·综述·

药物流行病学研究中稀疏数据偏倚的机制、识别与处理

郜 鑫[1,2,3]，龚 硕[3]，邓慧雯[1]，黄 悦[4]，詹思延[1,2,3,5]，裴正存[4]，刘志科[1,2,3]

1. 北京大学公共卫生学院流行病与卫生统计学系（北京 100191）
2. 重大疾病流行病学教育部重点实验室（北京大学）（北京 100191）
3. 北京大学鄞州数智健康联合研究院（浙江宁波 315100）
4. 天津大学医学院（天津 300072）
5. 北京大学第三医院临床流行病学研究中心（北京 100191）

【摘要】药物流行病学研究常使用医疗保险、电子健康记录和登记系统等大规模真实世界数据，但罕见暴露、罕见结局、特殊人群、短风险窗及高维调整可使特定效应估计所需的信息不足。稀疏数据偏倚可导致最大似然估计偏离真值，影响药物安全性信号的解读及有效性判断。本文围绕药物流行病学常见数据结构，阐述稀疏数据偏倚的概念与形成机制，区分其与分离、正值性违背及估计不精确性的关系，从数据结构、估计与推断异常、替代模型及敏感性分析三个层次评价稀疏性及其影响，并比较研究设计优化、Firth 偏倚减少、数据增广、倾向评分及贝叶斯方法的适用条件和局限。处理时应先明确估计目标并评估其数据支持度，再优化设计与协变量结构，并根据稀疏性的来源选择估计方法。报告关键计数、惩罚或先验设定及敏感性分析，有助于判断结论对数据和模型假设的依赖程度。

【关键词】药物流行病学；稀疏数据偏倚；分离；惩罚似然；贝叶斯方法

【中图分类号】R181.3+5 【文献标志码】A

Mechanisms, identification and management of sparse-data bias in pharmacoepidemiologic studies

GAO Xin[1,2,3], GONG Shuo[3], DENG Huiwen[1], HUANG Yue[4], ZHAN Siyan[1,2,3,5], PEI Zhengcun[4], LIU Zhike[1,2,3]

1. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China
2. Key Laboratory of Epidemiology of Major Diseases(Peking University), Beijing 100191, China
3. Peking university-Yinzhou Regional joint institute for Advanced Medical Intelligence and Digital health, Ningbo 315100, Zhejiang Province, China
4. Medical College, Tianjin University, Tianjin 300072, China
5. Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing 100191, China
Corresponding authors: LIU Zhike, Email: liu-zhike@bjmu.edu.cn; PEI Zhengcun, Email: peizhengcun@tju.edu.cn

DOI: 10.12173/j.issn.1005-0698.202609095
基金项目：国家自然科学基金重点项目（82330107）；国家自然科学基金青年科学基金项目（82204175）；国家资助博士后研究人员计划项目（GZC20240064）
通信作者：刘志科，副研究员，博士研究生导师，Email：liu-zhike@bjmu.edu.cn
裴正存，副教授，博士研究生导师，Email：peizhengcun@tju.edu.cn

【Abstract】Pharmacoepidemiologic studies often use large-scale claims data, electronic health records and registries. However, rare exposures or outcomes, special populations, short risk windows and high-dimensional adjustment may leave insufficient information for estimating a specific effect. Sparse-data bias can cause maximum-likelihood estimates to deviate from the true effect, affecting the interpretation of safety signals and treatment effects. This review describes the concepts and mechanisms of sparse-data bias in common pharmacoepidemiologic settings, distinguishes it from separation, positivity violations and imprecision, and evaluates sparsity and its implications through data structure, estimation and inference diagnostics, and alternative models and sensitivity analyses. It compares the assumptions and limitations of design optimization, Firth bias reduction, data augmentation, propensity-score approaches, and Bayesian methods. Management should begin by defining the estimand and assessing its data support, then improve the design and covariate structure, and select an estimation method appropriate to the source of sparsity. Reporting key counts, penalty or prior specifications and sensitivity analyses helps readers assess the extent to which conclusions depend on the observed data and modeling assumptions.

【Keywords】Pharmacoepidemiology; Sparse-data bias; Separation; Penalized likelihood; Bayesian methods

药物流行病学利用电子健康记录、医疗保险理赔、疾病或药品登记等真实世界数据，评价药物的安全性、有效性和利用模式，为临床和监管决策提供依据[1]。在新上市药物、罕见严重不良事件或特殊人群研究中，暴露与结局的组合往往只涉及少量观察；进一步分层或调整协变量后，效应估计可能面临稀疏数据问题。

稀疏数据偏倚可使比值比（odds ratio，OR）、相对危险度（relative risk，RR）或风险比（hazard ratio，HR）等效应估计偏离真实值，其方向与程度取决于数据结构、模型及评价尺度[2-3]。若仅因调整后比值效应估计远离 1 就认定混杂已得到控制，可能误判安全性信号或药物疗效差异。稀疏数据常同时伴随宽置信区间和低检验效能，使研究难以排除有临床意义的风险。这种不确定性可能影响药品说明书修订、临床用药决策及获益与风险评估。

