

# 如何撰写高质量的流行病学研究论文

## 第四讲 免疫基因组学研究报告规范——STREIS 介绍

陈茹 徐传辉 詹思延

**【导读】** 现代高通量的人类白细胞抗原和杀伤细胞免疫球蛋白样受体分型技术产生了大量的免疫基因组学数据,对组织相容性和免疫相关的疾病关联研究带来深刻影响。而免疫基因组学数据的高度多态性为不同数据集的信息合成带来了独特的方法学和计算挑战。“加强免疫基因组学研究报告规范”(Strengthening the Reporting of Immunogenomic Studies statement, STREIS)以 STREGA 为基础进行了扩展,专门说明免疫基因组学数据的相关问题,以加强研究的透明化报告和一致性解释。本文旨在介绍 STREIS 清单并对扩展条目进行解读。

**【关键词】** 免疫基因组学研究报告规范;免疫基因组学;人类白细胞抗原;杀伤细胞免疫球蛋白样受体

**How to write high-quality epidemiological research paper IV. Strengthening the Reporting of Immunogenomic Studies statement—introduction and explanation** CHEN Ru<sup>1</sup>, XU Chuan-hui<sup>2</sup>, ZHAN Si-yan<sup>1</sup>. 1 Department of Epidemiology and Biostatistics, School of Public Health, 2 Department of Rheumatology and Immunology, People's Hospital, Peking University Health Science Center, Beijing 100191, China

Corresponding author: ZHAN Si-yan, Email: siyan-zhan@bjmu.edu.cn

This work was supported by a grant from the Beijing Natural Science Foundation (No. 7111006).

**【Key words】** Strengthening the Reporting of Immunogenomic Studies statement; Immunogenomic; Human leukocyte antigen; Killer immunoglobulin like receptor

1. STREIS 的产生背景:免疫基因组学(immunogenomics)是基因组学(genomics)与免疫组学(immunomics)的结合,并与免疫蛋白质组学(immunoproteomic)和免疫信息学(immunoinformatics)共同构成免疫组学。随着人类基因组计划的完成,免疫基因组学迅速发展,其研究内容涵盖了免疫相关的全套分子及其作用靶分子的基因组学,而人类白细胞抗原(HLA)和杀伤细胞免疫球蛋白样受体(KIR)基因由于在适应性免疫和固有免疫中具有举足轻重的地位,是免疫基因组的核心。现代高通量的 HLA 和 KIR 分型技术产生了数量可观的免疫基因组学数据,将为组织相容性领域、免疫相关疾病和人群遗传研究带来深刻影响。

近年来一些论文报告指南,如“加强观察性流行病学研究报告的质量”(STROBE)<sup>[1]</sup>和“加强遗传关联性研究报告的质量”(STREGA)<sup>[2]</sup>的提出和应用,

极大推动了遗传流行病学领域研究的规范化报告。这些声明中描述的许多数据报告问题均与组织相容性和免疫遗传学研究有关,但主要是关于大型队列和基于单核苷酸多态性(SNP)的研究。免疫基因组学的特殊性,例如 HLA 和 KIR 基因位点的高水平多态性、各种 HLA 和 KIR 基因分型系统的使用、移植研究的复杂性和主要组织相容性复合体在疾病易感性上的特殊作用,都超出了 STROBE 和 STREGA 声明的报告标准和建议,需要额外考虑。

2. STREIS 的制定、目标和应用:在第 16 届国际 HLA 和免疫基因组学讨论会中,免疫基因组数据分析工作组(Immunogenomics Data Analysis Working Group, IDAWG)提出了 STREIS (STrengthening the REporting of Immunogenomic Studies)<sup>[3]</sup>,即“加强免疫基因组学研究的报告规范”。STREIS 声明是 STREGA 声明在免疫基因组学研究的扩展,将 STREGA 原有的 22 项条目至少增加 4 项,以专门说明免疫基因组学数据的相关问题。

STREIS 的目的并不在于规定免疫基因组学研究的设计方案,而是确保研究的一致性和透明化报告,促进不同研究间 HLA 和 KIR 数据的合并。

DOI: 10.3760/cma.j.issn.0254-6450.2013.011.022

基金项目:北京市自然科学基金(7111006)

