· 药物流行病学理论与方法 ·

电子健康记录分析中分布式算法的应用现状及分析

陈枭　郭晓晶　许金芳　韦连慧　陈晨鑫　梁际洲　郑轶　迟立杰　贺佳　叶小飞
（海军军医大学军队卫生统计学教研室　上海 200433）

摘　要　随着电子健康记录在医疗保健研究中的广泛应用，为了整合多个临床中心的数据进行分析和预测，以达到实现多中心临床试验的研究目的，分布式算法应运而生。本文通过查阅相关文献的研究方法，对数篇文献中提及的分布式算法进行了归纳、总结、整合和比较，初步得出了一些结论。在不考虑迭代通信的情况下首选不产生信息损失的 GLORE 算法，偏倚为 0，标准差为 1；若在其他临床中心的数据仅可被借用一次，则首选 ODAL 算法；而与原始 ODAL 算法相比，如果存在潜在的外部临床中心，则首选 Robust-ODAL。因此应根据各临床中心给出数据的不同情况，灵活选择不同的分布式算法，以发挥不同算法的优势。

关键词　电子健康记录；分布式算法；网格二进制逻辑回归；ODAL
中图分类号：R181.3^{+}5　**文献标识码**：A　**文章编号**：1005-0698(2021)11-0767-05
DOI：10.19960/j.cnki.issn1005-0698.2021.11.011

Application Status and Analysis of Distributed Algorithms in Electronic Health Record Analysis

Chen Xiao, Guo Xiaojing, Xu Jinfang, Wei Lianhui, Chen Chenxin, Liang Jizhou, Zheng Yi, Chi Lijie, He Jia, Ye Xiaofei
Department of Health Statistics, Naval Medical University, Shanghai 200433, China

ABSTRACT　With the widespread application of electronic health records in medical care research, distributed algorithms have emerged in order to integrate data from multiple sites for analysis and prediction to achieve the research purpose of multi-center clinical trials. This paper summarizes, integrates and compares the distributed algorithms mentioned in several documents by consulting the research methods of related literatures, and draws some preliminary conclusions. Without considering iterative communication, the GLORE algorithm that does not produce information loss is preferred, with a deviation of 0 and a standard deviation of 1; if the data at other sites can only be borrowed once, the ODAL algorithm is preferred; compared with the original ODAL algorithm if there are potential peripheral clinical sites, Robust-ODAL is the first choice. Therefore, different distributed algorithms should be flexibly selected according to different data conditions to give full play to the advantages of different algorithms.

KEY WORDS　Electronic health record; Distributed algorithm; Grid binary Logistic regression; ODAL

电子健康记录（EHR）包含有关各种健康结果和危险因素的广泛信息，因此已广泛用于医疗保健研究。整合来自多个临床中心的 EHR 数据可以通过在更广泛的人群中提供更大的样本量来加速发现和预测风险，有可能减少临床偏倚并提高估计和预测准确性。为了克服患者级别数据共享的障碍，分布式算法得以开发，以通过仅共享汇总数据并在多个临床中心进行统计分析。2015 年国家药品不良反应监测中心发起建设中国医院药物警戒系统，并组建国家药品不良反应监测哨点联盟。“十三五”国家药品安全规划中明确提出“利用医疗机构电子数据，建立药品医疗器械安全性主动监测与评价系统。在综合医院设立 300 个药品不良反应和医疗器械不良事件监测哨点”。如何对于不同的监测哨点数据进行分析是一个较大的挑战。分布式算法在其数据分析中发挥一定的作用。本综述对现有的几种分布式算法进行了比较，旨在对比各种方法的基本原理、使用条件、优势劣势，为国内研究应用分布式

基金项目：国家自然科学基金项目（编号：82073671）；上海市卫计委优秀青年医学人才培养计划项目（编号：2018YQ47）；上海市公共卫生学科带头人项目（编号：GWV-10.2-XD22）；上海市公共卫生优青计划项目（编号：GWV-10.2-YQ33）；上海市公共卫生体系建设三年行动计划学科建设项目“大数据与人工智能应用”（编号：GWV-10.1-XK5）；军队双重建设项目-03
通信作者：叶小飞　Tel：(021)81871442　E-mail：yexiaofei@smmu.edu.cn

