性感染后数月乃至数年持续出现且时有波动的症状（Wang et al.，2026年；2025年）。临床严重程度通常通过多项症状指标进行概括，例如PASC评分；这些指标会随时间不规则变化，并且在患者之间存在显著异质性。一个核心科学问题是，医疗干预（尤其是新冠疫苗接种）如何影响长新冠严重程度的变化轨迹，以及这种影响如何取决于疫苗接种时间、患者病史和潜在生理状态。

[1]美国国家医学图书馆 [2]史蒂文斯理工学院 [3]威斯康星大学麦迪逊分校。通讯作者：Jing Wang<jing.wang20@nih.gov>。

预印本。*2026*年*2*月*3*日。

具体而言，两名患者早期的PASC变化轨迹和合并症情况相似。一名患者在症状出现后不久接种疫苗，随后症状严重程度逐渐降低；另一名患者则在晚得多的时候接种同样剂量，症状没有改善，甚至出现短暂恶化。要理解这类现象，需要对反事实纵向变化轨迹进行推理：如果在不同时间接种疫苗，或完全不接种，患者未来的严重程度会如何变化。这个问题本质上具有序贯性、时间依赖性和异质性，无法简化为静态的平均治疗效应。

形式上，我们观测一个随时间跟踪的 N 个体队列。对于个体 i 在时间 t，令$\mathcal{H}_{i,\,t}$ 表示观测历史，其中包括人口统计信息、合并症和高频可穿戴设备测量数据。表示在时间给药的疫苗剂次。表示观测到的 PASC 严重程度结果。令$Y_{i,\,t}(a)$表示在时间 t接受 a 治疗时的潜在结果。我们的目标是估计反事实条件均值$\theta_{i,\,t}(a)$，以及更一般地，估计不同疫苗接种方案下的反事实严重程度轨迹。

经典因果推断方法通常通过结构因果模型（SCM）和有向无环图来表述这类问题。尽管这些框架功能强大，但它们要求采用强参数假设，或明确恢复因果结构；而在具有高维、非规则采样纵向数据的生物医学场景中，这些任务尤其具有挑战性。近期研究指出，在许多应用领域，回答定义明确的因果问题并不需要明确恢复因果图（Mlodozeniec 等人）。相反，只要纵向数据生成过程的概率模型具有足够的表达能力，并结合正确的时间顺序和条件独立性假设，就足以支持有效的反事实推理和干预推理。这一观点在医学领域尤为重要，因为干预受时间约束，协变量观测丰富，而且因果问题关注的是轨迹对比，而非抽象的图识别。

在本研究中，我们采用这一观点。我们不试图学习支配长新冠进程的因果图，而是寻求

通过对条件分布 $p(Y_{i,t}|H_{i,t},A_{i,t})$ 建模，在序贯一致性、正值性和可忽略性等标准假设下，估计反事实结果轨迹。我们关注的是，概率建模与严谨的非参数推断相结合，如何足以在复杂的纵向场景中进行因果推理。

挑战 尽管这一概念上的简化降低了问题的复杂性，但在实践中估计 $\theta_{i,t}(a)$ 仍然很困难。以下三项挑战在长新冠研究中尤为突出：

- 高维病史。患者病史包含数百个临床变量和密集的可穿戴设备时间序列，使得直接进行非参数匹配变得不可行。例如，有 500 页康复参与者的资料。

- 时变混杂。疫苗接种时间取决于不断变化的健康状况，而健康状况本身也能预测未来结局，例如药物使用和疾病进展。

- 异质性与状态转变。从经验来看，患者呈现出不同的严重程度状态，包括长期随访期间病情稳定且严重程度较低、对疫苗有反应的病程轨迹，以及持续处于高严重程度状态。

统计学领域近期的理论研究表明，当病史具有低维潜在结构时，最近邻匹配能够在序列实验中产生一致的反事实估计量。然而，这些研究通常假设能够获取充分的潜在状态，或是在状态维度适中的场景下进行，而现代生物医学数据集并不满足这些假设。

因此，我们提出一种新的算法框架，通过整合以下三个组成部分，将最近邻反事实推断扩展到高维纵向数据：

- 潜在历史压缩。变分自编码器（VAE）学习患者历史的低维潜在表示。该表示经过训练，在保留与结局相关信息的同时，作为治疗分配的近似充分统计量。

- 局部非参数推断。在学习到的潜在空间中，通过 K 近邻匹配估计反事实结局，从而实现个体化的轨迹级推断。

- 双重稳健去偏。我们结合结果回归与倾向得分建模，得到的估计量在任一干扰成分设定正确时均具有一致性。

关键在于，VAE 并非用作黑箱预测器。相反，它充当概率状态抽象，使原本基于低维假设开发的经典非参数估计量能够应用于复杂的纵向环境。我们证明，在近似潜在充分性的条件下，这种组合可得到一致的反事实估计量，而现有理论研究尚未涵盖这一情形。我们将该方法应用于一个大型纵向长新冠队列，该队列包含 13,511 名以上参与者，随访时间最长达 700 天。除临床协变量和疫苗接种记录外，该数据集还包含连续的可穿戴设备测量数据，用于记录睡眠、活动和生理信号。这些测量数据可作为潜在健康状态的高频代理指标，减少未观测混杂，并增强条件反事实比较的可信度。

2. 问题设定

我们考虑一个包含 N个体的纵向观察性设定，其中每个个体 $i \in [N]$ 都在离散时间范围 $t \in [T]$内接受追踪。在每个时间步 t，我们记录特征向量$X_{i,\ t} \in \mathcal{X} \subseteq \mathbb{R}^d$ 该向量由静态合并症和源自可穿戴传感器的高维生理时间序列构成。令$A_{i,\ t} \in \mathcal{A}$ 表示干预（例如疫苗剂量接种），令$Y_{i,\ t} \in \mathbb{R}$表示观测到的临床结局。