适应证混杂、时间相关偏倚、错分类和选择偏倚已有较系统的方法学讨论[1]，稀疏数据偏倚则容易被等同于小样本或模型不收敛。要避免上述误判，仅检查样本量和模型收敛状态并不足够，还需评估数据能否支持研究问题，以及分析与报告是否充分反映信息不足。本综述的框架与我国 2025 年更新的《药物流行病学研究方法学指南（第 2 版）》中关于明确研究问题、评估数据适用性、合理选择分析方法和透明报告的要求一致[4]。本文结合典型实例，解释药物流行病学设计与分析过程如何产生局部信息不足，并将其来源与诊断及处理方法对应，为研究方案和统计分析计划的制订提供参考。

1 稀疏数据偏倚的概念与产生机制

1.1 定义

稀疏数据偏倚是指暴露、结局与协变量某些水平组合中的观察信息不足，使回归参数的最大似然估计（maximum likelihood estimation，MLE）在有限样本中系统性偏离真值的现象[2-3]。这种数据稀疏性可以影响回归系数及其指数变换后的效应值估计。需注意，偏倚被定义为效应估计在重复抽样下的期望与目标参数真值之差。仅凭一次研究的单元格计数或点估计，不能判定是否存在稀疏数据偏倚，也不能确定其方向和大小；偏倚的表现还取决于数据结构、模型和评价尺度。

1.2 产生机制

稀疏数据偏倚与支持特定参数估计的信息不足有关。以新药队列为例，估计药物效应需要暴露组与对照组的结局比较；暴露组仅有少量事件时，增加大量对照者仍不能充分弥补这一信息不足。进一步调整协变量，需要在相应条件下估计药物效应，关键暴露与结局组合中的少数观察可能对估计产生较大影响[2-3]。

在模型正确且满足常规正则条件时，MLE 具有一致性，但不保证有限样本无偏。在 Logistic 等非线性模型中，计数的随机波动经非线性估计过

程转换后，正、负方向的偏移不一定相互抵消，因而除增加估计变异外，还可能使估计系统性偏离真值。信息不足时，对数似然的局部二次近似也可能较差，使常规标准误和Wald区间的近似不够准确；这属于推断可靠性问题，与点估计偏倚的产生机制不同[2-3]。

随着协变量数量、分类变量水平数、样条基函数及交互项的增加，待估参数数目随之增加；若现有证据不足以支持这些参数的估计，有限样本偏倚可能加重。此时，即使模型形式正确、混杂因素选择合理，调整后的估计仍可能受到稀疏性的影响[2-3]。此外，OR等比值指标由回归系数经指数变换得到，非线性变换会改变偏倚的表现，因此，对数系数的偏倚减少不等同于比值尺度上的无偏估计[3]。

1.3 与相关概念的区分

小样本描述的是数据的总体规模，稀疏性则取决于具体参数可获得的信息。结局罕见并不意味着一定存在严重偏倚：在各比较组有足够事件且模型复杂度适当时，估计仍可能较稳健。宽置信区间反映估计不精确，不能据此判定点估计的偏倚方向或大小；偏倚与不精确性可同时存在[2-3]。

分离涉及Logistic回归系数是否存在有限MLE。若某一协变量或协变量的线性组合能将所有事件与非事件严格分开，则为完全分离。若部分事件和非事件落在同一分界面上，而分界面外的观察仍可按结局分开，则为准完全分离。两种情况下均有至少一个系数不存在有限MLE[5]。分离在稀疏数据中较常见，但稀疏数据偏倚也可在无分离且模型收敛时存在，二者不能等同。

正值性要求在目标人群相关协变量层内，各比较治疗均有非零的接受概率。若因禁忌证而从不使用某药，该层缺乏相应治疗的观察比较，属于结构性正值性违背，涉及因果效应的识别条件；若理论上可接受两种治疗，但有限样本中一种治疗极少或未出现，则可能属于实际或随机正值性违背，并与稀疏性并存。前者不能仅通过增加样本量或施加惩罚解决，后者则可能随相关治疗观察的增加而缓解[6-7]。

协变量调整后效应值的变化还需与混杂及OR的非可折叠性区分。控制负向混杂可使效应值远离1；即使不存在混杂，调整结局相关因素后的条件OR也可因非可折叠性而不同于边际OR。因此，调整后OR增大本身不能证明存在稀疏数据偏倚，需结合数据结构及模型设定判断[3]。

2 药物流行病学中的常见场景与实践问题

2.1 常见数据结构与高风险研究问题

药物流行病学研究中的数据局部稀疏，常见于三类场景：低频暴露、罕见结局和分层分析。在低频暴露方面，新上市药物、特定剂量或给药途径及联合用药可能仅涉及少量使用者。在罕见结局方面，严重皮肤不良反应、急性肝衰竭、粒细胞缺乏及先天畸形等事件数也可能不足。这类稀疏在具体设计中会被进一步放大。例如比较效果研究中，较少使用的二线药物组与常用活性对照药组的样本量可能严重失衡，妊娠、儿童或罕见病研究按孕期、年龄或疾病亚型限制人群后，可比较的对象还会减少；在自身对照设计中，若某一风险窗内事件过少，即使纳入的病例总数较多，也不足以判断该窗口的估计是否稳定。