算法提供一些参考。

1　分布式算法的开发背景

EHR 包含作为临床护理记录在内的一部分定期收集的信息,包括诊断、治疗药物、治疗程序、影像学和临床记录等数据。自 2009 年以来,电子病历的使用在全国范围内迅猛增长,也可以实现使用 EHR 数据进行有意义的研究[1]。这些医疗保健数据集存储在数据库中,而数据库通常使用多种数据模型以及本地语法构建。跨多个不同数据库的分析必须调整分析方案以适应每个基础数据模型和语法,或将数据库数据模型转换为通用数据模型(CDM)[2,3]。在许多情况下,跨中心共享患者级别数据或将数据提供给各中心是不可行的,特别是如果临床中心位于不同国家/地区则更是如此。为解决这些问题,观察健康数据科学与信息学(OHDSI)联盟成立,其主要目的是开发可在多个临床中心之间共享的开源工具。OHDSI 还开发了通用的 CDM,以使每个临床中心都可以将其本地数据映射到通用的可共享框架。这样一来,单个脚本即可在多个临床中心运行而且无需更改。同时将数据库转换错误的可能性降到最低(将脚本从一个数据库结构转换到另一数据库结构以提取相同类型的结果时),也加快了转换结果的时间。根据不同的场景,OHDSI 开发了数个分布式算法,包括 GLORE, ODAL 等。这些算法将每个临床中心内的计算任务分解为多个部分,而不需要共享患者级别的信息[4]。

2　分布式算法介绍

2.1　GLORE

在以患者为中心的可扩展性全国有效性研究网络(pSCANNER)项目的激励下,开发的一种用于进行 Logistic 回归的分布式算法,称为网格二进制 Logistic 回归(GLORE)[2,5,6]。GLORE 算法没有将数据带到中央存储库中进行计算,而是将计算带到了数据中。GLORE 模型集成了可分解的部分元素或非隐私敏感的预测值,以获得模型系数、方差协方差矩阵、拟合优度检验统计量以及受试者工作特征(ROC)曲线下面积(*AUC*)[2]。如果在任何临床中心都无法访问单个患者级别的数据,则首选 GLORE。如果网络中不考虑迭代通信,则首选无损方法(例如 GLORE)[2,6]。此算法可保证信息无损, 其中偏倚为 0,标准误等于 1。而 GLORE 也存在相应的不足:数据需要跨中心迭代,直到达到收敛为止,而在涉及更多协变量的情况下,可能需要进行大量迭代才能实现收敛,这在跨临床中心的交流中造成了巨大负担,其方法本身忽略跨临床中心的数据异质性,这在研究结果与暴露之间的关联时会导致很大偏倚。该算法的原理流程见图 1,保存在不同机构(即 A、B、C)的数据集,通过同一个虚拟引擎(即 GLORE 代码)在本地进行处理,计算非敏感的中间结果,通过交换合并在中心得到最终的全局模型参数。

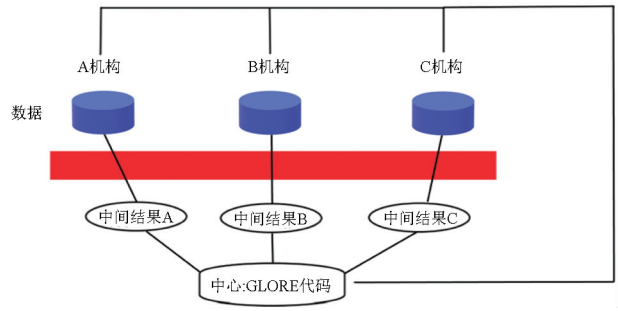

图 1　GLORE 模型流程图

案例辨析:爱丁堡学者开展了一项心肌梗死研究,数据主要结局为是否发生心肌梗死,共收集 48 个变量和 1 253 条记录,记录分布于两个临床中心,分别有 627 条和 626 条[2]。用 GLORE 算法计算结果与传统的中心化 Logistic 回归模型计算出的统计量没有差异,而传统方法需要进行 7 次牛顿-拉弗森(Newton-Raphson)迭代才能以 10^{-6} 的精度收敛。而使用 GLORE 算法计算 *AUC* 值为 0.965,与传统模型得出的值同样没有差异[2,6]。