作者单位:100191 北京大学公共卫生学院流行病与卫生统计学系(陈茹、詹思延);北京大学人民医院风湿免疫科(徐传辉)

通信作者:詹思延, Email: siyan-zhan@bjmu.edu.cn

STREIS 的推出受到免疫学界广泛关注,目前已被 *International Journal Immunogenetics* 等专业杂志的编辑所推荐<sup>[4]</sup>。本文对 STREIS 解读。

3. STREIS 的清单内容:STREIS 声明主要基于 STROBE 和 STREGA,扩展了 STREGA 在免疫基因组学研究中的标准,包括 4 个补充条目,主要涉及 HLA 和 KIR 的命名(条目 7)、数据验证(条目 8)、模糊度确定(条目 8、19)和高度多态性遗传系统的分析(条目 12、19)。表 1 比较了 STROBE、STREGA 和 STREIS 在四项条目中的内容。

4. STREIS 清单解读:

**条目 7(方法变量)** 国际免疫遗传学数据库 (IMGT Database)是人类主要组织相容性复合系统序列专题数据库,包括 WHO HLA 系统因子命名委员会正式认可并命名的全部 HLA 序列<sup>[5]</sup>。自 1999 年以来,每次 HLA 等位基因的序列定义和命名规则均在 IMGT/HLA 数据库中集成。该数据库每 3 个月更新一次,并为每次更新提供特定的版本号。

在过去 20 年中,HLA 的命名规则几经更改<sup>[6]</sup>。在 1987 年的命名规则中,使用 4 位数字区分编码不同蛋白质的 HLA 等位基因,前两位数字描述每个等位基因的血清特异性,后两位数字确定在该血清特异性等位基因内的连续的蛋白变异体。1990 年加入了第五个阿拉伯数字,代表同义核苷酸多态性。2002 年命名扩展到四对数字,分别代表血清学、蛋白质、同义和非编码核苷酸多态性。2010 年命名引入了冒号(:)以分隔不同的多态性区域,使得每个区域内可容纳超过 99 个变异体。

命名规则的改进可使某些等位基因的名称发生改变(例如 B\*1522 变为 B\*3543),但这些更改并不以

一致的方式应用到现有数据中,尤其是大型数据库的更新非常耗时,且频繁更新容易引入错误。这就导致在命名改变后的很长一段时间内,年代较早版本命名的 HLA 数据仍继续分析、发表和流传。由于 IMGT/HLA 数据库的版本号往往不会在报告中公布,在数据集合并/或添加新的基因型时,很可能纳入不相容的 HLA 的数据(例如同时包含 B\*1522 和 B\*3543 等位基因),给分析带来困扰。因此必须按照 WHO 命名委员会对 HLA 系统因子的命名法描述 HLA 等位基因,并明确与数据相关的 IMGT/HLA 数据库版本号。

与 IMGT 的 HLA 数据库一样,免疫多态性数据库 (Immuno Polymorphism Database, IPD)中的 KIR 数据库是 KIR 序列和术语信息的中央资料库,每个 IPD-KIR 数据库版本号是在该特定命名规则下,为一组特定的 KIR 等位基因序列命名提供参考。对于 KIR 基因,也需要按照 IPD-KIR 数据库进行描述,并明确与数据相关的 IPD-KIR 数据库版本号。

**条目 8(数据来源/测量)** 出版组织相容性、免疫遗传学和免疫基因组学研究的期刊应该采取联合数据归档政策,将出版的数据存储在公共数据库中,并允许个人访问和获得。

HLA 和 KIR 分型的多样化特点使相关的数据管理方法(包括免疫基因组学数据的确定、记录、编码、存储和交换等)千差万别。因此,必须描述用于储存、管理、验证基因型和等位基因数据以及用于数据分析的系统,以确定各研究组间数据的等效关系。

HLA 多态性的特点和基因分型方法的局限性导致了 HLA 基因位点的模糊性(ambiguity),即对于某个 HLA 位点,其样品的每种分型结果均包括多个基因型(基因型模糊性,genotype ambiguity),而每个基因