我们将截至时间 t 治疗分配时刻的观测历史定义为滤过：

$$\mathcal{H}_{i,t} \doteq \{X_{i,1:t}, A_{i,1:t-1}, Y_{i,1:t-1}\} \tag{1}$$

遵循潜在结果框架，对于每个行动 $a \in \mathcal{A}$，令 $Y_{i,t}(a)$ 表示反事实结果。我们的主要估计是条件反事实均值：

$$\theta_{i,t}(a) = \mathbb{E}[Y_{i,t}(a) \mid \mathcal{H}_{i,t}] \tag{2}$$

我们假设序贯正值性、可忽略性与一致性等标准纵向因果条件成立 $\forall t, i$。

假设 2.1. （一致性）。若 $A_{i,t} = a$，则$Y_{i,t} = Y_{i,t}(a)$。

它断言存在从潜在结果空间到观测数据流形的确定性映射。

假设 2.2. （序列可忽略性）。对于所有 $a \in \mathcal{A}$，$Y_{i,t}(a) \perp\!\!\!\perp A_{i,t} \mid H_{i,t}$。

该假设断言，$H_{i,t}$ 中的观测协变量足以阻断动作 $A_{i,t}$ 与未来潜在结果之间的所有“后门”路径。也就是说，在有向无环图（DAG）中，不存在未观测变量 U 使得 $A_{i,t} \leftarrow U \rightarrow Y_{i,t}$。

假设 2.3.（正值性）。存在 $\delta > 0$ 使得

$$\delta \le P(A_{i,t} = a \mid H_{i,t}) \le 1 - \delta \quad \text{for all } a, i, t. \tag{3}$$

给定上述假设，我们令 Θ 表示支配潜在结果的函数映射。本研究考虑上下文老虎机，$\theta_{i,\,t}(a)$可能表示个体 i 在时间 t 采取动作 a时的潜在结果。目标是估计完整的因果参数集合，使得

$$\Theta = \{\theta_{i,t}(a) : i \le N, t \le T_i, a \in \mathcal{A}\}, \tag{4}$$

从而能够进行个体化反事实轨迹推断及后续的干预对比。

2.1. 大语言模型增强的历史压缩

直接根据原始历史记录 $H_{i,\,t}$ 估计因果参数 $\theta_{i,\,t}(a)$会受到维度灾难的阻碍。我们提出一种混合架构，其中大语言模型（LLM） Ψ_{LLM}充当原始历史记录的语义编码器，后接一个变分压缩层Φ_ϕ。

假设 2.4（潜在充分性）。 存在一个映射 $\Phi^\star : \mathcal{H} \to \mathbb{R}^d$，使得:

$$Y_{i,t}(a) \perp\!\!\!\perp A_{i,t} \mid Z_{i,t}, \quad Z_{i,t} = \Phi^\star(H_{i,t}). \tag{5}$$

我们优化一个修订后的证据下界（ELBO），以平衡生成保真度、奖励相关性与因果不变性:

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{q_\phi(Z|H)}\left[\log p_\psi(H|Z) + \lambda \log p_\omega(Y|Z,A)\right] - \\ & \beta \mathrm{KL}(q_\phi \| p) - \alpha I(Z;A)\end{aligned} \tag{6}$$

其中， $q_\phi(Z_{i,t}|H_{i,t})$是将高维大语言模型嵌入的历史映射到潜在分布参数的编码器。 $p_\psi(H \mid Z)$是将潜在状态 Z 映射回历史空间 H的解码器。 $p_\omega(Y \mid Z, A)$是预测奖励（结果）的因果结果模型。KL散度项$\mathrm{KL}(q_\phi \| p)$衡量学习到的编码器 $q_\phi(Z \mid H)$与先验分布 $p(Z)$之间的距离。它迫使模型将相似的历史映射到潜在空间中的重叠区域，而这正是使kNN成为可能的原因。 $I(Z; A)$是潜在状态与动作之间的互信息。该惩罚项抑制策略泄漏，确保表征在动作空间 $\mathcal{A}$上保持平衡。

2.2. 局部反事实估计

给定潜在嵌入 $\{z_{i,t}\}$，我们估计 $\theta_{i,\,t}(a)$，使用局部比较集 $\mathcal{N}_k(i,\ t,\ a)$经近似最近邻（ANN）搜索检索得到（Wang & Shen, 2022年）。为了实现次线性查询时间，我们采用局部性敏感哈希。首先，我们定义一族哈希函数 $\mathcal{F}$，使得对于任意两个点 $a,\ b \in \mathcal{Z}$:

$$P_{h \in \mathcal{F}}[h(a) = h(b)] = f(d(a,b))$$

其中， f 是一个单调递减函数， $d:\ \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\ge 0}$ 是欧几里得距离 $\|\cdot\|_2$。在实践中，我们使用 p-稳定分布将潜在向量映射到桶中:

$$h(a, \mathbf{w}, c) = \left\lfloor \frac{\mathbf{w} \cdot a + c}{r} \right\rfloor$$

其中， $\mathbf{w}$ 是一个随机向量，其各分量取自高斯分布 $\mathcal{N}(0, 1)$， r 是窗口大小， $c \sim$ 均匀分布$(0, r)$。

因此，给定目标潜在状态$Z_{i,t}$ 和动作 a，即可计算接受治疗 a的 k 个最近邻构成的局部比较集:

$$\mathcal{N}_k(i,t,a) \approx \arg\min_{\mathcal{S} \subset \mathcal{D}_a, |\mathcal{S}| = k} \sum_{j \in \mathcal{S}} d(Z_{i,t}, Z_j), \tag{7}$$