2.2　ODAL

跨中心迭代这种局限性促使研究人员开发了非迭代的分布式算法,根据 Jordan 等[7]提出一种创新的一站式分布式计算框架,其主要思想是通过使用本地临床中心的患者级别数据和其他中心的汇总信息来构建替代似然函数,在具有稀疏性的高维回归的分布式分析中也提出了这个想法。在 Logistic 回归中行使替代可能性的想法,并开发了一种单次分布式算法来执行 Logistic 回归,称为 ODAL[4,7]。

ODAL 算法的主要优势在于只需要一次从多个临床中心综合汇总统计数据,无需在中心之间进行迭代通信或传输患者级别的数据,与需要在中心之间进行迭代通信的算法相比,将其部署在研究联盟中更为实用。因此具有较高的通信效率,信息传递成本低且利于保护隐私,更有效地利用了来自本地临床中心(可访问患者级别数据的中心)的信息,最终得到的估计值与合并的估计值高度一致。而 ODAL 也存在一些不足:该方法需要访问一个临床

中心各个患者级别的数据，与 GLORE 相类似，该算法忽略跨中心的数据异质性，这在研究结果与暴露之间的关联时会导致很大的偏倚。该算法的原理流程见图 2，使用来自本地中心（即中心 1）的数据，计算本地估计量 $\bar{\beta}$ 并将其传输到其他中心。将中间量 $\nabla L_j(\bar{\beta})$ 在每个中心 j（j = 2，…，K）进行评估，并将其传回本地中心。结合 $\nabla L_1(\bar{\beta})$ 和 $L_1(\beta)$，研究者在本地中心构造似然函数 $L(\beta)$ 并通过最大化该函数获得 ODAL 估计量 β。

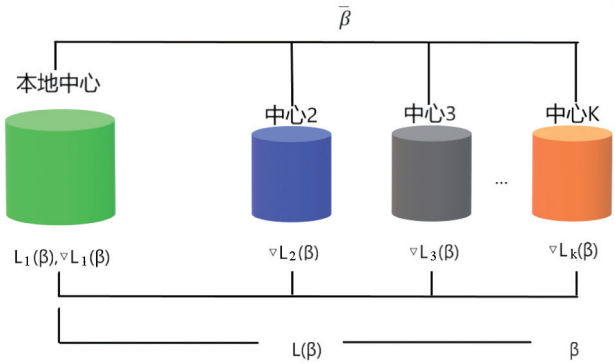

图 2　ODAL 算法的示意图

ODAL 有效地利用了来自本地中心（可访问患者级别数据的中心）的信息，并结合来自其他中心的似然函数一阶（ODAL1）和二阶（ODAL2）梯度，以构建一个估计量，而无需在中心之间进行迭代通信或传输患者级别的数据[2]。ODAL1 优势在于：仅需要进行一轮通信，因此通信成本低于 GLORE，可达到与合并估计量相当的估计精度相比 ODAL2 传输更少的数字。相应地，ODAL1 也存在以下缺点：需要访问一个临床中心的各个患者级别的数据。实际上，当拟合一个相对低维的模型时，转移 p×p 数的额外通信成本可以忽略不计。在这种情况下，ODAL2 将是首选[2]，可达到与合并估计量相当的估计精度，而且比 ODAL1 具有更好的精度，且几乎与金标准估计量相同。ODAL2 算法的缺陷则在于与 ODAL1 相比需要传输少量的额外汇总信息。

案例辨析：使用宾夕法尼亚大学卫生系统（UPHS）胎儿流产数据集，其共包含 30 810 例正常妊娠和 4 763 例流产病例（流产率为 13.43%），影响流产的可能因素包括年龄、体重、体重指数（BMI）、种族 4 个变量[2,8]。