大型医保和电子健康记录数据库则可能呈现另一类稀疏性。为控制潜在混杂，研究者常采用高维倾向评分方法，在预先指定的混杂因素基础上纳入更多协变量。然而，罕见暴露与高维二元特征可能导致治疗模型过拟合，产生接近0或1的倾向评分及极端权重[6-8]。此时，稀疏性不仅影响暴露模型的稳定性，还可通过极端权重降低有效样本量，进而影响加权效应估计的准确性。

2.2 当前实践中的识别不足

上述稀疏性可随研究限制逐步累积。例如，新上市二线药使用者本身较少，若进一步采用新使用者设计、限定短风险窗并按年龄分层，某些比较可能只剩少量事件。因此，可行性评估宜随纳入限制和分析分层逐步进行，并权衡这些限制带来的偏倚控制收益与信息损失。

从应用现状看，药物流行病学研究常报告总样本量和最终调整模型，但对于关键暴露-结局计数、分离诊断、参数数量、模型未收敛时的处理方式及惩罚敏感性分析，往往报告不足。Gosho等[9]的模拟与实例研究表明，在事件数少且协变量分布严重不平衡时，MLE、Firth和精确法的OR均可有明显偏倚，具有适当log-F先验的

贝叶斯方法在部分场景中表现更好。因此，应将稀疏性的识别、方法选择和透明报告纳入分析流程。这也与近年对医学统计透明报告的倡议一致[10]。

3 稀疏数据结构及其对估计的影响评估

稀疏数据是信息分布的特征，稀疏数据偏倚则是估计量的重复抽样性质。单次研究通常不能直接测得其偏倚方向和大小，故应依次评估数据支持度、估计与推断异常，以及结论对合理模型设定的依赖程度[3, 9]。

3.1 识别稀疏数据结构与信息不足

建模前应列出暴露与结局的交叉计数，并按主要混杂因素或匹配因素查看分层分布。Cox 回归还需关注事件发生时风险集中的暴露构成，Poisson 回归需检查各暴露水平的事件数与人时；条件 Logistic 回归则应识别真正提供暴露对比信息的不一致匹配集[2-3]。还需检查低频分类水平、极端连续值及可能导致分离的变量组合[3, 5]。

每变量事件数（events per variable，EPV）仅为粗略预警指标，应按事件数与实际估计的预测变量参数数目之比计算，通常不计截距。一个 5 分类变量通常需 4 个系数，样条基函数、交互项及主要暴露项也应计入；Logistic 回归在结局不罕见时，还需关注事件与非事件中较少的一类[3, 11-12]。EPV≥10 并非通用安全界限，预测模型的样本量建议也不能直接用作因果效应估计的合格标准[12]。因此，EPV 需结合暴露分布、关键单元格计数及分离诊断解释。

关键 2×2 表中存在零单元格，或某单元格计数不超过 5、主要暴露组事件极少、参数负担过重，以及倾向评分接近 0 或 1，均可作为进一步检查的触发条件，而非确诊或排除偏倚的统一阈值[3, 7]。加权分析可计算基于权重离散程度的有效样本量（effective sample size，ESS）：$ESS=(\sum w)^2/\sum w^2$，其中 w 为个体权重[13]。应分别报告各治疗组的 ESS 及实际事件数；ESS 并非结局信息量的完整度量，也不能替代事件分布及重叠检查[13]。

3.2 诊断估计与推断过程的异常

拟合时需关注不收敛、迭代次数达上限、系数与标准误异常大，以及效应估计的置信区间跨越多个数量级等现象，并区分分离、共线性与数值计算问题。Wald 区间与轮廓似然区间差异显著，提示局部正态近似可能不充分；调整后比值效应估计随协变量逐步加入而持续远离 1，也需结合混杂控制及非可折叠性解释[3, 5]。这些现象可提示估计或推断不稳定，均不能单独证明稀疏数据偏倚。

可按预先规定的顺序加入协变量，观察主要暴露系数的变化，并通过留一法评估单个事件的影响。若删除一个事件即可改变效应方向或数量级，说明结论对少数观察敏感，但不能据此确定偏倚方向或删除该观察。低频类别合并或病例重新分类仅宜用于有实质依据的敏感性分析，不能为改善结果而修改数据[3]。

3.3 替代模型及敏感性分析

在保持目标人群、效应尺度和必要混杂控制尽可能一致的前提下，可比较 MLE 与 Firth、log-F 惩罚或弱信息先验下的估计，并改变合理的惩罚强度或先验设定[9, 14]。结果差异较大提示结论依赖分析假设，但不能据此确认偏倚的大小，或认定收缩后、更接近 1 的估计更真实。加权分析还应比较权重处理前后的结果、分组 ESS 及事件分布，并识别估计目标是否改变[6]。