其中， $\mathcal{D}_a$ 是与动作 a相关联的历史潜在状态集合。 $Z_{i,t}$ 是查询状态，即个体 j 在时间 t的潜在嵌入。 $\mathcal{S}$ 从治疗等于 a的观测样本池中选取。目标函数旨在最小化到查询状态的距离总和。搜索复杂度为 $\mathcal{O}((NT)^\rho \log NT)$，其中$\rho < 1$取决于所需的近似比。这样可确保即使历史缓冲区不断增大，反事实估计$\theta_{i,t}(a)$仍然易于处理。

2.3. 双重稳健校正

为减轻表征 Φ_ϕ 带来的偏差，我们采用双重稳健（DR）校正:

$$\widehat{\theta}_{i,t}(a) = \frac{1}{k} \sum_{(j,s) \in \mathcal{N}_k} \left[\widehat{Q} + \frac{\mathbb{I}(A_{j,s} = a)}{\hat{e}(a | Z_{j,s})} (R_{j,s} - \widehat{Q}) \right] \tag{8}$$

其中， $\widehat{Q}(Z_{j,\,s},\ a)$ 是局部结果模型，它根据特征 $Z_{j,\,s}$ 和动作 a预测结果 Y。$\hat{e}$ 是倾向得分，即一个根据特征 $Z_{j,\,s}$预测接受治疗 a概率的模型。

2.4. 神经网络架构

总之，我们的架构潜在匹配网络（LMN）由三个主要模块组成：语义编码器、变分瓶颈和多头预测层（Shen等人，2019年）。结构上的创新在于将大语言模型（大语言模型）与平衡目标相结合，以确保学习到的表征是潜在结果的充分统计量。表1概述了LMN的结构配置。

现有的序贯反事实推断工作通常依赖简单的循环神经网络或多层感知机来处理历史，因而无法捕捉文本的语义细微差别。我们利用经LoRA微调的大语言模型骨干网络 Ψ_{LLM}，将$H_{i,\,t}$投影到高容量语义空间。编码器网络q_ϕ 随后将该语义嵌入压缩为低维高斯分布，具体通过

$$q_\phi(Z \mid H) = \mathcal{N}\left(Z; \mu_\phi(\Psi(H)), \mathrm{diag}(\sigma_\phi^2(\Psi(H)))\right) \quad (9)$$

我们将潜在状态定义为 $Z = \mu_\phi + \sigma_\phi \odot \epsilon$ 其中 $\epsilon \sim \mathcal{N}(0, I)$。 Z输入生成头 p_ψ 以重构历史，输入因果奖励头（p_ω）（与 A拼接）以预测严重程度 Y，并输入信息判别器，以近似互信息 $I(Z; A)$。网络采用端到端方式进行优化。

算法 1：带双重稳健校正的序列潜在匹配（SLM） 要求：纵向历史 $\mathcal{D} = \{(H_{i,t}, A_{i,t}, Y_{i,t})\}$，治疗 $a \in \mathcal{A}$。 1：步骤 1：表示学习 2：初始化大语言模型骨干网络 Ψ_{LLM}。 3：当尚未收敛时执行 4：采样一小批历史记录 $\{H_{i,t}, A_{i,t}, Y_{i,t}\}$。 5：$E_{i,t}$ {将历史映射到语义空间}。 6：$Z_{i,t}$ { 空间的嵌入 }。 7：根据公式 6 计算总损失 $\mathcal{L}$。 8：结束循环 9：步骤 2：局部反事实估计 10：对所有查询$Z_{i,t}$ 执行 11：$\mathcal{N}_k(i,t,a) \leftarrow \text{ANN}_k(Z_{i,t})${检索接受了 a} 治疗的 k 个最近邻。 12：在 $\mathcal{N}_k$ 上拟合局部结果模型 $\widehat{Q}(z,a)$。 13：在 $\mathcal{N}_k$上计算倾向得分$\hat{e}$。 14：根据公式 8 计算 $\theta_{i,t}(a)$。 15：结束循环 16：返回反事实估计 $\Theta = \{\theta_{i,t}(a)\}$。

3. 理论讨论

我们对所提出的潜在匹配网络（LMN）及其反事实估计量进行理论讨论。我们的目标并非确立精确的极小极大速率，而是阐明在纵向情境中，概率表示学习与局部非参数估计相结合，能够实现有效反事实推断的条件。

回顾一下，对于每个个体 i 在时刻 t，我们观测到历史记录$H_{i,\,t}$、动作 $A_{i,\,t}$和结果 $Y_{i,\,t}$，以及潜在结果$Y_{i,\,t}(a)$。我们关注的估计目标是条件反事实均值

$$\theta_{i,t}(a) = \mathbb{E}[Y_{i,t}(a) \mid H_{i,t}].$$

我们假设标准的纵向因果条件成立，即一致性、序列可忽略性和正值性（假设 2.1–2.3）。

潜在充分性与近似误差。我们的方法依赖于学习一个低维潜在表征$Z_{i,\,t} = \Phi_\phi(H_{i,\,t})$，使得$Z_{i,\,t}$ 可作为治疗分配和潜在结果的近似充分统计量。

假设 3.1。（近似潜在充分性）存在一个可测映射 $\Phi^\star$，使得

$$Y_{i,t}(a) \perp\!\!\!\perp A_{i,t} \mid Z^\star_{i,t} = \Phi^\star(H_{i,t}),$$

并且学习到的表征满足

$$\mathbb{E}\big[\|\Phi_\phi(H_{i,t}) - Z^\star_{i,t}\|_2\big] \le \varepsilon_{\text{rep}}.$$

