arXiv:2601.07977v1 [stat.ME] 12 Jan 2026

两种随机过程的联合建模：应用于从污水病毒浓度中学习住院动态

K. Ken Peng[a, c]、Charmaine B. Dean*[b]、Robert Delatolla[c]、X. Joan Hu[a]和Elizabeth Renouf[c]

[a]西蒙菲莎大学统计学与精算科学系，加拿大不列颠哥伦比亚省本拿比市University Dr 8888号，邮编 V5A 1S6[b]滑铁卢大学统计学与精算科学系，加拿大安大略省滑铁卢市 University Ave W 200号，邮编 N2L 3G1[c]渥太华大学土木工程系，加拿大安大略省渥太华市

摘要

在 COVID-19 后疫情时代，住院仍然是首要的公共卫生问题，而污水监测已成为在社区层面监测其动态的重要工具。然而，通常没有足够的信息来了解导致污水病毒信号和住院人数的感染过程。这一关键挑战促成了本文所提出的统计框架。我们通过个体层面的潜在感染过程，构建随时间变化的污水病毒信号与住院人数之间的联系。我们提出了一种处理汇总数据的策略，这是监测数据的一种典型形式。此外，我们利用感染过程的可用信息（其中可能存在漏报），通过联合建模简化了传统的统计学习流程。模拟研究表明，在不同程度的漏检情况下，所提方法均能产生稳定的推断结果。对加拿大渥太华的 COVID-19 监测数据分析显示，在若干报告假设下，该框架能够恢复感染率和变异株特异性住院风险的一致时间模式。

关键词：环境监测、间接统计学习、多状态过程、伪似然估计、污水流行病学

*通讯作者

1 引言

COVID-19大流行凸显了开发可靠且可扩展的方法、在人群层面监测疾病传播的必要性。传统监测方法，如病例报告和住院记录，通常存在滞后，并且会因检测可及性和就医行为而产生偏差（Starnini 等，2021；Sherratt 等，2021）。作为一种替代方案，污水与环境监测（WES）作为一种具有成本效益且能够领先指示社区传染病变化的方法，受到了广泛关注（Parkins 等，2024）。通过分析污水中的病毒信号，研究人员可以估计感染率的变化趋势，而无需依赖个体层面的检测数据（Holcomb 等，2024；Pappu 等，2025；Mohring 等，2024）。在本文中，我们将污水病毒信号定义为污水样本（例如初沉污泥）中测得的SARS-CoV-2 RNA浓度，以每毫升基因拷贝数表示。该指标反映了一种汇总性的环境信号，其来源是相关人口中感染者排出的病毒。

将 WES 应用于 COVID-19，已经让我们对社区传播动态有了诸多认识。研究表明，污水病毒浓度与报告病例数和住院人数相关，且常能在感染激增前发出预警信号（Larsen & Wigginton 2020，Peng 等，2023；Hegazy 等，2022）。研究人员采用了多种方法对这些关系进行建模，包括分布滞后模型（Galani 等，2022；Peccia 等，2020；Kaplan 等，2021；Zulli 等，2022；Schoen 等，2022；Xie 等，2022；Peng 等，2023、2025）、时间序列模型（Jeng 等，2025；Lai 等，2025；Simone 等，2024）以及分舱流行病学模型（Mohring 等，2024；Meadows 等，2025；Miyazawa 等，2024）。尽管这些方法证明了 WES 的实用价值，但它们往往依赖于关于污水病毒信号与临床指标之间直接关系的强假设。

将污水病毒信号与住院联系起来的一个核心挑战在于，二者之间的关联是动态的且具有变异株依赖性，而非稳定或可直接解释的。相同的污水病毒信号在不同时间可能对应截然不同的住院负担，这反映了感染率、个体层面的病毒排出、变异株特异性住院风险，以及任一时点人口中多个病毒变异株共存等因素的变化，而不是由某个离散转变或阈值界定的单一主导变异株（例如，D’Aoust 等，2022）。因此，直接对二者的关联进行建模会将不同过程混为一谈，并限制推断的可解释性。应对这些挑战，需要将感染建模为一种中介过程，在个体层面同时支配病毒排出和住院风险。然而，监测数据很少能够直接支持此类建模。在实践中，个体级信息不可获得，观测数据也只能在社区层面进行汇总。此外，即使能够获得报告病例数等感染信息，由于无症状病例、有限的检测能力以及不断变化的检测政策，这些信息通常也会低估真实感染（Lau 等，2021；Albani 等，2021）。这两种限制性情形促使我们构建一种统计框架，仅利用不完美的群体层面监测数据来学习个体层面的疾病过程。

为同时分析多个相关结局或过程并考虑它们之间的依赖关系，联合建模广泛应用于生物统计学和流行病学中。传统的

联合建模通常通过潜在结构，将针对重复测量（如生物标志物）的纵向子模型与针对生存时间结局的生存或事件时间子模型联系起来（Tsiatis & Davidian 2004）。Wu 等人（2012）对联合模型的推断程序作了全面综述。常见应用包括将生物标志物与生存结局联系起来、对治疗过程中的疾病进展进行建模，以及在临床研究中整合异质数据源，例如 HIV相关研究（Wu 等人 2010，Temesgen 等人 2018，Luvanda 等人 2023）和心血管疾病研究（Zelelew & Workie 2023，Gilani 等人 2017）。

在无法直接观测个体层面过程、只能获得汇总摘要的推断情境中，大量文献表明，若未妥善处理，汇总可能造成信息损失，并导致过度自信的推断（Lawless & McLeish 1984, Orcutt et al. 1968, Banks et al. 2020）。针对汇总数据的估计已在多种情境中得到研究，包括马尔可夫过程（Davis et al. 2002, Kelton & Kelton 1987）、伽马分布过程（Chen & Ye 2017）以及时空模型（Nguyen et al. 2023）。尽管一些理论结果表明，在严格条件下，汇总可能不会造成渐近信息损失，但实践中很少满足这些假设。据我们所知，这些问题尚未在基于污水的传染病监测背景下得到系统研究。

在本研究中，我们针对污水监测情境定制了一个联合建模框架，其中感染过程仅能被部分观测，且可用数据为人群层面的汇总数据。我们构建了一个简约的个体层面联合模型，其中感染过程介导污水病毒排出和住院风险，并提出一种估计策略，以便在实践中应对两种常见的数据限制：缺少个体层面观测，以及感染信息存在系统性漏报。我们的方法不依赖污水信号与临床结局之间的直接或确定性关系，而是将这两个过程都视为条件依赖于一个共同的潜在感染过程。我们通过在一个连贯的统计框架内综合这些来源的信息来实现推断，而不假定任何单一数据流都能完整反映疾病动态。

本文其余部分安排如下：第 2 节介绍我们提出的建模框架。第 3 节讨论估计程序。第4节报告一项模拟研究，评估所提出估计程序的性能，并考察方差估计策略。第 5 节将模型应用于渥太华的污水与住院数据，展示其实践应用及结果。第 6 节总结研究发现的意义、可能的扩展方向，以及基于污水的疾病监测未来的发展方向。

2 联合建模框架

2.1 记号与建模

我们考虑一个用于传染病监测中多个随机过程的联合建模框架。我们定义了三个随连续日历时间演变的过程$t \in (0, E)$，其中 E 表示研究期间的结束时间。

第一个过程 $s(t)$ 表示时刻 t的感染状态。该过程定义在有限状态空间上，可以是多状态过程，从而涵盖不同类型的感染（例如，感染变体）。第二个过程 $h(t)$ 表示时刻 t 的住院状态，我们将其视为一个二元过程，用以指示个体是否住院。第三个过程 $w(t)$ 表示时刻 t的监测目标，例如对病毒载量进行连续测量。此外，我们还允许存在随时间变化的协变量 $x(t)$，这些协变量可能影响上述一个或多个过程。

该联合模型假设：给定潜在感染过程和观测到的协变量后，监测目标过程与住院过程条件独立。在每个时点 t，感染过程 $s(t)$作为驱动过程，将监测目标 $w(t)$与住院过程 $h(t)$联系起来。具体而言，给定感染状态 $s(t)$和协变量 $x(t)$后， $w(t)$和 $h(t)$的分布分别通过 $[w(t) \mid s(t)\ x(t)]$ 和 $[h(t) \mid s(t)\ x(t)]$, 来指定，而 $s(t)$的演变则由 $[s(\cdot) \mid x(\cdot)]$控制。联合模型通过以下条件分解定义

$$[s(\cdot), w(\cdot), h(\cdot) \mid x(\cdot)] = [s(\cdot) \mid x(\cdot)][w(\cdot) \mid s(\cdot), x(\cdot)][h(\cdot) \mid s(\cdot), x(\cdot)]. \tag{1}$$

该分解是在一般个体层面定义的，是总体层面推断的构建基础。疫苗接种是感染动态和住院风险的重要决定因素。尽管受数据限制和汇总方式的影响，此处并未显式建模疫苗接种的作用，但所提出的框架可以轻松纳入与疫苗接种或免疫相关的协变量，作为自然的扩展。

这一表述构成了一个联合模型，其中感染在两个可观测过程之间发挥中介作用。病毒排出（由监测目标 $w(t)$ 捕获）和住院风险（由 $h(t)$ 捕获）均受感染状态驱动，但代表疾病进展的不同表现。在给定 $s(t)$ 的条件下，我们假设 $w(t)$ 与 $h(t)$ 相互独立，这反映出这样一种解释：感染状态足以概括在时间 t 时与病毒排出和住院均相关的疾病严重程度方面。该条件独立假设在保留感染所起关键机制联系的同时，简化了联合结构。

在实践中，个体层面的感染状态和遗传变体很难被完全准确地观测到，可能只能部分观测、存在延迟，或完全无法观测。相比之下，病毒载量测量和住院事件等监测目标更易于观测，并可提供有关感染状态的间接信息。例如，即使未直接测量感染状态，检测到阳性病毒信号或发生住院事件也表明存在感染。所提出的框架以个体为层面进行构建，并适用于过程中的某些组成部分只能被部分观测的情形。