表 1 报告观察性流行病学研究——分子流行病学的条目清单

| 条目      | 条目序号 | STROBE                                              | STREGA                                                                                                                       | STREIS 扩展                                                                                                                                                                               |
|---------|------|-----------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 方法变量    | 7    | 明确定义所有结局、暴露、预测因子、潜在的混杂因子和效应修饰因子(如果可能,给出诊断标准)        | 需用公认的命名法清晰定义遗传暴露(遗传变异),确定与人群分层(种族混杂)有关的变量                                                                                    | 按照 WHO HLA 系统因子命名委员会的命名法描述 HLA 等位基因,明确与数据相关的 IMGT/HLA 数据库版本号;按照 IPD-KIR 数据库描述 KIR 等位基因,明确与数据相关的 IPD-KIR 数据库版本号                                                                          |
| 数据来源/测量 | 8    | 对每个关心的变量,描述其数据来源和详细判定(测量)方法(如果有多组,还应描述各组之间判定方法的可比性) | 描述实验方法,包括 DNA 来源与保存方法、基因分型方法平台(包括等位基因判定算法和版本)、错误率和检出率,并说明基因分型实验室/中心的名称;如果是多中心实验室,还须说明不同实验中心方法的可比性,说明是否从所有研究对象或是部分研究对象获得基因型数据 | 为每个人提供原始的模糊基因分型数据使用权;描述用于储存、管理、验证基因型和等位基因数据及数据分析准备的系统;使用客观条目,确定每个基因的评估功能,并描述基因分型系统及其分析结果。避免使用随时间改变的主观条目(例如,低、中、高或等位基因分辨率)来描述基因分型系统和结果;记录所有用于解决模糊性的方法;定义任何用来代表模糊性的代码;描述任何将等位基因合并为共同分类的措施 |
| 统计学方法   | 12   | 描述所有统计方法,包括控制混杂的方法                                  | 说明使用软件版本和选择的选项(或设置)                                                                                                          | 讨论分析方法及修改的数据;当分析涉及免疫基因组数据时,记录任何与其有关的注意事项                                                                                                                                                |
| 局限性     | 19   | 结合潜在偏倚和无法确定的来源,讨论研究的局限性和潜在偏倚的方向及其大小                 | -                                                                                                                            | 讨论修改数据可能对分析带来的影响;当分析涉及免疫基因组数据时,讨论任何与其有关的注意事项;讨论任何模糊性解决措施对结果的潜在影响                                                                                                                        |

涉及合并数据时必须描述等位基因的合并措施,以便从方法学上判断其合并的合理性和结果真实性。在比较和合并不同研究结果及大规模的

(2)低频率等位基因:低频率等位基因(频率<2%,并取决于该数据集的大小)是典型的高度多态性免疫基因组学数据,处理不正确可导致虚假或无意义的结果。对于明确分配假设的检验(如 $\chi^2$ 检验),低频率等位基因和基因型影响其检验统计量的

渐近近似有效性。虽然低频等位基因可能不影响个体 SNP 或其他限制次要等位基因频率纳入的低水平多态性位点分析,但在处理 HLA、KIR 和 SNP 单倍型等多态性位点时必须加以重视。

(3)单倍型估计:在许多遗传学研究中,单倍型和单倍型频率估计具有核心作用,但免疫基因组学数据的多样性和复杂性给单倍型估计带来挑战。对 KIR 和 HLA 而言,等位基因频率、数据集样本大小、缺失信息的水平和连锁不平衡水平均影响估计的准确性。此外高水平的等位基因多态性可能导致估计的单倍型参数多于实际观察的表型数目。

(4)连锁不平衡:即相邻等位基因座上等位基因的非随机性相关。标准的成对连锁不平衡测量有多等位基因的扩展,但这些测量可能隐含对称假设而无法发现重要的关联信息。例如, $W_d$ 是 $r^2$ 连锁不平衡关联测量的多等位基因扩展,它对于两个位点是对称的,但在每个位点上报告的等位基因数目可能差别很大,从而导致不对称。其他常用的对连锁不平衡的测量,如  $D'$ ,也对罕见等位基因的存在具有高度敏感性,这是免疫基因组学位点的一大特征。解决的关键是从多种连锁不平衡测量探索结果,而对大量非对称等位基因中一对基因的特定连锁不平衡测量不要过分解释。