为了评估所提出算法的性能，研究者设计了以下两种模拟数据集。

A：从 K 个中心中随机生成数据。本地中心有 1 000 个样本，其他每个 K−1 中心有 $10^r \times 1\,000$ 个样本，r 从 −1～1 随机选择。对范围 2～100 的 K 的不同值进行单独的模拟。

B：随机生成 10 000 例患者数据，并将数据划分为 10 个子集，在其中将 n 个样本分配给本地中心，其他 9 个中心随机划分 $1\,000-n$ 个样本，对 n 的不同值（范围为 100～9100）执行单独的模拟。当本地中心的相对大小（与患者总数相比）从小百分比增加到很大比例时，此数据集将测试 ODAL 的性能。

本研究验证了 GLORE 是无损耗的，其中偏倚为 0，相对偏倚为所有情况下的标准误均为 1。在数据集 A 中，当 K 增加时，总样本量增加，而本地样本量保持不变。因此，局部估计量的相对偏倚和标准误会由于跨中心的总样本量增加而增加。与合并数据相比，ODAL1 的相对偏倚较小（<0.5%），相对标准误差在 1.02～1.25 之间。ODAL2 更准确，相对偏倚<0.1%，标准误<1.05。在数据集 B 中，当总样本量固定时，局部估计量的性能会随着局部样本量的增加而提高。另一方面，对于不同大小的所有本地样本数据集，ODAL1 和 ODAL2 的相对偏倚均小于 0.3%。

为了降低通信成本，GLORE 需要进行 5～7 轮通信，直到算法收敛为止；对于每次迭代，其需要将 p×p+p 个数字从每个中心传输到中心。ODAL1 和 ODAL2 仅需要进行一轮通信，其中 ODAL1 需要将 p 个数字从每个中心传输到本地中心，而 ODAL2 需要将 p×p+p 个数字从每个中心传输到本地中心。总而言之，ODAL1 和 ODAL2 均可达到与合并数据相当的估计精度，而 ODAL2 比 ODAL1 具有更强大的准确性。ODAL1 和 ODAL2 的通信成本均低于 GLORE。

2.3　Robust-ODAL

为了克服共享患者级别数据的障碍，出现了分布式算法，以通过仅共享汇总信息来跨多个临床中心进行统计分析。但是，现有中心的分布式算法通常会忽略中心之间数据的异质性，这在研究结果与暴露之间的关联时会导致很大的偏倚。

在 Robust-ODAL 研究中，提出了一种隐私保护和通信有效的分布式算法，该算法解决了由少数临床中心引起的异质性，以在异构健康系统内拟合 Logistic 回归而不需要共享患者级别的数据[2,9]。关键思想是通过传达对“外部研究”存在不那么敏感的可靠汇总统计信息来修改 ODAL 算法。通过使用 Janssen Research 的数据库进行的仿真研究和真实数据分析，发现新算法（Robust-ODAL 方法）对外部

研究的稳定性较强，并且产生的偏倚估计要比当前的 ODAL 方法和传统的 Meta 分析方法少。若存在潜在的外部临床中心，则首选 Robust-ODAL[2,9]。该算法的原理流程见图 3，使用来自中心 1（即本地中心）的数据来估计本地估计量 $\bar{\beta}$，并传输到其他中心，在其他中心计算中间量 $\nabla L_j(\bar{\beta})$，并传回中心 1，得到 $\nabla L_1(\bar{\beta})$。通过 $\nabla L_1(\bar{\beta})$ 和 $L_1(\beta)$，得到 Robust-ODAL 统计量 β。

Robust-ODAL 有很多优势：该算法解决了由少数临床中心引起的异质性，具有最宽的置信区间，可以正确反映异质性的潜在影响，具有更高的精度，优于传统的 Meta 分析，在罕见疾病的发生环境中偏倚较小。不需要在中心之间进行迭代通信，仅需要在单个临床中心中访问各个患者级别的数据，从而减少了通信成本和管理工作量。Robust-ODAL 仅需从其他中心传输汇总信息以构建替代似然函数，从而避免共享患者级别的信息，相对于 ODAL，Robust-ODAL 产生的回归系数估计偏倚要小得多[2]。但是 Robust-ODAL 算法也不是完美无缺的，当中心总数较少时，Robust-ODAL 可能无法很好地执行，外部临床中心在所有中心中的比例也会对 Robust-ODAL 产生影响，因为需要访问其中一个临床中心的各个患者级别的数据[2]。