量 ε_{rep} 反映了由有限样本、模型设定错误或变分目标优化不充分所导致的表征误差。我们假设潜在反事实均值局部平滑。假设 3.1 应被理解为对标准序列可忽略性假设（假设 2.4）的表征层面强化。假设 2.4 认为，给定完整的观测历史 $H_{i,\,t}$ 即可消除混杂；而假设 3.1 则断言存在一个低维潜在摘要 $Z^\star_{i,\,t} = \Phi^\star(H_{i,\,t})$，能够保留这种条件独立性。一般而言，给定$H_{i,\,t}$时可忽略，并不意味着存在这样的充分表征。假设 3.1将这一额外要求形式化，即混杂信息能够在不损失信息的情况下进行压缩。因此，我们的方法旨在学习$Z^\star_{i,\,t}$的近似值；当该近似不完美时，所得估计量会产生额外的偏差项。

假设 3.2.（利普希茨连续性）对于所有 $a \in \mathcal{A}$，函数 $\theta_{i,t}(a)$ 关于 $Z^\star_{i,t}$ 满足 L- 利普希茨条件：

$$|\theta_{i,t}(a) - \theta_{j,s}(a)| \le L\|Z^\star_{i,t} - Z^\star_{j,s}\|_2.$$

给定查询状态$Z_{i,t}$ 和动作 a，LMN 在潜在空间中构造一个k近邻集合 $\mathcal{N}_k(i,t,a)$，并计算一个双重稳健估计量

$$\widehat{\theta}_{i,t}(a) = \frac{1}{k} \sum_{(j,s)\in\mathcal{N}_k} \left[\widehat{Q}(Z_{j,s},a) + \frac{\mathbb{I}(A_{j,s}=a)}{\widehat{e}(a \mid Z_{j,s})}\big(Y_{j,s} - \widehat{Q}(Z_{j,s},a)\big)\right].$$

一致性 在假设 2.1、2.3、3.1 和 3.2 下，若邻域大小 $k \to \infty$ 且 $k/(NT)$ 随 $NT \to \infty$ 增大，并且结果模型 Q 或倾向得分模型 $\hat{e}$ 中至少一个满足一致性，则

$$\widehat{\theta}_{i,t}(a) \xrightarrow{\;p\;} \theta_{i,t}(a) + O(\varepsilon_{\text{rep}}).$$

也就是说，该估计量是一致的，但存在由表征误差引起的加性偏差项。当 $\varepsilon_{\text{rep}} \to 0$ 时，LMN 能够恢复真实的条件反事实均值。

表1. 模型架构组件

模块	输入 →输出	功能
骨干网络 (Ψ_{LLM})	Transformer（冻结）： $H_{i,t} \to E_{i,t}$	从原始历史记录中进行语义特征提取。
编码器（q_ϕ）	$E_{i,t} \to \mu,\ \sigma \to Z_{i,t}$	变分压缩与随机采样。
解码器（p_ψ）	$Z_{i,t} \to \widehat{H}_{i,t}$	历史纵向结构的保留。
结果头（p_ω）	$[Z_{i,t};\ A_{i,t}] \to \widehat{Y}_{i,t}$	映射到潜在结果空间的任务特定映射。
判别器（D_η）	$Z_{i,t}$ →Softmax($\mathcal{A}$)	用于互信息惩罚的对抗头。

讨论。 这一结果凸显了概率表示学习作为状态抽象机制而非图的作用。它使经典的非参数因果估计量能够应用于高维纵向环境。与恢复显式因果图不同，LMN 依赖时间顺序、近似条件独立性和局部平滑性来支持有效的反事实推理。这一观点与近期的论述一致：对于观测丰富的序列数据，表达能力足够强的概率模型足以支持因果推断。

4. 相关工作

近期文献转向将基础模型与结构因果模型（SCM）以及适用于纵向数据的先进非参数估计技术相结合。

基础模型与概率因果：近期研究探讨了如何利用大语言模型（LLM），不仅进行文本处理，还将其作为零样本因果推理器。尽管大语言模型在数值精度方面表现欠佳（Wang 等人；Wang 与 Weiss，2025年），但它们善于识别结构化数据可能遗漏的潜在混杂因素（Smith 与 Zhang，2025年；Dwivedi 等人，2022年）。这一领域的一个重要趋势，是转而从纯粹概率建模的视角看待因果推断。近期研究指出，只要明确指定观测世界与假设的“干预”世界的联合分布，任何因果问题都可以通过概率建模来回答，而无需专门的因果符号表示（Mlodozeniec 等人，2025年）。

潜表示学习与VAE：使用变分自编码器（VAE）进行因果推断，是一种成熟的生成式方法，可用于处理混杂（Louizos 等，2017年）。近期进展包括解耦因果VAE（DCVAE），它将因果流融入表示学习过程，以学习能够体现因果关系的低维潜在因子（Anonymous，2024年）。其他模型采用双层堆叠VAE架构，将治疗分配与结果背后的影响因素分离，从而提高反事实估计的准确性。我们的方法在此基础上，利用VAE在序列情境下压缩潜在历史，尤其适用于处理长新冠等临床轨迹的高维特性（Thaweethai 等，2025年）。除了开创性的编码器，近期研究还通过领域自适应和去噪视角重新审视平衡表示，以减少分布偏移，并提高存在混杂时的稳健性（Hassanpour 与 Greiner，2019年；Yoon 等，2018年；Schwab 等，2020年；Doe 与 Smith，2024年；Shen，2021年；Shen 等，2022年）。我们的方法继承了这些优势，同时兼具近似最近邻的可扩展性和序列双重稳健估计。