一项房性快速性心律失常患者血栓栓塞风险研究中，血管疾病与血栓栓塞事件关联的最大似然 OR 为 7.83［95%CI（1.19，52.53）］，但仅 5 名患者有血管疾病，其中 2 人发生事件。分别采用 log-F（1，1）和 log-F（2，2）先验进行数据增广后，OR 分别为 5.62［95%CI（0.76，34.57）］和 4.25［95%CI（0.65，24.36）］[15]。该例虽非药物效应研究，但说明少数观察支持的关联可对惩罚设定敏感，药物研究中低频合并症的解释也需作类似检查。

Choi 等[16]比较不同回归估计及检验方法发现，在发生完全分离的结局中，Firth 惩罚似然结合相应的似然比检验可改善估计与检验表现。因此，处理分离时还需明确区间和检验的计算方法。丹麦一项 HPV9 疫苗自身对照病例系列研究的源队列包含 854 586 名 10~17 岁青少年，其中 3 354 名接种且发生研究结局者进入分析。白塞病等部分结局在风险期和对照期的事件数均为个位数，白塞病率比为 15.50［95%CI（0.76，316.56）］[17]。

4 稀疏数据偏倚的处理原则与方法选择

处理方法应针对稀疏性的来源选择（表1）。治疗组间缺乏可比重叠时，应评估原估计目标能否由现有数据支持，再考虑补充数据或调整比较策略；若需改变目标人群，应重新界定估计目标。具有可比观察但信息有限时，可通过适当的模型简化与惩罚估计改善有限样本表现[3, 6-7]。

4.1 研究设计阶段的控制策略

研究方案阶段应明确研究问题、估计目标、比较策略和风险窗，并评估关键暴露与结局组合的可用信息。延长研究观察期、扩大地理覆盖范围或联合多个数据源可能增加事件数，但需评估数据质量和人群异质性。嵌套病例对照或病例队列抽样可提高资料收集效率，但不能增加源队列中已发生的结局事件数；匹配和限制也可能减少可用于分析的对象及事件，需权衡混杂控制与估计精度。对短暂暴露与急性结局，可考虑自身对照病例系列（self-controlled case series，SCCS）或病例交叉设计，并分别检查事件依赖暴露、观察终止及暴露时间趋势等假设。SCCS引入活性对照可能有助于处理部分时变混杂，但仍需论证比较策略的适用性[18]。

4.2 估计目标一致性与方法可比性

选择处理方法前，应明确目标人群、治疗比较、结局及风险窗，明确效应指标（如OR、RR、HR），并区分给定协变量的条件效应与对目标人群平均的边际效应。以倾向评分加权为例，总体平均处理效应、已治疗人群平均处理效应与重叠人群平均处理效应对应不同目标人群，不能仅因置信区间较窄就相互替代[19]。常规多变量Logistic回归的暴露系数通常对应条件OR，而标准化或适当加权可用于估计边际效应；即使无混杂，条件OR与边际OR也可能因非可折叠性而不同[3]。

比较降维、匹配、加权、限制人群及惩罚方法时，应核查分析样本、必要调整集与估计目标是否一致。删除必要的混杂因素可能导致残余混杂，限制或匹配可能改变估计结果适用的人群，调整权重也可能改变所估计的效应或引入偏倚。若无法保持一致，应分别说明新的估计目标与适用人群，不将方法间差异一概归因于稀疏数据偏倚的纠正[6, 19]。

4.3 协变量选择、降维与模型复杂度

在因果效应估计中，协变量选择应依据因果结构，而非当前样本的P值或预测能力。应基于领域知识和有向无环图（directed acyclic graph，DAG）预先识别最小充分调整集，估计总效应时避免调整中介和碰撞因子，也不宜仅为提高治疗预测能力而纳入不能帮助控制结局混杂的纯工具变量[20]。为了缓解稀疏性而按P值逐步删除变量，可遗漏重要混杂因素，并使保留系数的P值偏小、置信区间偏窄[21]。

降维与惩罚可减少模型的有效复杂度，但预测变量筛选不能直接替代混杂因素选择。LASSO将系数压缩为0，并不意味着该变量在因果上不重要[21]。在大规模倾向评分分析中，惩罚选择还需结合协变量平衡和目标效应的估计表现，不能仅依据治疗预测性能[8, 22]（表2）。

4.4 精确统计法

Fisher精确检验和精确Logistic回归可在简单的小样本或零单元格情形中提供不依赖常规大样本近似的推断[23]。但“精确”不等于点估计无偏，条件推断可能损失信息，区间还可能因数据离散而变宽。随着协变量及其取值水平增多，计算负担上升，限制了其在大型医疗数据库多参数

表1 药物流行病学常见稀疏数据场景及优先策略

Table 1. Common sparse-data settings and priority strategies in pharmacoepidemiology