下面我们将介绍式 (1) 中联合模型各个组成部分的一般形式。第2.2节中引入的具体参数选择是这一通用框架的特例。

感染过程。潜在感染过程 $s(\cdot) = \{s(t) : t \in (0, E)\}$ 被建模为一个有限状态空间 $\{0, 1, \ldots, K\}$ 上的连续时间多状态随机过程，

其中状态 0 表示未感染状态，状态 $k = 1, \ldots\ldots, K$ 对应不同的遗传变异株。 $s(\cdot)$ 的演化通过转移强度$\gamma_{jl}(t, \mathcal{H}_t)$来刻画，其中 $\mathcal{H}_t$ 表示截至时间 t 的过程历史。该模型既适用于转移强度 $\gamma_{jl}(t, \mathcal{H}_t)$仅取决于当前状态和日历时间的非齐次马尔可夫模型，也适用于允许转移强度取决于当前状态驻留时间的半马尔可夫扩展。

病毒载量过程。监测过程 $w(\cdot)$ 表示与感染相关的个体层面测量，例如病毒排出。在给定潜在感染状态和协变量的条件下， $w(t)$ 通过状态依赖的条件分布来设定，

$$w(t) \mid s(t), \mathbf{x}(t) \sim G_{s(t)}\big(\cdot;\, \mathbf{x}(t)\big),$$

其中 $\{G_k(\cdot; \mathbf{x}) : k = 0、\ldots\ldots、 K\}$ 表示一个由感染状态索引的分布族。这种表述将监测信号视为由潜在感染动态驱动的测量或发射过程。根据具体应用和数据特征， G_k 的选择可以有所不同，从而允许采用灵活的分布形式，而不限于本研究中用于说明的伽马分布设定。

住院过程。住院过程被建模为一个生存时间过程，用于描述首次住院入院的时间。等价地，它也可以视为一个包含状态 {未住院、已住院} 的两状态过程，其中已住院状态为吸收态。令 H 表示住院时间（若发生）。在给定感染状态和协变量的条件下，住院风险通过风险函数来刻画，

$$\lambda(t \mid s(t-), \mathbf{x}(t)),$$

定义在日历时间尺度上。从机制上看，住院风险取决于感染后的经过时间；这种依赖关系通过感染过程和感染起始时间 $\{T_{k,r}\}$隐式体现。因此，此处采用的日历时间表述等价于感染后时长表示，同时仍与群体层面汇总数据相兼容。

2.2 模型组成部分的具体设定

第 2.1 节所述的一般框架为感染、监测和住院过程提供了多种建模选择。本节介绍我们在分析中采用的具体参数化设定。这些选择基于现有污水与住院数据的结构，旨在以简约而灵活的方式刻画流行病动态。

2.2.1 感染过程 $s(\cdot) \mid x(\cdot)$

作为第 2.1 节中介绍的一般多状态感染模型的一个示例，我们考虑一种允许存在多个遗传变体的非齐次马尔可夫模型。状态转移仅限于未感染状态与感染状态之间，

从而排除合并感染以及不同遗传变体之间的直接转移。在任一给定时刻 t，多个变异株特异性感染强度 $\gamma_{0k}(t)$ 可以为正，因此不同遗传变体可在人群层面共同传播。

从未感染状态转变为遗传变体 k 的转变强度通过高斯函数混合建模，

$$\gamma_{0k}(t) = \sum_{m \in \mathcal{M}_k} a_{k,m} \exp\left\{ -\frac{(t - b_{k,m})^2}{2c_{k,m}^2} \right\}, \qquad k = 1, \ldots, K. \tag{2}$$

这里，$\mathcal{M}_k$ 为用于表示变体 k 的时间发病率曲线的高斯成分集合建立索引，成分数量的选择则反映了该变体所考虑的不同流行病时期的数量。这一设定能够灵活表示感染发病率随时间的上升和下降。为简便起见，感染过程中未纳入协变量效应；但如果有相关信息，这一通用框架可以轻松扩展以纳入此类效应。高斯成分能够平滑且简约地表示流行波，而混合模型则可在不强加突变式转变的前提下捕捉多个传播波或时期。这种选择为随时间变化的感染发病率提供了一种灵活且易于解释的近似。

2.2.2 病毒载量过程 $w(\cdot) \mid s(\cdot), \mathbf{x}(\cdot)$

我们将病毒载量过程设定为一个以感染状态和协变量为条件的状态依赖测量模型。具体而言，在存在感染时，我们使用伽马分布来实例化上文介绍的一般族 $G_k(\cdot;\mathbf{x})$。鉴于病毒载量测量的非负性和右偏特征，我们假设在存在感染时，监测信号服从伽马分布，

$$w(t) \mid s(t), \mathbf{x}(t) \sim \begin{cases} 0, & s(t) = 0, \\ \mathrm{Gamma}\Big(\alpha_k,\, \exp\{\beta_0^{(k)} + \mathbf{x}(t)^\top \boldsymbol{\beta}^{(k)}\}\Big), & s(t) = k, \end{cases} \tag{3}$$

这一设定提供了灵活的均值结构，同时允许个体间的病毒排出存在较大差异。如果数据有此需要，也可以在同一框架内采用其他分布形式。

2.2.3 住院过程 $h(\cdot) \mid s(\cdot), \mathbf{x}(\cdot)$

尽管从概念上看，住院可表示为一种二元状态（住院或未住院），但现有住院数据只记录住院入院情况，不记录出院信息。因此，我们考虑一种简化情形，将住院建模为首次住院入院的时间。这对应于一个双状态过程，个体从未住院状态转变为吸收性住院状态。

在给定感染状态的条件下，我们将日历时间尺度上的住院风险规定为

$$\lambda(t \mid s(t-), \mathbf{x}(t)) = \begin{cases} 0, & s(t-) = 0, \\ \exp\{\lambda_0^{(k)} + \mathbf{x}(t)^\top \boldsymbol{\lambda}^{(k)}\}, & s(t-) = k, \end{cases} \tag{4}$$

在此设定下，$\exp\{\lambda_0^{(k)}\}$ 表示变异株特异性基线风险。假设对于感染了变异株 k的个体，在给定协变量$\mathbf{x}(t)$的条件下，该风险随时间保持恒定。之所以采用日历时间尺度来表示风险，是为了与现有的汇总监测数据保持一致，因为这些数据不包含个体感染后时间的信息。若要允许基线风险随感染后的经过时间变化，就需要引入基于感染起始的替代时间尺度；本文不作此探讨。

3 统计学习

3.1 完整个体层面数据下的推断

我们首先介绍一种理想化情形下的基于似然的推断，在该情形中，个体级信息完整可得。这一表述为后续内容提供了概念上的参照，并清晰地衔接了第 2 节引入的个体层面联合模型与后续章节中提出的汇总数据估计策略。

我们考虑一个由 N 个个体构成的人口，个体索引为 $i = 1$、.....、N。对于个体 i，令 $s_i(t)$、$w_i(t)$ 和 $h_i(t)$ 分别表示时间 $t \in (0,\ E)$ 的感染状态、监测测量和住院过程，并令其对应的协变量过程为 $\mathbf{x}_i(t)$。在本小节中，我们假设完整轨迹 $\{s_i(t)\ w_i(t)\ h_i(t) :\ t\ \in(0,\ E)\}$ 均被无误差地观测到。假设所有个体均以未感染状态开始，即 $s_i(0) = 0$。

对于个体 i，令 T_i 表示首次住院时间，令 E 表示研究期间的结束时间。观测到的随访时间为 $U_i = T_i \wedge E$，删失指标为$\delta_i = I(T_i < E)$。此外，假设个体层面的监测过程 $\{w_i(t)\}$ 在一组离散观测时间（例如每日）上均无缺失地观测到，且感染过程 $\{s_i(t)\}$ 在个体层面被完整观测到。

令参数向量 $\boldsymbol{\theta}_h$、$\boldsymbol{\theta}_w$ 和 $\boldsymbol{\theta}_s$ 分别控制与住院、监测和感染过程相关的参数。给定感染过程后，住院过程的似然贡献为

$$L_{h,i}(\boldsymbol{\theta}_h) = \lambda(U_i \mid s_i(U_i-), \mathbf{x}_i(U_i))^{\delta_i} \exp\left\{ - \int_0^{U_i} \lambda(v \mid s_i(v-), \mathbf{x}_i(v))\ dv \right\}, \tag{5}$$

其中 $\lambda(t \mid s_i(t-),\ \mathbf{x}_i(t))$ 表示住院的条件风险。

令 $\mathcal{T}_i \subset (0,\ E)$ 表示测量监测目标 $w_i(t)$ 的观测时间集合（例如，每日测量）。在条件模型 $[w_i(t)\ \mid s_i(t),\ \mathbf{x}_i(t)]$ 的密度为 $f_w(\cdot \mid s_i(t),\ \mathbf{x}_i(t))$ 时，监测似然为

$$L_{w,i}(\boldsymbol{\theta}_w) = \prod_{t \in \mathcal{T}_i} f_w\big(w_i(t) \mid s_i(t), \mathbf{x}_i(t)\big). \tag{6}$$

对于 (2.2.2) 中的示例性伽马分布设定，当 $s_i(t) = 0$ 时，该密度对应于 0 处的点质量；当 $s_i(t) = k$ 时，则对应于伽马密度。

感染过程被建模为一个具有转移强度 $\gamma_{jl}(t)$ 的连续时间多状态过程。令 $0 = t_{i0} < t_{i1} < \cdots < t_{iJ_i} < E$ 表示 $s_i(\cdot)$ 的转移时间，并且