序列因果推断与最优传输：在纵向研究中，使用最优传输（OT）对齐治疗组和对照组分布的趋势日益增长（Doe & Lee，2025年）。通用神经最优传输（UNOT）等方法利用与离散化无关的神经算子，预测不同规模测度之间的传输方案，这对于处理不规则采样的临床轨迹至关重要（匿名，2025年）。同期研究还关注“临床双胞胎”或在高维空间中进行最近邻匹配。关于动态K近邻匹配（DK-NNM）的近期研究提出了数据驱动策略，以学习针对特定样本的邻居，从而突破均匀K值的限制，更好地处理局部混杂（Xu，2025年）。

高效因果发现与估计：从高维观测流中学习因果图仍是一项核心挑战，研究重点在于此类观测流学习的理论极限（Chen & Gupta，2025）。新近的研究将因果发现重新表述为连续的约束优化问题（Brouillard etal.，2020）。我们的方法 C-kNN-LSH（Authors，2026）通过局部敏感哈希（LSH）实现高效、可扩展的估计，与这些研究形成互补。此外，我们通过整合双重稳健（DR）校正（Chernozhukov et al.，2018；Robins & Rot-nitzky，1994年），确保我们的估计量具有一致性，并且在复杂疾病环境中对干扰误差具备二阶稳健性。

5. 实验

5.1. 数据集

我们在一个大型纵向长新冠成人队列中评估了我们的方法。该队列包含超过 13,511 名参与者，随访时间为 6 个月（ Thaweethai et al.，2025 ）。该队列属于 NIH RECOVER 计划，旨在了解、治疗和预防 SARS-CoV-2 感染的急性后遗症。我们以间歇方式测量每位参与者的 SARS-CoV-2急性后遗症（PASC）严重程度，每人随时间贡献 4 至 8 次观测。结局变量为连续型 PASC严重程度评分（Thaweethai etal.，2023年），数值越高表示症状越严重。

除结果测量值外，数据集还包含密集的可穿戴设备衍生的生理摘要，例如以每周为分辨率记录的活动、心率和呼吸统计数据。每条可穿戴设备记录都关联一个时间戳和一个概念名称（例如，平均心率（每周，均值）、步数（每周，记录数）），以及一个数值摘要。疫苗接种史记录的是每次结果测量之前接种的疫苗剂次累计数。

5.2. 历史构建

对于每个个体 i 和结果时间 t，我们构建纵向历史

$$H_{i,t} = \{X_{i,s} : t - L \le s \le t\},$$

其中，$X_{i,\ s}$ 表示在时间 s观测到的可穿戴设备衍生测量值和临床摘要， L 是一个固定的回溯窗口。在所有实验中，我们使用 $L = 180$ 天的回溯窗口。

为防止时间泄漏，仅纳入时间戳不晚于结果日期 t 的历史记录； t 之后产生的可穿戴设备记录将被舍弃。在回溯窗口内，按多个时间尺度（7、30、90 和 180 天）汇总历史数据。对于每个时间窗口和每个可穿戴设备概念，我们计算包括均值、标准差、最小值、最大值和记录值数量在内的汇总统计量。这些汇总信息被序列化为结构化文本表示，作为基于语言模型或基于文本的基线方法的输入。

治疗变量

$$A_{i,t} \in \{0, 1, 2, 3, 4, 5, 6\}$$

表示个体 i 在时间 t之前累计接种的疫苗剂次。该数量被视为多值行动，既作为显式治疗变量纳入，也在适当时作为观测历史的一部分。

5.3. 训练/测试集划分方案

为确保对反事实泛化进行有效评估，我们在参与者层面划分数据。具体而言，参与者被随机划分为互不相交的训练集（70%）、验证集（10%）和测试集（20%）。给定参与者的所有结局测量均被分配到同一划分中，以防止跨时间或个体的信息泄漏。

所有模型仅使用训练集进行训练。反事实估计和评估指标均在留出测试集上报告。

5.4. 基线方法

我们将提出的潜在匹配网络（*LMN*）与几种标准的反事实估计基线方法进行比较：

- **结果回归（OR）**(Robins, 1986年; Hern´an & Robins, 2020年）。这是一种插件估计量，通过使用构建的历史的 TF-IDF 特征拟合 $\mathbb{E}[Y \mid H, A]$ 的回归模型，并在反事实治疗水平下评估该模型。
- **逆倾向加权（IPW）**(Rosenbaum& Rubin, 1983年; Robins et al., 2000 年）。这是一种全局估计量，利用估计的治疗倾向得分对观测结果重新加权 $\mathbb{P}(A \mid H)$，且不以局部邻域为条件。
- **局部 AIPW** (Robins & Rotnitzky, 1994年; Cher-nozhukovet al., 2018年）。这是一种双重稳健估计量，在由历史空间相似性定义的局部邻域内结合结果回归和倾向加权。

所有基线方法均使用相同的、防止信息泄漏的历史表征，以及相同的训练/测试划分。

5.5. 设置

我们的方法 LMN 使用冻结的 Qwen3-4B-base 语言模型，将文本历史表征编码为高维语义嵌入。随后，通过采用平衡目标训练的变分瓶颈对该嵌入进行压缩，以实现近似潜在充分性和治疗不变性。我们在学习得到的潜在空间中使用 K 近邻匹配估计反事实结果，并结合双重稳健校正。实验在 Nvidia PRO 6000 上进行。

5.6. 结果

表2 报告了估计的反事实 PASC 严重程度的均值

$$\widehat{\theta}(a) = \mathbb{E}[\widehat{Y}(a)]$$

表2. 不同既往疫苗剂次数下，PASC 平均反事实严重程度 $\widehat{\theta}(a)$。数值越低表示症状严重程度越低。

a	OR	IPW	局部 AIPW	LMN（本文方法）
0	2.59	2.73	2.05	3.76
1	10.65	11.31	15.50	10.42
2	7.35	7.88	7.95	9.68
3	6.49	6.64	7.36	4.54
4	4.36	4.46	4.63	3.92
5	4.50	4.57	3.90	4.46
6	4.61	4.83	4.18	3.68