稀疏数据场景	主要风险	优先设计或分析策略
罕见暴露（新药、特定剂量、联合用药）	暴露组事件或非事件过少，混杂模型不稳定	扩大数据源；依据暴露分布评估可行性；减少不必要的模型参数
罕见结局（严重不良事件、先天畸形）	事件数不足，完全或准完全分离、极宽区间	多数据源联合；高效抽样收集协变量；惩罚似然或贝叶斯估计
短暂暴露-急性结局	风险窗内事件少，个体间混杂或时变混杂	SCCS或病例交叉；按设计检查事件依赖暴露、时间趋势与观察终止
多药物、多结局或多亚组	低频单元格、交互项和多重性叠加	预先指定主要对比；层级收缩；将信号发现与验证分离

表2 协变量选择与正则化方法的定位

Table 2. Positioning of covariate-selection and regularization methods

方法	优势	局限	建议定位
领域知识与DAG	直接面向混杂控制，假设透明可审查	依赖实质知识，因果结构可有争议	因果效应研究的首选基础
效应值改变法	简单直观，能反映变量对目标估计的影响	依赖样本、变量顺序和阈值；在稀疏数据中不稳定	只作补充敏感性分析，不作主要因果筛选
逐步回归	自动、实现简单	可误删混杂因素；选择后偏倚、*P*值和区间失真	不建议用于因果混杂控制
LASSO/弹性网回归	适用高维特征，改善预测稳定性	可压缩重要混杂因素；对相关特征和惩罚强度敏感	预测或高维倾向评分分析；保留必要混杂因素并检查重叠
贝叶斯模型平均	将模型不确定性纳入推断	计算与先验设定复杂；模型空间可迅速增大	中低维、明确候选模型的场景
贝叶斯正则化	通过先验连续收缩系数，可保留不确定性	结果依赖变量尺度和先验，需敏感性分析	稀疏或高维模型，尤其适合有可辩护背景知识时

模型中的应用。

4.5 惩罚似然与Firth偏倚减少

在Logistic回归中，Firth方法通过在对数似然中加入基于Fisher信息矩阵的Jeffreys先验惩罚项，消除MLE回归系数偏倚展开式中的一阶项；在设计矩阵满秩时，可在完全或准完全分离下得到有限估计[24-25]。当分析以预先指定的Logistic回归系数为目标、出现分离或明显有限样本问题，且未另行设定有实质依据的系数先验时，可优先考虑Firth法。该方法减少的是特定参数化下回归系数的偏倚，不能保证指数变换后的OR无偏；事件极少、暴露分布极端或真实OR较大时，所得OR仍可能存在明显偏倚[9]。Firth法不能消除未控制混杂、纠正模型错误或弥补结构性无重叠。报告时应将Firth点估计与轮廓惩罚似然置信区间配套使用，通常优先于Wald区间；在MLE存在且可比较时，可将两者并列[24]。

系数偏倚减少不等于绝对风险预测校准改善。标准Firth法可使预测概率向0.5偏移；在罕见结局中，平均预测风险可能偏高。若目标包括风险预测，可考虑截距校正（Firth Logistic regression with intercept correction，FLIC）、增加辅助协变量的校正（Firth Logistic regression with added covariate，FLAC）或其他校准方法，并在独立验证数据中或采用适当的内部验证方法，评估预测概率与实际结局发生率的一致性；不能由效应估计改善推断风险预测也得到改善[26]。

4.6 数据增广与log-F惩罚

数据增广（data augmentation）可将参数先验表示为伪观测，并与真实数据合并后利用常规软件进行惩罚似然拟合。Logistic回归中的log-F先验可用这一方式实现[27-28]。当需要按药理学或既往证据明确效应的合理范围、对不同系数设定不同收缩强度，或评价惩罚强度的敏感性时，可考虑log-F惩罚。log-F是一类先验分布，也可用于构建惩罚项。数据增广是实现这一先验惩罚的计算方式，并非另一种独立的偏倚校正原理。惩罚似然实现通常获得回归系数的后验众数；可采用轮廓惩罚似然区间，但不应将其直接称为后验分位数可信区间[27]。

例如，对某一已明确编码的暴露系数采用log-F（2，2）先验时，该OR先验分布的95%中心区间约为1/39~39[9, 27]。该设定容许正、负方向的关联，但对某些药物效应仍可能过宽或过窄，需结合背景知识判断。连续变量应按有临床意义的单位编码，如每10 mmHg血压变化，以便解释先验约束；同时说明哪些系数受到惩罚，以及截距是否单独处理[27]。

先验设定及敏感性分析方案宜在查看主要效应结果前确定。Gosho等[15]提出利用数据估计先验中心的经验性方法，提示固定以0为中心可能在部分场景中过度收缩。使用时需明确数据如何参与先验构造，并评估其对不确定性估计的影响。