令 s_{ij} 表示在 $(t_{ij}, t_{i,j+1})$ 上所处的状态。感染过程的似然贡献为

$$L_{s,i}(\boldsymbol{\theta}_s) = \left\{ \prod_{j=1}^{J_i} \gamma_{s_{i,j-1},\, s_{ij}}\big(t_{ij}, \mathcal{H}_{it_{ij}}\big) \right\} \exp\left\{ -\int_0^E \Gamma_{s_i(v)}(v, \mathcal{H}_{iv})\; dv \right\}, \tag{7}$$

其中，$\Gamma_{s_i(v)}(v, \mathcal{H}_{iv}) = \sum_{l \neq r} \gamma_{s_i(v)l}(v, \mathcal{H}_{iv})$ 是在时刻 v 从状态 $s_i(v)$ 离开的总离开强度。在我们的应用中采用的简化转移结构下（仅在状态 0 与状态 $1, \ldots, K$ 之间发生转移），$\Gamma_{s_i(v)}(\cdot)$ 中的求和范围也相应受到限制。

如果 $s_i(\cdot)$、$w_i(\cdot)$ 和 $h_i(\cdot)$ 均在个体层面上可观测，那么在条件独立假设 $h_i(\cdot) \perp w_i(\cdot) \mid s_i(\cdot)\ x_i(\cdot)$ 下，完整数据似然可分解为

$$L_i^{(c)}(\boldsymbol{\theta}_s, \boldsymbol{\theta}_w, \boldsymbol{\theta}_h) = L_{h,i}(\boldsymbol{\theta}_h)\, L_{w,i}(\boldsymbol{\theta}_w)\, L_{s,i}(\boldsymbol{\theta}_s).$$

如果感染过程 $s_i(\cdot)$ 未被观测，则推断基于对潜在路径积分所得的边际似然，

$$L_i^{(o)}(\boldsymbol{\theta}_s, \boldsymbol{\theta}_w, \boldsymbol{\theta}_h) = \int_{\mathcal{S}} L_{h,i}(\boldsymbol{\theta}_h)\, L_{w,i}(\boldsymbol{\theta}_w)\, L_{s,i}(\boldsymbol{\theta}_s)\, d[s_i(\cdot)],$$

其中，$\mathcal{S}$ 表示可容许感染轨迹的空间。假设个体之间相互独立，则完整边际似然为

$$L^{(o)}(\boldsymbol{\theta}_s, \boldsymbol{\theta}_w, \boldsymbol{\theta}_h) = \prod_{i=1}^{N} L_i^{(o)}(\boldsymbol{\theta}_s, \boldsymbol{\theta}_w, \boldsymbol{\theta}_h).$$

实际上，监测系统通常只提供 $w_i(\cdot)$ 和 $h_i(\cdot)$ 的汇总摘要，而非个体层面轨迹，因此需要采用下文提出的伪似然方法。

3.2 聚合监测数据下的伪似然

在实践中，监测系统很少能观测到感染状态和病毒排出的完整个体层面轨迹。相反，可用数据通常由在离散时间点收集的人口层面汇总数据构成。因此，我们采用一种估计策略，通过伪似然构造将第 2 节介绍的个体层面联合模型与聚合观测值联系起来。

我们考虑按离散时间索引的观测 $t = 0$，1，.....，E。令 H_t 表示时间 t与目标疾病相关的住院总人数，令 $W_t^{(k)}$ 表示与遗传变体 k相关的汇总监测信号，该信号通过对个体层面的贡献求和得到，

$$H_t = \sum_{i=1}^{N} h_i(t), \qquad W_t^{(k)} = \sum_{i=1}^{N} w_i(t)\, I\{s_i(t) = k\}.$$

此外，监测系统可能报告现有病例的观测数量 S_t^*，而该数量通常低于感染者的真实数量

$$S_t = \sum_{i=1}^{N} I\{s_i(t) \neq 0\}.$$

我们不假设 S_t^* 是 S_t的无偏或完整测量，而是将其视为有关感染率的部分信息。对病例查明程度的不同假设，会导致对伪似然中 S_t^*的不同解释；我们通过第5.1小节中的情景分析，系统地探讨这些假设。

为将个体层面模型与汇总数据联系起来，我们将时间离散化，并用离散时间对应形式近似连续时间似然。在个体之间相互独立的假设下，个体住院似然意味着：给定遗传变体后，每日住院人数服从泊松型似然。具体而言，令$H_t^{(k)}$ 和 $R_t^{(k)}$ 分别表示时间 t与遗传变体 k 相关的（未观测）住院人数和风险人群人数，则住院部分的似然贡献形式如下：

$$L_h(\boldsymbol{\theta}_h) = \prod_{t=0}^{E} \prod_{k=1}^{K} \Big[f(\mathbf{X}_t; \boldsymbol{\lambda}^{(k)})^{H_t^{(k)}} \exp\big\{ - f(\mathbf{X}_t; \boldsymbol{\lambda}^{(k)})\, R_t^{(k)} \big\} \Big], \tag{8}$$

其中$\mathbf{X}_t$ 表示群体层面协变量。

类似地，在个体监测测量值的条件伽马模型下，聚合信号 $W_t^{(k)}$ 服从伽马分布，其形状参数与类型为 k的感染者人数成正比，

$$L_w(\boldsymbol{\theta}_w) = \prod_{t=0}^{E} \prod_{k=1}^{K} g\Big(W_t^{(k)}; S_t^{(k)} \alpha_k,\, f(\mathbf{X}_t; \boldsymbol{\beta}^{(k)}) \Big), \tag{9}$$

其中， $g(\cdot)$表示伽马密度， $S_t^{(k)}$ 表示在时间 t感染类型 k 的个体数量。

基于期望的近似。 这些量 $H_t^{(k)}$ 、 $R_t^{(k)}$ 和 $S_t^{(k)}$ 无法直接观测。为了构造一个易于处理的似然函数，我们用给定聚合观测值的条件期望替代这些潜在变异株特异性量。具体而言，我们根据时间 t的感染构成分解群体层面计数，

$$\begin{aligned}
E\Big(S_t^{(k)}\Big) &= S_t\, P\{s(t) = k \mid s(t) \neq 0\}, \\
E\Big(H_t^{(k)}\Big) &= H_t\, P\{s(t) = k \mid s(t) \neq 0\}, \\
E\Big(R_t^{(k)}\Big) &= E\Big(S_t^{(k)}\Big) \left(1 - \frac{C_t}{N}\right), \qquad C_t = \sum_{u=0}^{t} H_u,
\end{aligned}$$

其中

$$P\{s(t) = k \mid s(t) \neq 0\} = \frac{P(s(t) = k)}{\sum_{l=1}^{K} P\{s(t) = l\}}.$$

对 $E(R_t^{(k)})$ 的近似隐含地假设：在人群层面，既往住院史与当前感染状态及变异株类型近似独立。在这一平均场假设下，$1 - C_t/N$ 可作为在时间 t仍有首次住院风险的个体比例的代理量，进而按比例应用于当前感染变异株 k的预期个体数。

概率 $P\{s(t) = k\}$ 由感染过程模型通过其转移强度结构确定。具体而言，令

$$\boldsymbol{\rho}(t;\boldsymbol{\theta}_s) = \big(\rho_1(t;\boldsymbol{\theta}_s), \ldots, \rho_K(t;\boldsymbol{\theta}_s)\big)$$

表示时间 t 的状态占据概率向量，则有

$$P\{s(t) = k\} = \rho_k(t;\boldsymbol{\theta}_s),$$

其中，$\boldsymbol{\rho}(t;\boldsymbol{\theta}_s)$ 是通过对连续时间转移强度矩阵进行离散时间近似，并将初始状态分布向前传播得到的。

当仅能获得 S_t^* 关于 S_t 的下界时，我们在二项分布模型下用其条件期望近似 S_t，

$$S_t \sim \text{Binomial}\{N, P(s(t) \neq 0)\},$$

从而得到

$$E(S_t \mid S_t \geq S_t^*) \approx \mu_s(t) + \sigma_s(t)\frac{\phi((S_t^* - \mu_s(t))/\sigma_s(t))}{1 - \Phi((S_t^* - \mu_s(t))/\sigma_s(t))},$$

其中 $\mu_s(t) = NP(s(t) \neq 0)$，且 $\sigma_s^2(t) = NP(s(t) \neq 0)\{1 - P(s(t) \neq 0)\}$。这里，$\phi(\cdot)$ 和 $\Phi(\cdot)$ 分别表示标准正态分布的概率密度函数和累积分布函数，它们源自对二项分布的正态近似。

伪似然。代入这些基于期望的近似，得到伪似然：

$$L^{\text{PL}}(\boldsymbol{\theta}_s, \boldsymbol{\theta}_w, \boldsymbol{\theta}_h) = \prod_{t=0}^{E} L_h\Big(E(H_t^{(k)}), E(R_t^{(k)}); \boldsymbol{\theta}_h\Big)\, L_w\Big(E(S_t^{(k)}); \boldsymbol{\theta}_w\Big)\, L_{S^*}(S_t^*; \rho(t;\boldsymbol{\theta}_s))\,,$$

其中，所有潜在变异株特异性量均由其条件期望替代，而对 $\boldsymbol{\theta}_s$ 的依赖通过感染状态概率 $\rho_k(t;\boldsymbol{\theta}_s)$ 引入。分量 $L_{S^*}(S_t^*;\ \rho(t;\boldsymbol{\theta}_s))$ 根据用于潜在感染人数 $S_t \sim$ 的二项工作模型构造：$S_t \sim \text{Binomial}\Big(N, \sum_{k=1}^{K} \rho_k(t;\ \boldsymbol{\theta}_s)\Big)$。由于报告病例数 S_t^* 系统性地低估真实感染，因此我们将 S_t^* 视为提供单侧约束，而非精确观测。具体而言，当报告病例数被视为存在漏报时，其贡献形式为

$$L_{S^*}(S_t^*; \rho(t;\boldsymbol{\theta}_s)) = \Pr(S_t \geq S_t^* \mid \rho(t;\boldsymbol{\theta}_s)),$$