案例辨析：强生詹森研究所关于急性心肌梗死危险因素的研究，共有 5 个数据库，纳入 Logistic 回归模型的风险因素包括：肥胖、酒精依赖、高血压疾病、重度抑郁症、2 型糖尿病和高脂血症。分别用 ODAL、Robust-ODAL、Meta 分析方法与合并数据方法计算比值比（*OR*）以及 95% 置信区间。ODAL 与 Robust-ODAL 算法的估计之间存在较大差异。对于大多数风险因素，ODAL 提供了更接近合并数据分析的 *OR* 值的点估计。中心研究者认为在所有中心之间拟合联合 Logistic 回归模型可能会导致偏倚，因为其忽略了中心之间的差异。合并数据分析的估计值可能会有偏倚。Robust-ODAL 算法旨在解决这种异质性问题，其被证明具有最宽的置信区间，因此可以正确反映异质性的潜在影响[2,10]。

3　总结与展望

3 种分布式算法的比较见表 1。如果在任何临床中心都无法访问单个患者级别的数据，则首选 GLORE，因为该方法是无损的[2]。

表 1　3 种分布式算法的比较

算法	提出时间	数据要求	优点	缺点
GLORE	2012 年	各中心中间数据	无损，无偏倚	通信成本高，忽略中心之间异质性
ODAL	2019 年	无需患者一级的数据	通信效率高，成本低，结果接近合并数据分析	需要访问一个中心的原始数据
Robust-ODAL	2020 年	无需患者一级的数据	考虑中心之间数据的异质性，置信区间最宽	中心总数少时难以执行

OHDSI 联盟由许多合作机构组成，在这些合作机构中，不允许患者级别的数据共享，因为这经常与地区立法相冲突。在这种情况下，单个研究人员可能会在其给定中心上获得患者级别的数据，但随后将在其他临床中心上部署他们的算法而无需访问患者级别的数据。此时，ODAL 是理想的选择，因为从其他临床中心收集的汇总信息只能借用一次[2]。

如果存在潜在的外部临床中心，则首选 Robust-ODAL。当数据被认为相对均匀时，ODAL 方法是首

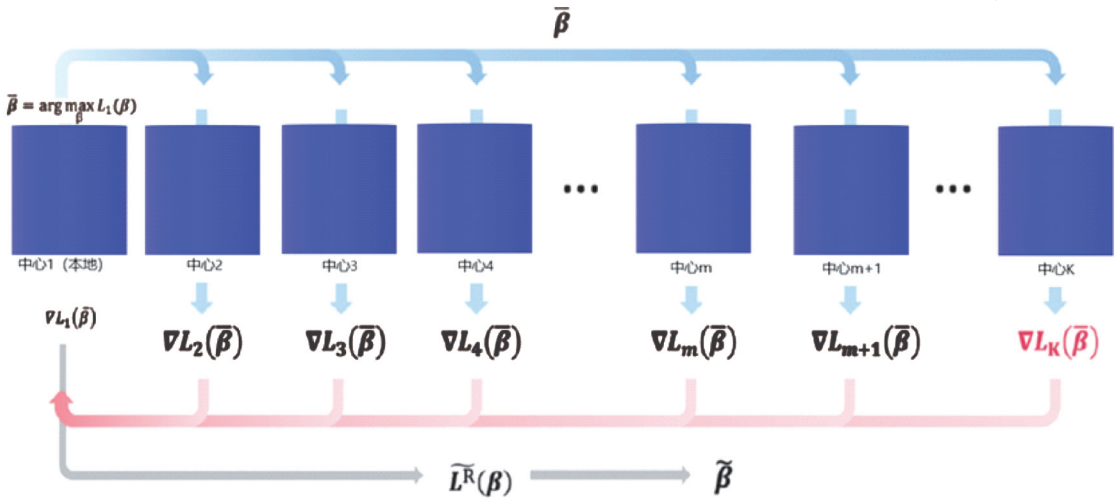

图 3　Robust-ODAL 流程图

选。基于实际数据应用,表明临床中心总数可能会影响所提方法的性能。当中心总数很小时,Robust-ODAL 可能效果不佳,因为当中心数量较少时,中位数会更敏感(变化较大)[2]。外部临床中心在所有中心中的比例也会对所提出的方法产生影响,当外部临床中心比例较大时,应考虑其他方法。