(5)关联研究:免疫基因组学数据的特点要求慎重考虑疾病的关联研究。低频率等位基因对传统疾病研究的关联检验,是对比较病例组和对照组使用 $\chi^2$ 检验方法的挑战。 $\chi^2$ 检验是整体位点水平检验的标准方法,但当预期的基因型在列联表中的频数很低时,这种检验可能导致错误或拒绝零假设。标准的做法是建立低频率等位基因的合并分类,也可以在分析中完全排除含有低频率等位基因的基因型,或对基因进行分级分析,也可以使用非参数重采样(nonparametric resampling)和精确检验。此外,对免疫基因组学位点的大量等位基因多重比较需做特定修正,这些校正取决于 GWAS 测试的标志物数目,也可能受到各免疫基因组学关联性研究标志物检测等位基因数目的影响。

潜在的人口分层也是免疫基因组学数据的关联研究需要着重考虑的问题,因为自然选择和人口学特征决定了特定种族和地域内群体的等位基因频率分布。如果样品收集时未注意到家系背景的同质性,其结果可能对病例和对照遗传差异作出错误的解释,人群分层相关的等位基因频率分布差异也可被误认为与一个特定的位点有关联。而全基因组标

记物可用于检测该分层问题。

(6)H-W 遗传平衡检验:许多分析方法(如单倍型的估计)均以数据集的 H-W 平衡假设为前提,因此 H-W 遗传平衡检验的准确性对结果分析至关重要;在高度多态性数据集中,H-W 检验尤其具有挑战性。关联研究中, $\chi^2$ 检验已成为适合 H-W 平衡比例的检验标准。尽管观察到的最小基因型数目必须是一个,但预期的基因型最小数目可能远小于 1,因此仅观察一次且预期计数 $<1$ 的基因型可能造成明显偏离 H-W 遗传平衡的情况。采取以下三种方法可提高检验的准确性:①低频率等位基因可以合并统一,并有效采用 $\chi^2$ 检验;②采取精确检验,列出所有完整表格,包括所有可能与观察到的等位基因频率一致的基因型;③通过重采样获得上述完整的近似值。

(7)进化分析:应用等位基因和单倍型频率生成的进化树(或树状图)分析种群关系,检验与人类历史和迁移相关的假设。然而分析时,由于未考虑到等位基因之间序列差异的程度,结果将血缘相关和不相关的等位基因均以相同方法评价;在大多数进化频率的分析中,忽略了源于序列的等位基因亲疏关系的信息。对于血缘相关的低频率等位基因,可采用低于外显子水平的等位基因名称合并方法,但该方法不适用于更常见的等位基因。

6. 结语:将 STREIS 原则应用于免疫基因组学数据处理和分析的时机已到来。与 STREGA 相同,STREIS 目的并非规定免疫基因组学研究的设计,而是要求研究结果的一致性及其报告的透明,以促进 HLA 和 KIR 数据在研究中的整合。本文更多信息可参考 IDAWG 网站([www.immunogenomics.org](http://www.immunogenomics.org))。

## 参 考 文 献

- [1] von Elm E, Altman DG, Egger M, et al. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*, 2007, 370(9596): 1453-1457.
- [2] Little J, Higgins JP, Ioannidis JP, et al. STrengthening the REporting of Genetic Association studies (STREGA)—an extension of the STROBE statement. *Eur J Clin Invest*, 2009, 39(4): 247-266.
- [3] Hollenbach JA, Mack SJ, Gourraud PA, et al. A community standard for immunogenomic data reporting and analysis: proposal for a STrengthening the REporting of Immunogenomic Studies statement. *Tissue Antigens*, 2011, 78(5): 333-344.
- [4] Travers PJ, Middleton D. Community standards: a new policy to ensure consistency and transparency in immunogenomics. *Int J Immunogenet*, 2012, 39(4): 279-281.
- [5] Lefranc MP, Giudicelli V, Regnier L, et al. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform*, 2008, 9(4): 263-275.
- [6] Marsh SG, Albert ED, Bodmer WF, et al. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 2010, 75(4): 291-455.

(收稿日期: 2013-10-13)

(本文编辑: 张林东)