在报告病例数被假定为完整的时期，则使用常规的二项概率质量函数。这种构造使报告病例数能够为感染率提供信息，同时不会将其解读为真实感染过程的无偏实现。

通过最大化由此得到的伪对数似然，获得参数估计值。虽然可以利用伪似然的观测费舍尔信息进行方差估计，但我们在第4节的模拟研究表明，这种方法往往会低估不确定性。因此，我们建议采用参数自助法程序量化不确定性；第4节的实证结果表明，在本文考察的汇总与近似方案下，该方法能够提供稳定的不确定性估计。在每次自助抽样重复中，根据拟合模型生成个体层面数据，将其重新汇总至观测数据尺度，然后重新拟合模型；所得参数估计值的经验变异性用于近似计算标准误。数据生成机制与模拟研究中使用的机制相同，模型参数固定为其拟合值（见补充材料）。

4 仿真研究

4.1 研究动机与设计

本模拟研究评估了所提出的伪似然方法在现实监测限制下的有限样本性能。设计主要考虑两项挑战。首先，感染轨迹无法在个体层面观测到，因此推断必须依赖汇总数据。其次，由于检测不完整和监测实践不断变化，报告病例数可能会大幅低估真实感染数。

为反映这些特点，数据根据第 2 节所述的个体层面联合模型生成，但估计仅使用汇总数据。通过两个参数引入漏报。参数 r_1 表示报告病例数可视为近乎完整的时间点所占比例，而 $1-r_1$对应存在漏报的时期。参数 r_2表示漏报时期感染者中的报告率。此设定使我们能够考察一系列监测情景，从病例报告频繁且可靠的情况，到报告长期受限且漏检严重的情况。

4.2 仿真设置

我们考虑一个规模为 $N = 100,000$ 的封闭人群，在长度为 $E = 200$ 天的研究期间进行观察，其中包含 $K = 2$ 种遗传变异株。个体层面数据根据第 2 节所述的联合模型生成，该模型由多状态感染过程、条件病毒排出过程和感染依赖的住院风险率组成。参数值经过选择，以形成两个部分重叠且住院风险各不相同的流行波。

为评估漏报的影响，我们考虑由 $(r_1,\ r_2)$ 索引的四种监测情景。参数 $r_1 \in \{0.2, 0.8\}$ 表示报告病例数近乎完整的时间点所占比例，而 $r_2 \in \{0.2, 0.8\}$ 表示漏报时期感染者的报告率。由此得到四种组合，

$$(r_1, r_2) \in \{(0.2, 0.2), (0.2, 0.8), (0.8, 0.2), (0.8, 0.8)\},$$

其感染率的漏报程度从轻微到严重不等。这些数值旨在涵盖一系列合理的监测条件，从检测基本易于获取且报告率较高的情境，到检测长期受限且个体层面报告率较低的情景。

对于每种情景，我们进行 200 次相互独立的模拟重复。在每次重复中，先模拟个体层面数据，再将其汇总，以获得每日病例数、住院人数和污水病毒信号。随后，仅使用这些汇总摘要估计模型参数。数据生成机制和参数取值的完整细节见补充材料 S1。

4.3 估计程序

对于每个模拟数据集，我们采用两种方法估计参数：(i) 朴素方法采用第 3.2 小节中的伪似然框架，但在所有时间点都将报告病例数 S_t^* 视为等于感染者的真实人数 S_t。也就是说，该方法不考虑漏报，并在似然构造中直接以 S_t^* 替代 S_t；(ii) 所提出的伪似然方法会明确考虑漏报。

对于每种方法，我们采用两种方法量化不确定性：基于伪似然的观测费舍尔信息逆矩阵计算的标准误，以及参数自助法标准误。自助法程序基于从拟合模型生成的 200 个重抽样数据集。

4.4 结果

结果基于 200 次模拟重复进行汇总。表 1 按每种漏报情景列出了参数估计值的平均值和经验标准差。结果同时展示了朴素方法和所提方法，以及两种标准误估计值。

在所有情景中，忽略漏报都会导致感染和住院参数出现系统性偏差。漏报程度较轻时，偏差较小；但随着漏报频率（r_1）或严重程度（r_2）增加，偏差会变得显著。相比之下，所提伪似然方法在所有情景下均能准确恢复参数，包括漏报最严重的情形。值得注意的是，在情景（r_1，r_2）$=$（0。8，0。8）下的表现，与漏报机制更严重时观察到的表现相当；该情景对应轻度漏报和接近完全的病例查明。这表明，即使漏报极少，所提方法仍能保持稳定；当真实报告过程更接近完整观测时，其表现也不会下降。

关于不确定性量化，基于费舍尔信息矩阵逆矩阵计算的标准误持续低估估计量的经验变异性。参数自助法标准误与所有参数和情景下的经验标准差都十分接近。基于这些发现，后续真实数据分析采用基于自助法的推断。

5 案例研究：渥太华的 COVID-19 监测

我们将所提出的联合建模框架应用于加拿大渥太华的 COVID-19 监测数据，旨在展示在病例查明不完整的情况下，污水病毒信号与住院记录如何通过潜在感染动态关联起来。分析聚焦于从奥密克戎前变异株向奥密克戎及其亚变异株过渡的时期；在此期间，临床检测实践和报告完整性发生了显著变化。

可用的监测数据由在人群层面观测到的三个聚合时间序列组成：每日污水病毒信号、每日新增 COVID-19 病例、相关住院人数以及报告的现存病例数。图1展示了这三个序列随时间的变化，分别对应监测过程、住院过程和报告感染过程的观测实现值。报告病例曲线中的虚线段表示检测可及性有限的时期，而竖直虚线则标记了分析中使用的人群层面时期指标；这些指标根据污水等位基因比例推断出的变异株优势地位，并依据文献中的既定标准加以定义（Graber 等（2021））。这些时期的详细定义，包括具体日期和阈值，见补充材料。

在整个案例研究中，我们在基线分析中考虑两个主要变异株组，分别以 $k = 1,2$ 编号，对应于奥密克戎前变异株和奥密克戎（包括早期亚变异株）。具体而言，奥密克戎前组包括野生型、阿尔法和德尔塔变异株，而奥密克戎组包括奥密克戎 BA.1、BA.2 和 BA.$3+$。变异株组成的时间演变受到根据监测数据推断出的已知优势时期的约束，同时允许存在多个变异株可能共同传播的过渡窗口期。变异株混合设定的详细信息见第 2 节和补充材料 S2。

由于现有监测数据中无法观察到个体层面的康复时间，因此无法将从感染到康复的转移强度 γ_{k0} 与感染强度分别识别。为施加必要的可识别性约束，我们将所有变异株组的 $\gamma_{k0} = 0.07$ 固定为同一值。该值对应的平均传染期约为 14 天，与 SARS-CoV-2 传染期的常见估计一致（He 等（2020））。重要的是，固定 γ_{k0} 主要决定潜在感染过程的总体规模，而感染、污水病毒信号与住院人数之间的时间模式及估计关系基本保持不变。

尽管整个研究期间都常规收集污水信号和住院人数，但众所周知，报告病例数会大幅低估真实感染数，尤其是在检测政策发生变化之后。为反映这一现实，我们的分析明确考虑病例查明的不确定性，而不将报告病例视为感染率的无偏测量。

5.1 病例 ascertainment 情景分析

将污水病毒信号和住院数据与潜在感染动态联系起来的一项核心挑战，是报告的现有病例数存在不确定性。检测的变化

可及性和资格标准随时间变化，意味着报告病例通常只是对真实感染人数的下限估计，而非完整统计。为评估这种不确定性对推断的影响，我们开展了情景分析，对报告的现有病例 S_t^* 与感染者真实人数 S_t之间关系的不同假设进行形式化。

我们引入一个二元指标 $L(t)$，用于标示检测可及性较高的时期。当 $L(t) = 1$时，假设检测广泛可及，并将报告病例视为近乎完整；当 $L(t) = 0$ 时（对应图1中报告病例曲线的虚线部分），假设报告病例低估了真实感染。该指标根据安大略省公开的检测政策变化信息构建，仅用于界定第5.1小节中的病例查明情景，而不作为直接的建模输入。

我们考虑三种情景：

- **情景1（完全可观测）**。假设报告的现有病例在所有时点均等于真实感染，$S_t^* = S_t$。该情景作为基准，对应于忽略漏报的分析。

- **情景 2（政策信息辅助的可观测性）**。假设仅在检测可及性较高的时期，报告病例才是完整的：

$$S_t^* = S_t \text{ if } L(t) = 1, \qquad S_t^* < S_t \text{ if } L(t) = 0.$$

  该情景纳入了有关检测政策的外部信息。

- **情景 3（持续漏报）**。假设在整个研究期间，报告病例都低估了真实感染，$S_t^* < S_t$ 对所有 t均如此。该情景代表最保守的设定。

在每种情景下，均使用第3节中的同一伪似然估计程序；不同情景的区别仅在于如何将报告病例数解释为有关潜在感染率的信息。情景1对应于将报告病例视为完全观测到的感染（$S_t^* = S_t$）这一特殊情形，而情景2和情景3则放宽了这一假设，分别将 S_t^* 视为在政策规定的时期内或整个研究期间 S_t 的下限。如模拟研究（第4节）所示，即使仅间歇性地获得可靠的病例信息，所提方法仍保持稳定。

其他变异株分组。为展示模型的灵活性，我们还拟合了其他变异株混合结构，以进一步区分奥密克戎亚变异株（例如，将 BA.1 与后续的奥密克戎谱系分开）。相应的设定和结果见补充材料 S2。

5.2 真实数据分析结果

5.2.1 估计感染率和变异株组成

图2展示了三种病例确认情景下估计的感染率随时间的变化，以及报告的现存病例数。结果按

两个变异株组进行分层，并允许存在共同传播时期。报告病例曲线中的虚线段对应于情景2和情景3中视为漏报的时期，定义见第5.1小节。

在所有情景下，推断出的感染比例均能捕捉观测到的流行波的发生时间和总体形状。在过渡时期，估计的状态概率表明多种变异株同时传播，且变异株之间呈平缓过渡，而非突然取代。