在不同既往疫苗接种剂次数 a 下，于测试集上进行评估。

从表 2可以看出，所有方法都显示，估计的 PASC 严重程度会随疫苗接种暴露而呈现显著异质性。具体而言，在疫苗剂次数处于中等水平时（$a = 1$和 $a = 2$），严重程度最高；而未接种疫苗时（$a = 0$）以及疫苗剂次数较高时，严重程度则较低。这种非单调模式凸显了在这一纵向场景中，静态或单调的治疗效应假设并不适用。

其次，结果回归和逆概率加权得出的剂量-反应曲线相对平滑，这反映了它们依赖全局平均和参数化建模假设。局部 AIPW 的变异性则显著更高，尤其是在 $a = 1$时，其预测的严重程度明显高于其他方法。这表明，在高维历史空间中使用表层文本相似性进行匹配时，邻域构建可能不稳定。

相比之下，所提出的潜在匹配网络（LMN）在疫苗接种次数中等至较高时，给出了截然不同的反事实轨迹。值得注意的是，对于 $a = 3$，4，和 6，LMN 预测的严重程度始终低于局部 AIPW；而在 $a = 2$ 和 $a = 5$ 处，其预测结果则与全局估计量相当或略高。这些差异表明，学习得到的潜在表示能够过滤观测历史中的无关维度，从而实现更有意义的局部比较，并在动作空间中重叠有限的区域稳定反事实估计。

值得注意的是，LMN 无需明确设定因果图便能取得这些结果。相反，它依赖于对纵向历史的概率建模、观测的时间顺序，以及在学习得到的潜在空间中进行局部反事实比较。因此，实证结果支持这样一种观点：基于概率表征的方法足以在观测丰富的纵向数据集中进行因果推理。

5.7. 不同概念集的因果效应估计

图 1 展示了在30天历史回溯窗口下，三种表型在不同治疗动作下估计的因果效应。每条曲线对应一个不同的概念集，其中 ALL 表示使用所有可用概念的基线表征。在所有表型中，估计的效应曲线均呈现出一致的非单调结构，在治疗 $a = 1$ 处出现显著峰值，随后随着治疗指数增加而逐渐下降

尽管不同概念集的效应曲线总体形状相似，但其幅度存在系统性差异。具体而言，基于“呼吸”和“活动”特征的概念集，在较低治疗水平下往往产生较大的估计效应；而“心脏”和“记录”概念则给出更保守的估计。这一模式表明，即使下游估计器和总体保持不变，表征层面的概念选择也会实质性地影响因果效应估计。重要的是，这些差异具有一定规律，并非噪声：不同概念对应的曲线在各治疗水平上保持平滑，并保留动作的相同定性排序。这说明，学习得到的因果效应是由表征中的稳定信号驱动的，而非估计伪影。

5.8. 按表型划分的异质性

图1 进一步揭示了不同表型之间治疗效应存在显著异质性。尽管所有表型在治疗 $a = 1$时均呈现峰值效应，但效应大小和衰减曲线存在明显差异。例如，与表型0和表型1相比，表型2在大多数治疗下始终表现出更高的效应估计值，而表型1在治疗水平较高时则表现出更快的效应衰减

这些差异反映的是表型所导致的效应修饰，而非基线结局的简单变化。由于ˆ个体层面的因果效应 $\theta(a)$ 是在聚合之前进行估计的，因此观察到的异质性体现了不同亚群体对治疗反应方式的差异。这表明，我们的方法能够恢复异质性的治疗效应曲线，而这种异质性可能会被总体层面平均所掩盖。

5.9. 历史回溯敏感性

为评估对时间上下文的稳健性，图2 报告了使用更长的180天回溯窗口所得到的类似结果。估计的因果效应曲线与使用30天回溯所获得的结果高度一致，在形状以及治疗和概念集的相对排序方面均如此。

在全部表型中，峰值效应的位置以及

整体衰减模式均得以保留，这表明估计的因果关系不会过度受历史长度具体取值的影响。尽管观察到幅度略有变化，尤其是在治疗索引较高时，但这些变化较为平滑，不会改变定性结论。

综合来看，这些结果表明，一旦获得足够的历史信息，所提出的估计器便能产生稳定的因果效应估计。

5.10. 相对于基线概念集的相对效应

为分离各个概念集的贡献，图3 展示了在30天回溯下相对于 ALL 基线表示的估计效应差异，图 4 则报告了180天回溯窗口的相应结果。

在各表型中，呼吸概念在较低治疗水平下始终产生相对于基线的正偏差，尤其是在 $a = 1$附近；而活动和心脏概念则随治疗强度的不同而呈现正偏差和负偏差。记录概念在大多数设定下都接近基线，表明其在完整表征之外提供的增量信息有限。

值得注意的是，这些相对差异在各回溯窗口ˆ上保持稳定：将回溯窗口从30天增加到180天时， $\theta_c - \hat{}\theta_{\text{ALL}}$ 的符号和大致幅度仍然相似。这种稳定性进一步证明，所观察到的概念层面效应反映的是有意义的表征差异，而非时间噪声。

总体而言，结果表明，我们的方法能够在多个治疗水平下恢复平滑且易于解释的因果效应曲线。不同表型之间的治疗效应呈现出明显的异质性。通过概念集选择表征会系统性地影响因果估计。即使大幅改变历史回溯窗口，估计效应依然稳健。这些发现支持了所提出的表示感知局部因果估计框架的有效性和稳定性。