ODAL 可以应用在需要进行分布式分析的许多其他环境中。如对于利用行政数据(如出生和死亡记录)的人口和全球健康研究,人们越来越关注地方或国家卫生部门发布的数据[2]。此外,来自前瞻性队列的数据,尤其是环境健康研究,通常由参加者居住地进行收集,暴露时间和结果的时间以及其他信息采用特殊标识符而无法在上级研究之外共享[12~14]。将来,研究者可以着手计划将 ODAL 方法扩展到其他类型的研究维度,如作为分类和事件发生时间数据。另外,研究者正逐步将研究扩展到高维数据集,其中与总样本量大小相比,协变量的数量被认为是很大的,仍有许多困难需要克服。研究者们也正致力于开发用于在分布式网络中直接应用 ODAL 的开源软件包,相信这些算法可以作为对现有分布式算法的有效补充与拓展。

基于我国 EHR 的研究已逐步展开,在此类研究中可尝试建立良好的分布式数据网,以改善研究的可重复性、透明性,以及减小偏倚,尤其是跨临床中心研究带来的偏倚。需要注意的方面包括:建立通用的协议或 CDM 来规范分析过程;使用盲法来减少偏倚,在最终结果得出前各中心的数据结果保持对其他临床中心的研究者不可见,避免研究者被其他中心的结果导向而产生偏倚[11]。

目前,国内多中心研究面临的主要困难在于数据共享的障碍,医院独立管理医疗数据,第三方研究机构与药品制造方难以获取患者诊疗数据,给研究带来比较大的困难,此问题有待国家后续出台相关政策法规来规范医疗数据共享,以及研究者、医院之间的沟通和协调。由于国内未见基于 EHR 的跨国多中心研究,故目前的研究推荐采用 GLORE 方法和 ODAL 方法来进行数据分析。当多中心研究进一步发展后,Robust-ODAL 和其他新的算法将成为强大的工具。

参 考 文 献

1 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records[J]. N Engl J Med, 2010, 363(6):501-504

2 Kennedy RL, Fraser HS, Mcstay LN, et al. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: Derivation and evaluation of logistic regression models[J]. Eur Heart J, 1996, 17(8): 1181-1191

3 Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research [J]. J Am Med Inform Assoc, 2012, 19(1):54-60

4 Duan R, Boland MR, Moore JH, et al. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites[J]. Pac Symp Biocomput, 2019, 24: 30-41

5 Ohno-Machado L, Agha Z, Bell DS, et al. pSCANNER: Patient-centered Scalable National Network for Effectiveness Research[J]. J Am Med Inform Assoc, 2014, 21(4): 621-626

6 Wu Y, Jiang X, Kim J, et al. Grid Binary LOgistic REgression (GLORE): Building shared models without sharing data[J]. J Am Med Inform Assoc, 2012, 19(5): 758-764

7 Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference[J]. J Am Stat Assoc, 2019, 114 (526): 668-681

8 Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm[J]. J Am Med Inform Assoc, 2020, 27(3): 376-385

9 Tong J, Duan R, Li R, et al. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data[J]. Pac Symp Biocomput, 2020, 25: 695-706

10 Lanas F, Avezum A, Bautista LE, et al. Risk factors for acute myocardial infarction in Latin America: The INTERHEART Latin American study[J]. Circulation, 2007, 115 (9): 1067-1074

11 Platt RW, Platt R, Brown JS, et al. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias[J/OL]. Pharmacoepidemiol Drug Saf, 2019 doi: 10.1002/pds.4722 [Epub ahead]

(2021-06-25 收稿　2021-10-02 修回)