不同病例查明情景之间的差异主要影响感染率的估计水平，而时间变化模式保持稳定。当将报告病例视为漏报时（情景 2 和情景 3），推断出的感染率始终高于观测计数，同时仍保留流行高峰的发生时间和相对幅度。假设病例完全查明（情景1），所得估计值则与报告数据更为接近。

图 3 通过比较观测到的污水病毒信号和每日住院人数与政策信息情景（情景 2）下拟合模型模拟出的轨迹，对模型进行了补充评估。对于每种结果，均使用拟合模型的参数估计值生成 100 个重复数据集。在两种情况下，观测时间序列均落在模拟轨迹的 95% 中央范围内，这表明联合模型能够充分捕捉这两种监测过程的幅度、时间变异性和主要高峰。虽然此处仅展示了住院总人数，但拟合模型通过潜在感染过程区分了变异株特异性住院贡献；相应的变异株分层结果，以及在其他病例查明情景下的重复分析结果，均见补充材料。

5.2.2 参数估计值

表 2 报告了三种病例确认情景下病毒载量模型的参数估计值。结果分别针对两个变异株组（$k = 1$ 奥密克戎前期和 $k = 2$ 奥密克戎）给出，并按用作群体层面协变量的时期指标分层；所列变异株（例如 BA.1）有所重叠，反映了两个组之间允许共同传播。报告的均值和方差列是根据估计的分布参数计算得出的派生量，表示每次感染每天对污水病毒信号的平均贡献。

各情景下，变异株间的对比结果一致。在奥密克戎前组中，与德尔塔/BA.1 基线相比，野生型变异株对应的推断污水信号均值较高，而阿尔法与基线相近。在奥密克戎组中，BA.2 和 BA.3 的回归系数远低于 BA.1，这意味着在原始尺度上推断的均值信号更高。这些发现的方向与既往污水监测研究的结果一致；相关研究记录了德尔塔占主导时期 SARS-CoV-2 排毒率较高，而奥密克戎时期排毒率较低（例如 Prasek等（2023）），尽管估计目标以及群体层面与个体层面的解释存在差异。

病例发现假设主要影响估计的均值和方差参数的总体规模，而非变异株的相对排序。随着漏报假设变得更加严格（从情景1到情景 3），估计的每次感染

污水信号会降低，而变异株对比则保持稳定。

表3报告了三种病例 ascertainment 情景下的住院风险参数估计值，以及相应的首次住院时间风险率。报告的风险率是由估计的模型参数计算得出的派生量。

在所有情景下，奥密克戎变异株组的基线住院风险率均显著低于奥密克戎前组。在奥密克戎组内，与 BA.1 相比，BA.2 和 BA.3 始终对应更高的住院风险率，其中 BA.3 的风险率最高。在奥密克戎前组中，与德尔塔/BA.1 基线相比，阿尔法的住院风险率较高，而野生型变异株的差异较小。

6 讨论

本研究构建了一个建模框架，通过潜在感染过程联合分析污水病毒信号和住院数据。该框架的构建缘于临床监测的结构性局限，以及基于污水的传染病监测日益重要。尽管模型是在个体层面构建的，推断仅使用汇总的群体层面数据进行，这反映了大多数现实世界监测系统的数据结构。本研究的主要贡献在于展示了如何将这一框架付诸实践，从感染无法被完全观测且缺乏个体层面数据的情境中提取可解释的流行病学信号。具体而言，该框架凸显了环境监测如何在临床病例数系统性漏报时，支持对社区层面疾病负担的推断。

本研究要解决的一个核心挑战，是缺乏个体层面的感染轨迹。通过将感染视为未观测到的驱动过程，并在给定感染的条件下将污水信号与住院情况建模为条件独立，该框架能够利用这些互补的数据来源共同揭示传播动态。这种结构避免了在环境监测与临床指标之间强加直接的机制联系，因为这种联系很可能随时间变化而不稳定；相反，该框架将二者的关联归因于不断变化的感染率和变异株组成。

第二个重大挑战是汇总临床病例数数据存在系统性漏报。我们没有将报告的临床病例数视为感染率的无偏观测，而是通过情景分析明确考虑了病例未被完整识别的情况，并利用有关检测可及性的外部信息为分析提供依据。这一设定反映了现实中的监测情境：临床病例不太可能被完整识别，尤其是在新型病原体出现期间；这种情况并非只是暂时的局限，可以通过改进检测加以解决。通过区分检测普及时期与检测受限时期，该模型将报告的临床病例数视为潜在感染过程的不完整观测，而非精确测量。模拟结果和渥太华案例研究均表明，在各种漏报假设下，这种方法都能得到稳定的推断，同时保留关键的时间模式和变异株特异性模式。

若干局限　　值得考虑。由于推断依赖于汇总数据，

模型无法捕捉与年龄、合并症、疫苗接种状态或既往免疫力相关的个体层面异质性。此外，当完全无法观测到感染率时，若没有外部信息，推断出的感染绝对水平便无法识别。虽然所提方法能够恢复相对时间模式并进行变异株对比，但若要识别绝对感染负担，则需要纳入更多数据来源，例如血清流行率研究或经过校准的感染估计值。另一个相关局限是病例查明方面的不确定性。未来可通过更多数值研究，考察在漏报假设设定错误时，所提方法的效率和稳健性；这部分工作留待后续开展。在实际监测场景中，模型诊断也是一项值得关注的实务问题。正式的诊断工具对于评估模型适用性十分重要，但用于常规监测的模型必须受限于数据可用性和分析复杂度。因此，在公共卫生监测系统中部署潜在感染模型时，如何开发兼顾统计严谨性与简便性的诊断方法，是未来研究的一个重要方向。

从应用角度看，对于临床病例确定不完整或不断变化的监测场景，本文提出的框架尤其适用。例如，在检测能力有限的疫情早期阶段，以及常规病例报告减少或停止后的后期监测中，都可以使用该框架。对潜在感染过程进行明确建模，还使该框架能够适应多个变异株或毒株共同传播的时期。尽管本文以 COVID-19 数据为例进行说明，但该框架也广泛适用于其他传染病：在这些疾病中，污水监测能够提供有意义的群体层面信号，而临床结局仅以汇总形式观测；只需进行适当调整，以反映病原体特异性传播动力学、排 shedding 行为及可用监测数据即可。

未来研究可从多个方向拓展这一框架。借助更丰富的数据——无论来自污水、血清学研究还是其他群体层面数据源——感染过程和住院过程都可以纳入更灵活的转移结构、延迟或协变量效应，从而更好地反映临床和行为异质性。从监测角度看，将实时废水监测与短期即时预测或预测相结合，或可进一步提升其在公共卫生决策中的效用，尤其是在临床检测能力有限或快速变化的时期。更广泛地说，针对严重汇总和漏报情况下的潜在感染过程开发推断策略，仍是方法学研究的重要领域。

致谢

本研究部分由加拿大自然科学与工程研究委员会（NSERC）向 Hu 和 Swartz 提供的资助，以及由 Swartz 领导的加拿大统计科学研究所（CANSSI）体育分析 CRT（协作研究团队）提供的资助支持。作者感谢山东鲁能泰山足球俱乐部前技术总监 Daniel Stenz 提供本文所讨论的数据。