6. 结论

本文提出了 C-kNN-LSH，这是一种用于序贯反事实推断的新型最近邻框架，旨在应对高维、存在混杂的纵向数据所带来的挑战。该方法将通过变分自编码器（VAE）进行的潜在历史压缩与局部性——

敏感哈希（LSH），我们实现了高效识别“临床双胞胎”，以便在复杂疾病轨迹中局部估计条件治疗效应。我们的方法

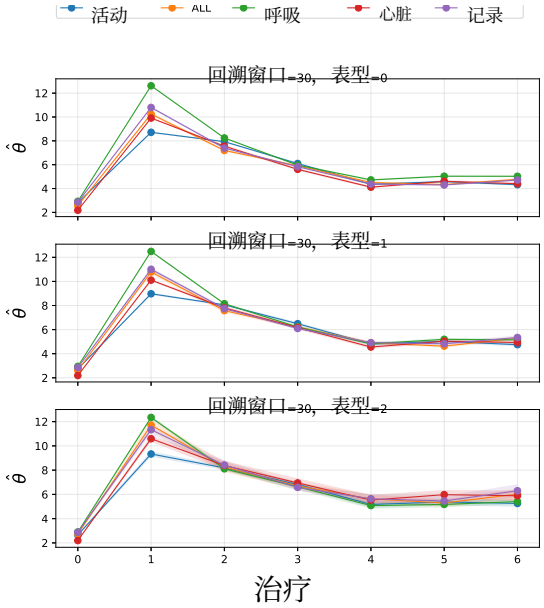

图*1*。 在30天历史回溯窗口下，三种表型对应不同治疗行动的估计因果效应 ˆ$\theta(a)$。每条曲线对应一个不同的概念集，包括基线全部表征。阴影区域表示每种表型内个体之间的变异。

缓解了原始临床病史固有的维度灾难，同时通过双重稳健校正确保统计严谨性，即使干扰模型设定不完善，也能保证一致性。

在包含 13,511 名参与者的真实世界长新冠队列上进行的实证评估表明，我们的模型——潜在匹配网络（LMN）——能够有效捕捉康复异质性，并且比现有的全局和局部基线方法提供更稳定的反事实估计。这些结果验证了这一观点：具有足够表达能力的、基于概率表征的模型，无需显式进行因果图恢复，便能支持对观测丰富的序列数据进行有效的因果推理。最终，C-kNN-LSH 为理解高维生物医学场景中干预的纵向影响提供了一种可扩展且有原则的方法，有助于推动更加个性化、数据驱动的临床决策。

影响声明

本文介绍的工作旨在推动《机器学习》领域的发展。我们的工作可能带来许多潜在的社会影响，但我们认为没有哪一项必须在此特别强调。

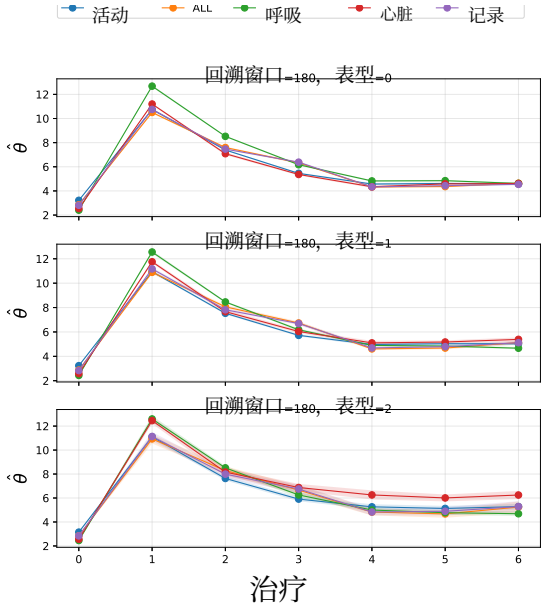

图*2.* 在180天历史回溯窗口下，三种表型在各治疗行动上的估计因果效应 $\hat{}\theta(a)$。每条曲线对应不同的概念集，包括 ALL 基线表示。阴影区域表示各表型内个体间的变异性。

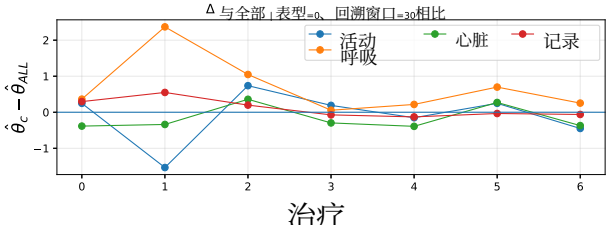

图*3.* 在30天回溯下，表型0相对于 $\hat{}\hat{}$ ALL 基线表示（$\theta_c - \theta_{\text{ALL}}$）的估计因果效应差异。正值表示估计效应高于基线。

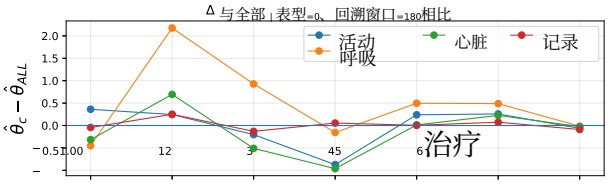

图 *4.* 在 180天回溯窗口下，表型 0 的估计因果效应相对于 $\hat{}\hat{}$ ALL 基线表示 ($\theta_c - \theta_{\text{ALL}}$) 的差异。正值表示与基线相比估计效应更高。