4.7 倾向评分、权重与高维方法

倾向评分将协变量信息归纳为接受某治疗的条件概率。在暴露较常见而结局罕见的队列中，使用倾向评分有时可减少结局模型的参数负担，但不会增加实际结局事件数；匹配还可能剔除部分对象及其事件，进一步减少信息[3, 22]。应按治

疗组报告匹配或限制前后的人数、事件数；加权分析同时报告各组ESS、权重分布和协变量平衡，不能将权重之和或加权事件数当作新增的独立观察[13, 29]。暴露近乎被协变量完全预测时，匹配或加权仍可能不稳定。应根据重叠程度考虑限制人群、匹配或调整权重，并说明估计目标是否改变[6, 19]。权重稳定化后仍可能存在极端权重，截断也可能引入偏倚[29]。高维惩罚及交叉拟合有助于控制过拟合，不能补足缺乏重叠的信息[8]。

4.8 贝叶斯方法与层级收缩

当需整合外部证据、层级借用信息或报告后验概率时，应进行完整贝叶斯后验分析，而非仅计算惩罚似然点估计。对Logistic模型，为所有系数（包括截距）设置适当的正态先验，通常可保证后验分布可归一化；但发生分离且采用Cauchy先验时，某些回归系数的后验均值可能不存在，需确认拟报告的后验汇总统计量是否有定义[30]。弱信息先验需与变量编码和单位匹配，不能对不同原始尺度的变量机械套用同一分布。

常用系数先验包括正态、Student-t（包括Cauchy）和log-F分布；因此，log-F先验也可用于完整贝叶斯分析[27, 31]。后验明显不对称时，可报告后验中位数及2.5%和97.5%分位数组成的95%等尾可信区间；如采用最高后验密度区间，应注明不能与频率学置信区间混称。在极端稀疏条件下，先验族、中心和宽度均可明显影响结果，没有适用于所有效应情形的统一最优先验[9]。采用抽样计算时，应检查链的收敛、有效抽样量及尾部行为；此处的有效抽样量不同于倾向评分加权的ESS。

同时评估多个药物、结局或亚组时，层级模型可通过部分汇聚，使信息较少的效应估计向组水平分布收缩，从而降低估计波动。其适用性取决于药理学、结局机制或人群特征是否支持可交换性；不宜为增加精度而合并机制明显不同的效应[27]。

4.9 方法选择

表3归纳了不同数据支持度及诊断结果下的优先处理路径。对于同一Logistic效应参数，发生分离且主要需要偏倚减少估计时，可考虑Firth法；需要明确或改变系数收缩范围时，可考虑log-F惩罚；需要层级建模或后验概率时，可采用完整贝叶斯分析。这些选择并非依据事件数设定的固定规则，也不能替代对估计目标、混杂控制及模型假设的检验[9, 14, 24, 27]。双重稳健估计不能自动消除稀疏性，其有限样本性能仍受倾向评分模型、结局模型的估计质量及治疗重叠程度影响。若采用不同的合理方法后结果仍有明显差异，应核查其估计目标和假设是否可比，并报告结论的敏感性，不宜按*P*值选择最终模型[6-8]。

5 报告规范建议

Mansournia等[32]建议统计报告中应评估稀疏数据偏倚，并结合效应量及区间解释结果。结合RECORD-PE及非干预性研究报告要求[10, 33]，研究宜重点报告以下内容。

5.1 研究样本与数据描述

按分组报告人数、事件数、随访人时，以及关键分层的暴露与结局交叉计数。说明纳入限制、匹配和缺失值处理造成的人数及事件数变化[10]。自身对照研究报告各风险窗与对照期的事件分布[10]。

表3 数据支持度及估计诊断结果与优先处理路径

Table 3. Data support and estimation diagnostics with priority responses

数据支持度或诊断结果	优先处理	不建议的做法
结构性无重叠	限制目标人群、改变比较者或估计目标；必要时停止估计	依靠Firth或强先验跨越不存在的数据支持
存在临床可比性但发生分离	在目标参数明确后考虑Firth；需设定收缩范围时用log-F惩罚；需后验概率时作完整贝叶斯分析	把软件收敛或有限点估计等同于可靠
罕见结局且混杂因素较多	保留必要混杂因素；可用倾向评分或适当降维；报告处理后的各组事件数	按*P*值删除混杂因素；认为匹配或加权增加了实际事件
极端倾向评分或权重	检查重叠、各组ESS及事件数；说明权重处理是否改变估计目标	只报告平衡；将加权事件数视为独立事件数
多药物、多结局或多亚组	预先排序并采用层级收缩；区分信号生成与验证	逐项筛选显著结果并忽略多重性和选择效应
以绝对风险预测为目标	评估校准；必要时采用FLIC/FLAC或再校准，并进行验证	仅报告Firth回归系数或比值效应