声明

作者声明不存在利益冲突。

A 图

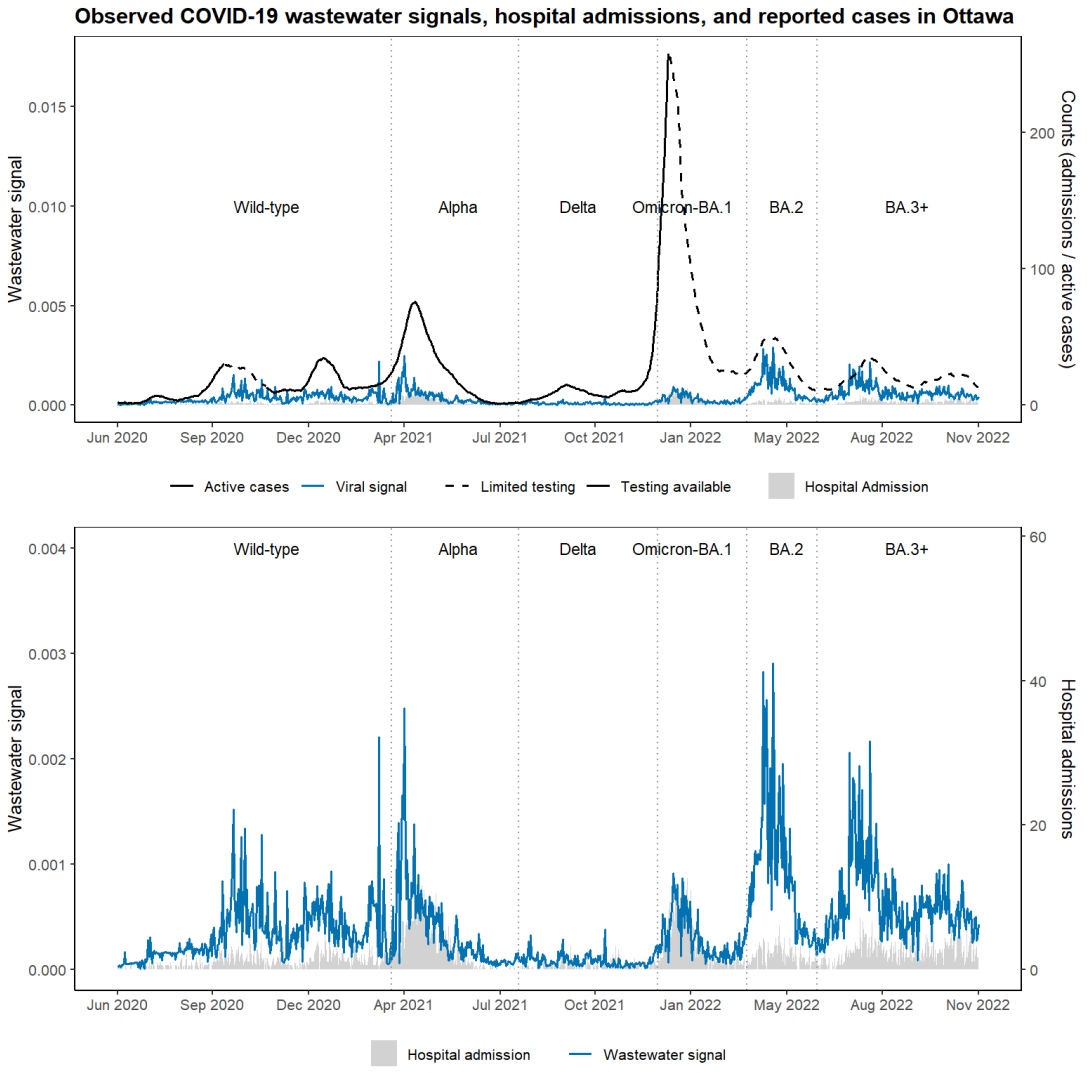

图1：研究期间渥太华的污水病毒信号、每日新增 COVID-19 住院人数和报告的 COVID-19 现存病例。为便于与污水信号和住院人数进行比较，现存病例数按原始值的 1% 重新缩放。现存病例曲线中的虚线段表示检测可及性受限的时期；在情景分析中，报告病例被视为少报。

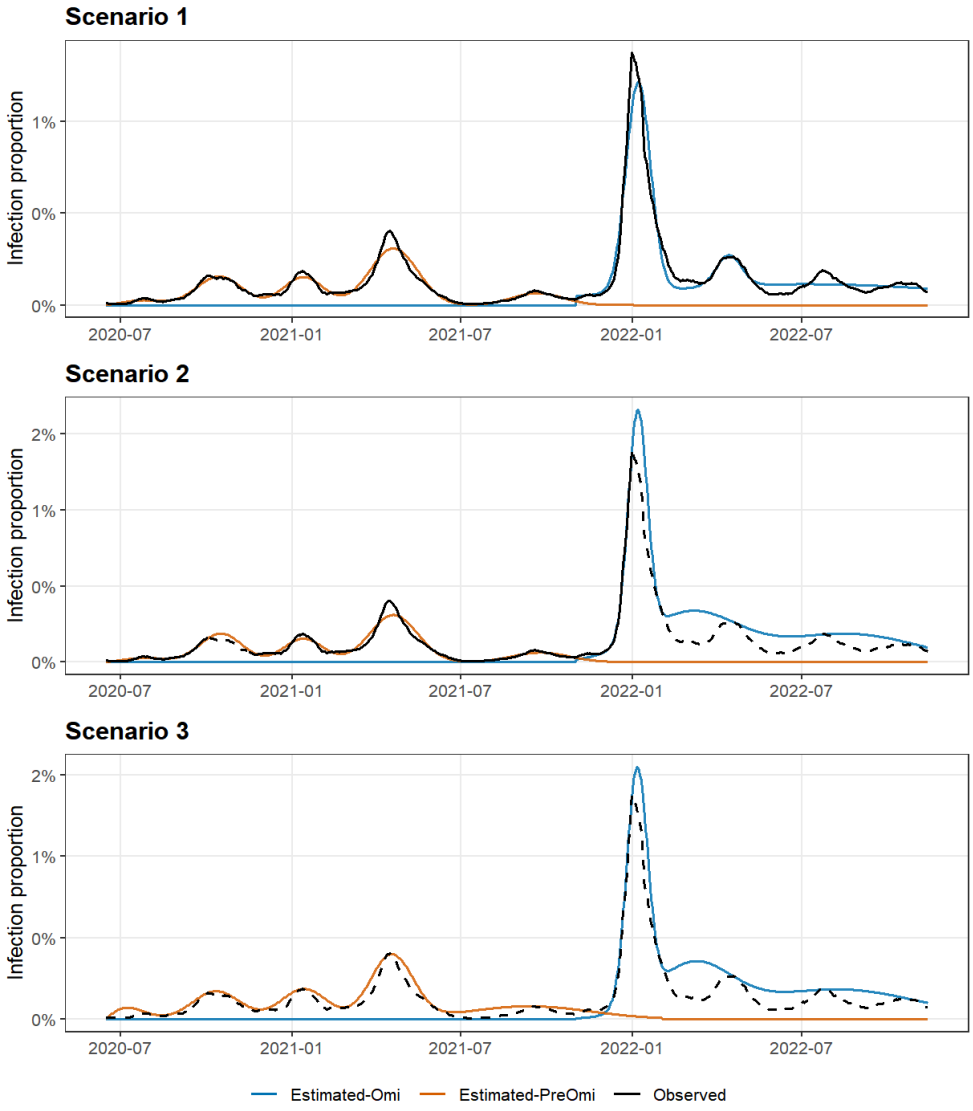

图2：第5.1节所述三种病例查明情景下随时间变化的感染率估计值，以及报告的现存病例数。两种变异株组以不同颜色的线条表示，并显示其共同传播时期。报告病例曲线中的虚线段表示被视为存在漏报的时期。

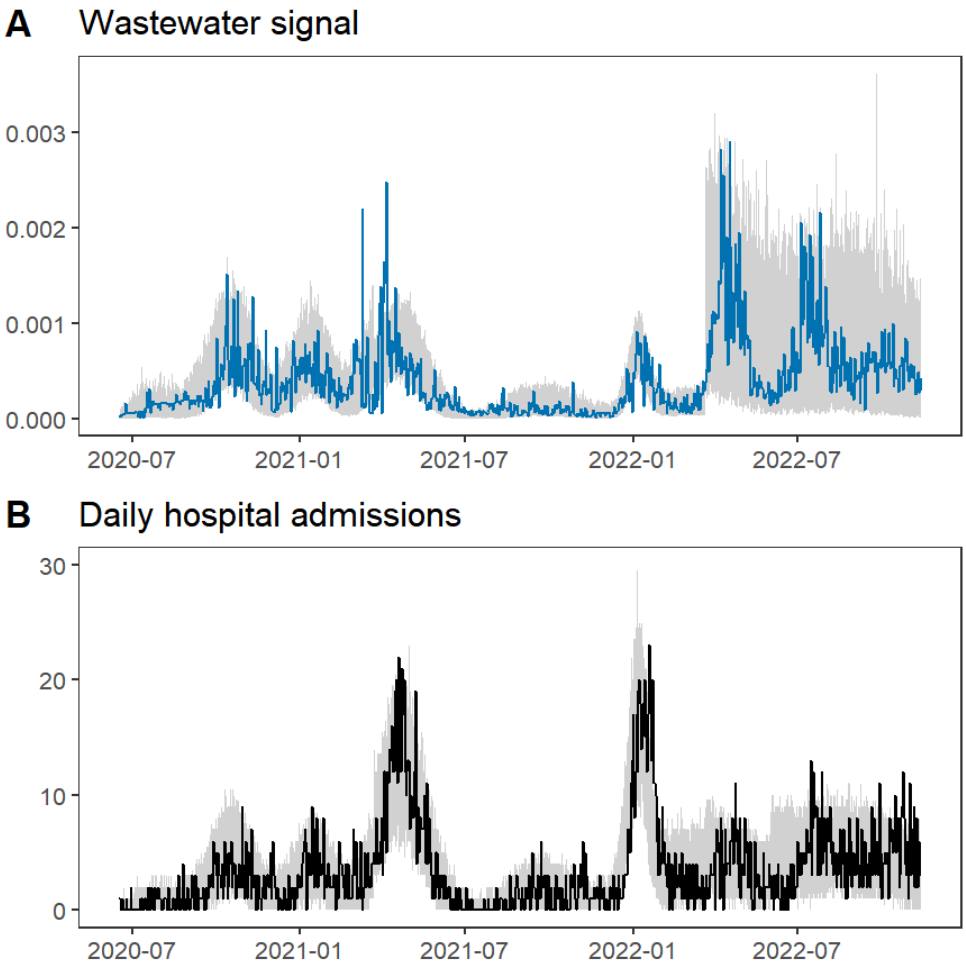

图 3：在政策信息引导的病例查明情景（情景 2）下，观测到的污水病毒信号和每日住院人数与拟合的联合模型模拟的轨迹对比。阴影带表示在给定拟合参数的条件下生成的 100 条模拟轨迹的 95% 中央范围，反映的是观测模型引起的变异，而非推断不确定性。

B 表格

(a) (r_1, r_2) =(0.8,0.8)

	True	忽略低报		所提方法			
参数		$\overline{\text{估计值}_{\text{naive}}}$	SD(估计值$_{\text{naive}}$)	$\overline{\text{EST}}$	SD(估计值)	$\overline{\hat{SE}_1}$	$\overline{\hat{SE}_2}$
λ_1 (10^{-3})	2.00	2.23	0.38	1.95	0.14	0.07	0.14
λ_2 (10^{-3})	5.00	5.44	1.25	4.55	0.25	0.13	0.29
α_1 (10^{-3})	1.00	1.68	1.59	0.96	0.41	0.09	0.34
β_1 (10^4)	1.00	1.69	1.66	0.98	0.49	0.10	0.41
α_2 (10^{-3})	5.00	5.24	4.88	4.72	1.45	0.53	1.20
β_2 (10^4)	2.00	2.22	1.96	1.92	0.63	0.22	0.59

(b) (r_1, r_2) =(0。2，0。8)