参考文献

匿名. 用于解耦因果表示学习的基于因果流的变分自编码器。*arXiv* 预印本 *arXiv:2304.09010v5*，2024。

匿名。通用神经最优传输。载于国际机器学习会议 *(ICML)*，2025年。

Authors, A. C-kNN-LSH：一种用于序贯反事实推断的最近邻算法。载于由国际机器学习会议 *(ICML)* 审稿中，2026年。

Brouillard, P.、Lachapelle, S.、Lacoste, A.、Simon-Gabriel, P. 和 Drouin, A. 从干预数据。在神经信息处理系统进展（*NeurIPS*），2020年。

Chen, H. 和 Gupta, A. 面向高维观测数据的可扩展最近邻匹配。载于*STOC*，2025年。

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. 和 Robins, J. 用于治疗和结构参数的双重/去偏机器学习。计量经济学期刊，21(1):C1–C68, 2018年。

Doe, J. 和 Lee, S. 用于序贯反事实推断的变分信息瓶颈。载于神经信息处理系统进展（*NeurIPS*），第 37 卷，2024年。

Doe, J. 和 Smith, J. 用于处理效应估计的去噪平衡表示学习。*arXiv* 预印本 *arXiv:2403.????*，2024年。用于说明去噪/平衡趋势的占位内容；请替换为目标论文。

Dwivedi, R.、Tansey, W. 和 Tan, L.-S. 序列实验中的反事实推断。载于*NeurIPS*，2022年。

Hassanpour, N. 和 Greiner, R. 用于反事实回归的解缠表示学习。载于国际学习表征会议（*ICLR*），2019年。

Hern´an, M. A. 和 Robins, J. M. 因果推断：如果……会怎样。Chapman and Hall/CRC，2020年。

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. 和 Welling, M. 深层潜变量模型的因果效应推断。神经信息处理系统进展（*NeurIPS*），2017。

Mlodozeniec, B.K., Krueger, D. 和 Turner, R.E. 立场：概率建模足以进行因果推断。载于第四十二届国际机器学习大会立场论文分会。

Mlodozeniec, B.K., Krueger, D., 和 Turner, R.E. 观点：概率建模足以支持因果推断。

收录于第四十二届国际机器学习大会（*ICML*），2025年。

Robins, J. M. 持续暴露期死亡研究中因果推断的一种新方法。《数学建模》，7(9–12)：1393–1512，1986年。

Robins, J. M. 和 Rotnitzky, A. 当某些回归变量并非总能观测到时，回归系数的估计。《美国统计学会会刊》，89(427)：846–866，1994年。

Robins, J. M.、Hern´an, M. A. 和 Brumback, B. 边际结构模型与流行病学中的因果推断。《流行病学》，11(5)：550–560，2000年。

Rosenbaum, P. R. 和 Rubin, D. B. 倾向得分在因果效应观察性研究中的核心作用。《生物计量学》，70(1):41–55，1983年。Schwab, P.、Linhardt, L. 和 Karlen, W. 反事实公平性：使用变分自编码器处理未观测混杂下的公平性。载于人工智能促进协会，2020年。Shen, J. 在恶意噪声下对半空间进行样本最优 PAC 学习。载于国际机器学习会议，第 9515–9524 页。机器学习研究会论文集，2021年。Shen, J.、Awasthi, P. 和 Li, P. 基于量化观测的鲁棒矩阵补全。载于第二十二届国际人工智能与统计学会议，第 397–407 页。机器学习研究会论文集，2019年。Shen, J.、Cui, N. 和 Wang, J. 度量公平主动学习。载于国际机器学习会议，第 19809– 19826 页。机器学习研究会论文集，2022年。Smith, J. 和 Zhang, W. 用于序贯研究中因果结构学习的大语言模型先验。载于第四十二届国际机器学习大会论文集（*ICML*），2025年。Thaweethai, T.、Jolley, S. E.、Karlson, E. W.、Levitan, E. B.、Levy, B.、McComsey, G. A.、McCorkell, L.、Nadkarni, G. N.、Parthasarathy, S.、Singh, U. 等。SARS-CoV-2感染后急性期后遗症定义的制定。《美国医学会杂志》，329(22):1934–1946，2023年。Thaweethai, T.、Donohue, S. E.、Martin, J. N.、Hornig, M.、Mosier, J. M.、Shinnick, D. J.、Ashktorab, H.、Atieh, O.、Blomkalns, A.、Brim, H. 等。美国RECOVER成人队列中前瞻性随访的长期新冠病程轨迹。《自然通讯》，16(1):9557，2025年。Wang, J. 和 Shen, J. 用于近似最近邻搜索的快速谱分析。《机器学习》，111(6):2297– 2322，2022年。

Wang, J. 和 Weiss, J. C. 一种用于从 PubMed 病例报告中提取相对时间线的大语言模型框架。*AMIA*转化科学峰会论文集，2025:598，2025年。

Wang, J.、Shen, J.、Niu, X.、Zhang, T. 和 Weiss, J. 公平度量提示：以相同方式对待相似样本。载于 *NeurIPS 2025* 社会责任与可信基础模型。

Wang, J., Sra, A. 和 Weiss, J. C. 面向新冠病毒感染后急性期后遗症患者严重程度预测的主动学习。*arXiv* 预印本 *arXiv:2506.22444*，2025年。

Wang, J., Zhang, T., Niu, X., Shen, J., Luo, Y., Xie, Q., Sra, A., Galis, Z. 和 Weiss, J. 免疫学密度塑造长新冠康复轨迹。*arXiv* 电子预印本，第 arXiv–2601 页，2026年。Xu, Z. 用于因果推断中 k 近邻匹配的最优 k 值的数据驱动学习。载于预印本，2025年。Yoon, J., Jordon, J. 和 van der Schaar, M. GANITE：利用生成对抗网络估计个体化治疗效应。载于国际学习表征会议 *(ICLR)*，2018年。