5.2 模型设定与稀疏性诊断

列明调整变量及其因果依据[20]、变量编码和实际参数数目；报告分离、收敛及主要模型假设的检查结果[5]。采用倾向评分时，展示治疗组间重叠及协变量平衡；加权分析还应报告权重分布[29]及各组ESS。限制或匹配后应列出各组实际事件数；加权分析仍应报告未加权的事件数。零单元格、无重叠区域和无法估计的效应均应如实报告；不能仅以EPV达到某一阈值或软件收敛说明信息充足[3, 12]，也不能将诊断异常直接等同于已证实的偏倚。

5.3 估计方法与分析实现

说明主要估计方法及选择理由。惩罚似然报告惩罚形式、作用参数与强度；数据增广报告伪观测构造与权重，伪观测不计入实际样本量或事件数[28]；贝叶斯报告先验分布、参数、尺度及依据，区分预设与数据参与构造的先验[27]，并注明后验汇总量与可信区间类型；Firth法注明是否采用轮廓惩罚似然置信区间[24]。报告软件及版本、关键选项，贝叶斯抽样另附收敛诊断。MLE不存在或算法未收敛时应明确标注。

5.4 效应量与不确定性的呈现

报告主要效应量及95%置信区间或可信区间；并列比较常规估计与偏倚减少估计时，标明样本、调整集和估计目标的差异；目标人群或条件效应或边际效应改变时，不将估计差异直接解释为偏倚减少。数据支持时可补充绝对风险或风险差，结局常见时不宜将OR解释为RR[32-33]。

*P*值应给出具体数值，不能将$P \geq 0.05$或区间跨越无效应值解释为无关联或无风险。例如，假设RR为1.8、95%CI为0.6~5.4，应说明估计不精确，仍不能排除有临床意义的风险增加。区间解释还应考虑模型假设及未计入的系统误差[32]。

5.5 敏感性分析与亚组分析结果

围绕主要稀疏性来源，改变惩罚或先验设定，检验个别观察及权重处理方式对估计的影响，并报告相应的效应量与区间，说明结论对模型和先验的依赖程度，不将估计收缩等同于偏倚已被消除，也不以显著性是否改变作为稳健性的唯一依据[3, 14]。亚组分析应区分预设与探索性分析，列出各组事件数及效应区间，交代多重比较处理；组间差异应通过交互作用项直接检验；数据允许时，可补充加法交互作用分析结果，不能根据一组显著而另一组不显著作出判断[32]。

综上所述，即使基于大规模人群数据的药物流行病学研究，在罕见暴露或罕见结局、特殊人群及细分亚组等情形下，支持特定效应估计的信息仍可能不足。本文提出从数据结构、估计异常与敏感性分析三个层次识别和评估稀疏数据偏倚：先判断稀疏的来源与性质，继而识别数据分离、系数估计不稳定等异常信号，并通过敏感性分析评估偏倚的方向与程度。处理时应先明确估计目标、评估治疗组间的可比性，优先通过优化研究设计与合理选择协变量缓解数据稀疏；在此基础上，再针对数据分离、系数估计不稳定等具体情形，分别选用Firth罚似然回归、log-F惩罚估计或全贝叶斯分析。需要强调的是，统计调整无法从根本上弥补特定亚组或数据稀疏区域的局部信息不足。因此，研究结论的强度应与数据支持程度相称，并透明地报告效应估计对模型假设的依赖。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 Acton EK, Willis AW, Hennessy S. Core concepts in pharmacoepidemiology: key biases arising in pharmacoepidemiologic studies[J]. Pharmacoepidemiol Drug Saf, 2023, 32(1): 9-18. DOI: 10.1002/pds.5547.

2 Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis[J]. Am J Epidemiol, 2000, 151(5): 531-539. DOI: 10.1093/oxfordjournals.aje.a010240.

3 Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight[J]. BMJ, 2016, 352: i1981. DOI: 10.1136/bmj.i1981.

4 中国药学会. 药物流行病学研究方法学指南(第2版): T/CPHARMA 002-2025[S]. 北京: 中国药学会, 2025.

5 Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic regression: causes, consequences, and control[J]. Am J Epidemiol, 2018, 187(4): 864-870. DOI: 10.1093/aje/kwx299.

6 Zhu Y, Hubbard RA, Chubak J, et al. Core concepts in pharmacoepidemiology: violations of the positivity assumption in the causal analysis of observational data: consequences and statistical approaches[J]. Pharmacoepidemiol Drug Saf, 2021, 30(11): 1471-1485. DOI: 10.1002/pds.5338.

7 Petersen ML, Porter KE, Gruber S, et al. Diagnosing and responding to violations in the positivity assumption[J]. Stat Methods Med Res, 2012, 21(1): 31-54. DOI: 10.1177/0962280210386207.

8 Wyss R, van der Laan M, Gruber S, et al. Targeted learning with an undersmoothed LASSO propensity score model for large-scale covariate adjustment in health-care database studies[J]. Am J Epidemiol, 2024, 193(11): 1632-1640. DOI: 10.1093/aje/kwae023.

9 Gosho M, Ohigashi T, Nagashima K, et al. Bias in odds ratios from logistic regression methods with sparse data sets[J]. J Epidemiol, 2023, 33(6): 265-275. DOI: 10.2188/jea.JE20210089.