	True	忽略低报		所提方法			
参数		$\overline{\text{估计值}_{\text{naive}}}$	SD(估计值$_{\text{naive}}$)	$\overline{\text{EST}}$	SD(估计值)	$\overline{\hat{SE}_1}$	$\overline{\hat{SE}_2}$
λ_1 (10^{-3})	2.00	2.66	0.24	1.96	0.14	0.07	0.15
λ_2 (10^{-3})	5.00	5.67	0.51	4.51	0.26	0.13	0.32
α_1 (10^{-3})	1.00	1.06	0.35	0.96	0.45	0.11	0.42
β_1 (10^4)	1.00	0.92	0.33	1.04	0.52	0.13	0.50
α_2 (10^{-3})	5.00	4.91	1.89	4.86	1.86	0.55	1.32
β_2 (10^4)	2.00	1.69	0.64	1.95	0.77	0.24	0.66

(c) (r_1, r_2) =(0.8,0.2)

	True	忽略低报		所提方法			
参数		$\overline{\text{估计值}_{\text{naive}}}$	SD(估计值$_{\text{naive}}$)	$\overline{\text{EST}}$	SD(估计值)	$\overline{\hat{SE}_1}$	$\overline{\hat{SE}_2}$
λ_1 (10^{-3})	2.00	2.55	0.25	2.08	0.20	0.07	0.17
λ_2 (10^{-3})	5.00	5.52	0.31	4.78	0.38	0.13	0.31
α_1 (10^{-3})	1.00	1.23	0.34	1.04	0.36	0.10	0.41
β_1 (10^4)	1.00	1.02	0.31	1.03	0.38	0.12	0.44
α_2 (10^{-3})	5.00	4.74	1.04	4.80	1.52	0.51	1.46
β_2 (10^4)	2.00	1.62	0.36	1.89	0.63	0.22	0.57

(d) (r_1, r_2) = (0.2，0.2)

	True	忽略低报		所提方法			
参数		$\overline{\text{估计值}_{\text{naive}}}$	SD(估计值$_{\text{naive}}$)	$\overline{\text{EST}}$	SD(估计值)	$\overline{\hat{SE}_1}$	$\overline{\hat{SE}_2}$
λ_1 (10^{-3})	2.00	5.99	0.51	2.20	0.22	0.07	0.22
λ_2 (10^{-3})	5.00	13.06	0.92	4.93	0.41	0.14	0.33
α_1 (10^{-3})	1.00	2.78	1.30	1.04	0.34	0.12	0.32
β_1 (10^4)	1.00	1.14	0.60	0.96	0.37	0.13	0.32
α_2 (10^{-3})	5.00	2.09	1.60	5.07	1.82	0.58	1.49
β_2 (10^4)	2.00	0.39	0.31	2.09	0.69	0.23	0.52

表 1：第4.4节四种（r_1，r_2）情景下的模拟结果。

时期	β，估计值（标准误）	均值，估计值（标准误）	方差，估计值（标准误）
情景1：假设观测到的 S_t 是真实感染人数			
奥密克戎出现前时期（形状参数：***0.004***（***0.0002***））			
截距 β_0:	9.57(0.094)		
Delta与BA.1（基线）	.	28。1(2。21) $\times 10^{-8}$	19.5(3.22) $\times 10^{-12}$
Wild	−0.51(0.0860.0860.086)( )( )	46.7(1.65) $\times 10^{-8}$	54.0(4.68) $\times 10^{-12}$
阿尔法	−0.01(0。088)	27.8(1.07) $\times 10^{-8}$	19.1(1.77) $\times 10^{-12}$
奥密克戎时期（形状参数：***0.001 (0.0001)***）			
截距 β_0:	9.79(0.079)		
BA.1（基线）	.	59。7(2。69) $\times 10^{-9}$	33.3(3.71) $\times 10^{-13}$
BA.2	−2.27(0.1010.1010.101)( )( )	57.5(5.17) $\times 10^{-8}$	31.0(5.91) $\times 10^{-11}$
BA.3	−2.35(0.0880.0880.088)( )( )	62。7(4。74) $\times 10^{-8}$	36.8(6.05) $\times 10^{-11}$
情景 2：考虑 $L(t)$			
奥密克戎出现前时期（形状参数：***0.004 (0.0002)***）			
截距 β_0:	9.63(0.092)		
Delta与BA.1（基线）	.	26.3(2.02) $\times 10^{-8}$	17.3(2.80) $\times 10^{-12}$
Wild	−0.51(0.0840.0840.084)( )( )	44。0(1。51) $\times 10^{-8}$	48.3(4.11) $\times 10^{-12}$
阿尔法	−0.05(0.084)	27.8(1.08) $\times 10^{-8}$	19。2(1。78) $\times 10^{-12}$
奥密克戎时期（形状参数：***0.001 (0.0001)***）			
截距 β_0:	9。99(0。078)		
BA.1（基线）	.	45.6(1.86) $\times 10^{-9}$	20.8(2.20) $\times 10^{-13}$
BA.2	−2.16(0.0880.0880.088)( )( )	39。6(3。06) $\times 10^{-8}$	15.7(2.64) $\times 10^{-11}$
BA.3	−2.17(0.0750.0750.075)( )( )	40.0(2.51) $\times 10^{-8}$	16.1(2.28) $\times 10^{-11}$
情景 3：考虑 S_t 始终低于真实感染			
奥密克戎出现前时期（形状参数：***0.003 (0.0001)***）			
截距 β_0:	9.97(0.083)		
德尔塔与BA.1（基线）	.	12.4(0.80) $\times 10^{-8}$	5.79(0.81) $\times 10^{-12}$
Wild	−1.06(0.0750.0750.075)( )( )	35.8(1.36) $\times 10^{-8}$	48.3(4.47) $\times 10^{-12}$
阿尔法	−0.55(0.0770.0770.077)( )( )	21.4(0.90) $\times 10^{-8}$	17.2(1.71) $\times 10^{-12}$
奥密克戎时期（形状参数：***0.001***（***0.0001***））			
截距 β_0:	10。17(0。077)		
BA.1（基线）	.	46.5(1.73) $\times 10^{-9}$	17.7(1.77) $\times 10^{-13}$
BA.2	−2.14(0.0790.0790.079)( )( )	39.5(2.76) $\times 10^{-8}$	12。8(1。98) $\times 10^{-11}$
BA.3	−2.17(0.0680.0680.068)( )( )	40.8(2.33) $\times 10^{-8}$	13。6(1。80) $\times 10^{-11}$

表 2：三种病例 ascertainment 情景下参数估计值的比较。第一列报告回归系数的估计值 β，括号内为标准误。第二列报告变异株特异性伽马分布的均值估计值，该值通过代入估计的形状参数和率参数获得，标准误采用多元 delta 方法计算。第三列报告伽马分布对应的方差估计值，标准误采用类似方法计算。情景1假设报告的现有病例 S_t 等于感染的真实数量；情景 2通过 $L(t)$考虑随时间变化的检测可及性；情景 3假设在整个研究期间病例持续漏报。

时期	λ，估计值（标准误）	风险率，估计值（标准误）
情景1		
奥德赛出现前时期		
截距 λ_0:	−5.77 (0。090)	
Delta与BA.1（基线）	.	0.0031 (0.0003)
Wild	−0.063 (0.098)	0.0029 (0.0001)
阿尔法	0。302 0。302 0。302 (0。0970。0970。097)( )( )	0.0042（0.0002）
奥密克戎时期		
截距 λ_0:	−6.53 (0.038)	
BA.1（基线）	.	0.0015 (0.0001)
BA.2	0.437 0.437 0.437 (0.0730.0730.073)( )( )	0.0023 (0.0001)
BA.3	1。146 1。146 1。146 (0。0530。0530。053)( )( )	0.0046 (0.0002)
情景 2		
奥密克戎出现前时期		
截距 λ_0:	−5.74 (0.084)	
Delta与BA.1（基线）	.	0.0032 (0.0003)
Wild	−0.14 (0.095)	0.0028 (0.0001)
阿尔法	0.283 0.283 0.283 (0.0940.0940.094)( )( )	0.0042 (0.0002)
奥密克戎时期		
截距 λ_0:	−6.81 (0.039)	
BA.1（基线）	.	0.0011 (0.0001)
BA.2	0.356 0.356 0.356 (0.0730.0730.073)( )( )	0.0016 (0.0001)
BA.3	0.989 0.989 0.989 (0.0530.0530.053)( )( )	0.0029 (0.0001)
情景 3		
奥密克戎出现前时期		
截距 λ_0:	−6.27 (0。076)	
Delta与BA.1（基线）	.	0.0019 (0.0001)
Wild	0.171 (0.086)	0.0022 (0.0001)
阿尔法	0.541 0.541 0.541 (0.0850.0850.085)( )( )	0.0033 (0.0001)
奥密克戎时期		
截距 λ_0:	−6.85 (0.040)	
BA.1（基线）	.	0.0011 (0.0001)
BA.2	0.384 0.384 0.384 (0.0740.0740.074)( )( )	0.0016 (0.0001)
BA.3	1.038 1.038 1.038 (0.0540.0540.054)( )( )	0.0030 (0.0001)

表3：三种病例查明情景下住院强度参数估计值的比较。第一列报告回归系数$\boldsymbol{\lambda}$ 的估计值，括号内为标准误。第二列报告相应的住院时间风险率估计值，该估计值通过将拟合的强度函数代入基线协变量值计算得出，标准误则通过Delta法得出。

参考文献

Albani, V., Loria, J., Massad, E. & Zubelli, J. (2021), 'COVID-19漏报及其对疫苗接种策略的影响', *BMC*传染病 **21**, 1–13.