10 Langan SM, Schmidt SAJ, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE) [J]. BMJ, 2018, 363: k3532. DOI: 10.1136/bmj.k3532.

11 Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis[J]. J Clin Epidemiol, 1996, 49(12): 1373-1379. DOI: 10.1016/S0895-4356(96)00236-3.

12 van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction models: beyond events per variable criteria[J]. Stat Methods Med Res, 2019, 28(8): 2455-2474. DOI: 10.1177/0962280218784726.

13 McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models[J]. Stat Med, 2013, 32(19): 3388-3414. DOI: 10.1002/sim.5753.

14 Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems—using weakly informative Bayesian priors[J]. Epidemiology, 2013, 24(2): 233-239. DOI: 10.1097/EDE.0b013e318280db1d.

15 Gosho M, Ishii R, Nagashima K, et al. Determining the prior mean in Bayesian logistic regression with sparse data: a nonarbitrary approach[J]. J R Stat Soc Ser C Appl Stat, 2025, 74(1): 126-141. DOI: 10.1093/jrsssc/qlae048.

16 Choi L, Carroll RJ, Beck C, et al. Evaluating statistical approaches to leverage large clinical datasets for uncovering therapeutic and adverse medication effects[J]. Bioinformatics, 2018, 34(17): 2988-2996. DOI: 10.1093/bioinformatics/bty306.

17 Faksová K, Laksafoss AD, Hviid A. Human papillomavirus nonavalent (HPV9) vaccination and risk of immune mediated diseases, myocarditis, pericarditis, and thromboembolic outcomes in Denmark: self-controlled case series study[J]. BMJ Med, 2024, 3(1): e000854. DOI: 10.1136/bmjmed-2024-000854.

18 Schultze A, Brown J, Logie J, et al. Overcoming time-varying confounding in self-controlled case series with active comparators: application and recommendations[J]. Am J Epidemiol, 2025, 194(1): 220-225. DOI: 10.1093/aje/kwae216.

19 Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting[J]. J Am Stat Assoc, 2018, 113(521): 390-400. DOI: 10.1080/01621459.2016.1260466.

20 Tennant PWG, Murray EJ, Arnold KF, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations[J]. Int J Epidemiol, 2021, 50(2): 620-632. DOI: 10.1093/ije/dyaa213.

21 Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement[J]. Eur J Epidemiol, 2019, 34(8): 725-730. DOI: 10.1007/s10654-019-00529-y.

22 Rassen JA, Glynn RJ, Brookhart MA, et al. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples[J]. Am J Epidemiol, 2011, 173(12): 1404-1413. DOI: 10.1093/aje/kwr001.

23 Mehta CR, Patel NR. Exact logistic regression: theory and examples[J]. Stat Med, 1995, 14(19): 2143-2160. DOI: 10.1002/sim.4780141908.

24 Heinze G, Schemper M. A solution to the problem of separation in logistic regression[J]. Stat Med, 2002, 21(16): 2409-2419. DOI: 10.1002/sim.1047.

25 Firth D. Bias reduction of maximum likelihood estimates[J]. Biometrika, 1993, 80(1): 27-38. DOI: 10.1093/biomet/80.1.27.

26 Puhr R, Heinze G, Nold M, et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? [J]. Stat Med, 2017, 36(14): 2302-2317. DOI: 10.1002/sim.7273.

27 Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions[J]. Stat Med, 2015, 34(23): 3133-3143. DOI: 10.1002/sim.6537.

28 Discacciati A, Orsini N, Greenland S. Approximate Bayesian logistic regression via penalized likelihood by data augmentation[J]. Stata J, 2015, 15(3): 712-736. DOI: 10.1177/1536867X1501500306.

29 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies[J]. Stat Med, 2015, 34(28): 3661-3679. DOI: 10.1002/sim.6607.

30 Ghosh J, Li Y, Mitra R. On the use of Cauchy prior distributions for Bayesian logistic regression[J]. Bayesian Anal, 2018, 13(2): 359-383. DOI: 10.1214/17-BA1051.

31 Gelman A, Jakulin A, Pittau MG, et al. A weakly informative default prior distribution for logistic and other regression models[J]. Ann Appl Stat, 2008, 2(4): 1360-1383. DOI: 10.1214/08-AOAS191.

32 Mansournia MA, Nazemipour M. Recommendations for accurate reporting in medical research statistics[J]. Lancet, 2024, 403(10427): 611-612. DOI: 10.1016/S0140-6736(24)00139-9.

33 U.S. Food and Drug Administration. M14 general principles on planning, designing, analyzing, and reporting of non-interventional studies that utilize real-world data for safety assessment of medicines: guidance for industry[EB/OL]. (2026-03)[2026-09-15]. https://www.fda.gov/media/179795/download.

收稿日期：2026 年 09 月 17 日 修回日期：2026 年 10 月 03 日
本文编辑：李绪辉 桂裕亮