Banks, H.、Meade, A. E.、Schacht, C.、Catenacci, J.、Thompson, W. C.、Abate-Daga, D. 与 Enderling, H.（2020），《利用汇总数据进行参数估计》，《应用数学快报》**100**，105999。

Chen, P. 与 Ye, Z.-S.（2017），《基于汇总寿命数据的现场可靠性估计》，《计量技术》**59**(1)，115–125。

D'Aoust, P. M., Tian, X., Towhid, S. T., Xiao, A., Mercier, E., Hegazy, N., Jia, J.- J., Wan, S., Kabir, M. P., Fang, W. 等（2022），《污水与临床病例（WC）比率：COVID-19可识别城市社区临床检测不足、关注变异株出现及人群免疫》，《全面环境科学》**853**, 158547. Davis, B., Heathcote, C. R. & O'neill, T.（2002），《根据汇总数据估计和插值马尔可夫链》，《生物统计学》**89**(1), 95–110. Galani, A., Aalizadeh, R., Kostakis, M., Markou, A., Alygizakis, N., Lytras, T., Adamopou- los, P. G., Peccia, J., Thompson, D. C., Kontou, A. 等（2022），《SARS-CoV-2污水监测数据可预测住院人数和重症监护病房收治人数》，《全面环境科学》**804**, 150151. Gilani, N., Kazemnejad, A., Zayeri, F., Hadaegh, F., Azizi, F. & Khalili, D.（2017），《人体测量指标作为冠心病风险预测因子：对纵向测量值和事件发生时间进行联合建模》，《伊朗公共卫生杂志》**46**(11), 1546.́Graber, T. E., Mercier, E., Bhatnagar, K., Fuzzen, M., D'Aoust, P. M., Hoang, H.-D., Tian, X., Towhid, S. T., Plaza-Diaz, J., Eid, W. 等（2021），《采用等位基因特异性引物延伸PCR策略，近实时测定污水中B.1.1.7占SARS-CoV-2总病毒载量的比例》，《水研究》**205**, 117681. He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X. 等（2020），《COVID-19病毒排出和传染性的时间动态》，《自然医学》**26**(5), 672–675. Hegazy, N., Cowan, A., D'Aoust, P. M., Mercier, E., Towhid, S. T., Jia, J.-J., Wan, S., Zhang, Z., Kabir, M. P., Fang, W. 等（2022），《了解大流行期间污水中SARS-CoV-2信号与临床指标之间的动态关系》，$medRxiv$。Holcomb, D. A., Christensen, A., Hoffman, K., Lee, A., Blackwood, A. D., Clerkin, T., Gallard-G´ongora, J., Harris, A., Kotlarz, N., Mitasova, H. 等（2024），《通过估计变化率解读疾病趋势的定量污水监测》，《全面环境科学》**951**, 175687.

Jeng, H. A., Diawara, N., Welch, N., Jackson, C., Singh, R., Curtis, K., Gonzalez, R., Jurgens, D. & Adikari, S.（2025）, 《利用高斯联结函数时间序列边际回归模型，基于污水SARS-CoV-2监测预测新冠肺炎病例、住院人数和死亡人数》, *COVID* **5**(2), 25. Kaplan, E. H., Wang, D., Wang, M., Malik, A. A., Zulli, A. & Peccia, J.（2021）, 《通过流行病模型对齐SARS-CoV-2指标：在住院人数和污水污泥中RNA检测方面的应用》, 《卫生保健管理科学》**24**(2), 320–329. Kelton, C. M. & Kelton, W. D.（1987）, 《比较基于微观数据与宏观数据估计的马尔可夫过程假设检验方法》, 《运筹学年鉴》**8**(1), 175–194. Lai, M., Wulff, S. S., Cao, Y., Robinson, T. J. & Rajapaksha, R.（2025）, 《预测中的时间交叉验证：利用污水数据预测新冠肺炎发病率的案例研究》, 《国际质量与可靠性工程》**41**(2), 672–688. Larsen, D. A. & Wigginton, K. R.（2020）, 《利用污水追踪COVID-19》, 《自然-生物技术》**38**(10), 1151–1153. Lau, H., Khosrawipour, T., Kocbach, P., Ichii, H., Bania, J. & Khosrawipour, V.（2021）, 《评估多个全球疫情中心新冠肺炎病例的严重漏报和检测不足》, 《肺病学》**27**(2), 110–115. Lawless, J. & McLeish, D.（1984）, 《马尔可夫链汇总数据中的信息》, 《生物统计学》*style>71*(3), 419–430. Luvanda, H. B., Mukyanuzi, E. N. & Akarro, R. R.（2023）, 《针对坦桑尼亚医疗与治疗中心就诊的艾滋病患者，估计病毒载量结局与死亡生存时间关联的联合生存模型》, 《*BMC*公共卫生》**23**(1), 2091. Meadows, T., Coats, E. R., Narum, S., Top, E. M., Ridenhour, B. J. & Stalder, T.（2025）, 《流行病学模型可根据农村社区的污水监测预测新冠肺炎疫情暴发》, 《水研究》**268**, 122671. Miyazawa, S., Wong, T. S., Ito, G., Iwamoto, R., Watanabe, K., van Boven, M., Wallinga, J. & Miura, F.（2024）, 《基于污水的再生数及对日本三个地区新冠肺炎病例的预测：2021年11月至2022年12月》, 《欧洲监测》**29**(8), 2300277. Mohring, J., Leith¨auser, N., Wlaz-lo, J., Schulte, M., Pilz, M., M¨unch, J. & K¨ufer, K.- H.（2024）, 《根据污水估计新冠肺炎患病率》, 《科学报告》**14**(1), 14384. Nguyen, K., Rutter, E. M. & Flores, K. B.（2023）, 《利用汇总时空数据估计具有竞争项的反应扩散方程的参数分布》, 《数学生物学通报》**85**(7), 62.

Orcutt, G. H., Watts, H. W. & Edwards, J. B.（1968），‘数据聚合与信息损失’，《美国经济评论》**58(4)**，773–787。

Pappu, A. R., Green, A., Oakes, M. & Jiang, S.（2025），‘通过基于污水的监测追踪人口较少社区的 COVID-19 趋势’，《全面环境科学》**970**，179007。

Parkins, M. D., Lee, B. E., Acosta, N., Bautista, M., Hubert, C. R., Hrudey, S. E., Frankowski, K. & Pang, X.-L.（2024），‘基于污水的监测作为公共卫生行动的工具：SARS-CoV-2 及其他’，《临床微生物学评论》**37(1)**，e00103–22。

Peccia, J., Zulli, A., Brackney, D. E., Grubaugh, N. D., Kaplan, E. H., Casanovas-Massana, A., Ko, A. I., Malik, A. A., Wang, D., Wang, M. 等（2020），《测量污水中的 SARS-CoV-2 RNA 可追踪社区感染动态》，*Nature Biotechnology***38(10)**，1164–1167。Peng, K. K., Dean, C. B., Hu, X. J. & Delatolla, R.（2025），《利用马尔可夫调制模型学习新冠肺炎住院与污水病毒信号之间的关联》，《流行病》**52**，100840。Peng, K. K., Renouf, E. M., Dean, C. B., Hu, X. J., Delatolla, R. & Manuel, D. G.（2023），《探讨加拿大渥太华污水病毒信号与新冠肺炎住院之间的关系》，《传染病建模》**8(3)**，617–631。Prasek, S. M., Pepper, I. L., Innes, G. K., Slinski, S., Betancourt, W. Q., Foster, A. R., Ya- glom, H. D., Porter, W. T., Engelthaler, D. M. & Schmitz, B. W.（2023），《污水中变异株特异性 SARS-CoV-2 排放率》，《全面环境科学》**857**，159165。Schoen, M. E., Wolfe, M. K., Li, L., Duong, D., White, B. J., Hughes, B. & Boehm, A. B.（2022），《SARS-CoV-2 RNA 污水沉降固体监测频率及其对使用分布滞后模型预测新冠肺炎发病率的影响》，*ACS ES&T Water*。Sherratt, K., Abbott, S., Meakin, S. R., Hellewell, J., Munday, J. D., Bosse, N., 工作组，C. C.-., Jit, M. & Funk, S.（2021），《探究估计再生数时监测数据的偏差：并深入了解英格兰新冠肺炎的亚群体传播》，《英国皇家学会哲学汇刊*B*辑》**376(1829)**，20200283。Simone, A., Cesaro, A. & Esposito, G.（2024），《从时间序列到可视性算法：研究污水中 SARS-CoV-2 传播的新方法》，《水处理工程杂志》**66**，106107。Starnini, M., Aleta, A., Tizzoni, M. & Moreno, Y.（2021），《数据准确性对新冠肺炎缓解政策评估的影响》，《数据与政策》**3**，e28。Temesgen, A., Gurmesa, A. & Getchew, Y.（2018），《纵向 CD4 计数与 HIV/TB 合并感染患者至死亡时间的联合建模：吉马大学专科医院案例》，《数据科学年刊》**5(4)**，659–678。Tsiatis, A. A. & Davidian, M.（2004），《纵向数据与事件发生时间数据的联合建模：综述》，《中华统计学志》，809–834。

Wu, L., Liu, W. & Hu, X.（2010），《存在测量误差时对 HIV 病毒动力学与免疫抑制进行联合推断》，*Biometrics* **66**(2)，327–335。Wu, L., Liu, W., Yi, G. Y. & Huang, Y.（2012），《纵向数据与生存数据分析：联合建模、推断方法及相关问题》，概率与统计杂志**2012**(1)，640153。Xie, Y., Challis, J. K., Oloye, F. F., Asadi, M., Cantin, J., Brinkmann, M., McPhedran, K. N., Hogan, N., Sadowski, M., Jones, P. D. 等（2022），《市政污水中的 RNA 揭示了由 SARS-CoV-2 关注变异株驱动的四波新冠肺炎疫情暴发规模》，*ACS ES&T Water*。Zelelew, A. N. & Workie, D. L.（2023），《埃塞俄比亚德布雷塔博尔医院高血压患者随访期间血压测量值与心血管疾病并发症生存时间的联合建模》，*Vascular Health and Risk Management*，621–635。Zulli, A., Pan, A., Bart, S. M., Crawford, F. W., Kaplan, E. H., Cartter, M., Ko, A. I., Sanchez, M., Brown, C., Cozens, D. 等（2022），《根据不同污水汇水区的 SARS-CoV-2 RNA 浓度预测每日 COVID-19 病例率》，*FEMS microbes***2**。