收稿日期：制作时添加 修订日期：制作时添加 接收日期：制作时添加DOI：xxx/xxxx

文章类型

采用广义成对比较方法分析阶梯楔形随机整群试验：一项模拟研究

Yohan Bard[1] Emilie Presles[2] Marc Buyse[3] Silvy Laporte[4] Paul Zufferey[5] Frederikus A. Klok[6] Olivier Sanchez[7] Francis Couturaud[8,9] Edouard Ollier[4]

[1]法国国家健康与医学研究院 U1059 单位 SAINBIOSE，法国圣艾蒂安

[2]临床研究中心 CIC 1408，圣艾蒂安大学医院，法国圣艾蒂安

[3]国际药物开发研究所（IDD），生物统计学，新鲁汶；哈瑟尔特大学 I-BioStat，哈瑟尔特，比利时

[4]圣艾蒂安大学医院临床研究单元（圣艾蒂安大学医院）；法国国家健康与医学研究院 U1059 SAINBIOSE，圣艾蒂安让·莫内大学，圣艾蒂安，法国

[5]麻醉与复苏科，北部医院，圣艾蒂安大学医院，F-42055，圣艾蒂安，法国

[6]莱顿大学医学中心（LUMC）医学系，莱顿，荷兰

[7]呼吸医学与重症监护科，巴黎西岱大学，乔治·蓬皮杜欧洲医院，巴黎公立医院集团，法国国家健康与医学研究院 UMR-S 970，F-CRIN INNOVTE，巴黎，法国

[8]肺病科，布雷斯特大学医院（布雷斯特大学医院），布雷斯特，法国

[9]法国国家健康与医学研究院 U1304 GETBO（布列塔尼西部血栓研究组），布雷斯特大学，F-CRININNOVTE，法国布雷斯特

通信作者
通信作者 Yohan Bard，电子邮箱：yohan.bard@inserm.fr

现地址
皮埃尔·孟戴斯-弗朗斯大道28号，42 270 圣普里斯特昂雅雷

arXiv:2603.02003v1 [stat.ME] 2 Mar 2026

摘要

阶梯楔形整群随机试验（SW-CRTs）越来越多地用于评估复杂干预，但目前关于使用广义成对比较（GPC）分析复合终点的方法学指导仍然有限。本研究探讨了在阶梯楔形整群随机试验典型的聚类、时间趋势及不同相关结构下，若干基于GPC的估计量的表现。我们开展了一项广泛的模拟研究，考察了一系列组内相关系数（ICC）、整群自相关系数（CAC）、时间效应和治疗效应大小。我们比较了八种分析方法，包括未调整估计量、整群分层胜算、应用于整群–期间胜算的混合效应模型，以及概率指数模型（PIM）。忽略时间或聚类的方法在控制I类错误方面表现明显不足，而只有两种方法能够始终维持名义错误率：包含序列层级和集群层级随机斜率的分层混合效应模型（b4），以及整群限制概率指数模型（c2）。我们进一步评估了这两种方法的统计功效。总体而言，c2的效率较高，尤其是在聚类程度强、CAC较低或存在时间趋势时；而当治疗效应较大时，两种方法的表现趋于相似。总之，我们的研究结果表明，b4和c2是用于阶梯楔形整群随机试验分析的最可靠的基于GPC的策略，并为其实际应用提供了指导，包括应用于正在进行的ETHER等试验。

关键词
阶梯楔形整群随机试验，广义成对比较，获胜优势比，混合效应模型，概率指数模型。

1 引言

广义成对比较（GPC）是一种现代统计方法，通过同时考虑多重结局来评估干预的效果。这种方法弥补了传统分析方法的局限性，例如经典复合终点会为每个组成事件赋予相同权重，即使这些事件在临床重要性上存在很大差异，

 © 2023 版权持有人名称

例如致死事件与非致死事件。因此，传统复合终点可能会忽略临床上有意义、但发生频率较低或发生时间晚于较不重要事件的事件[1]。

广义成对比较（GPC）允许研究者处理复合终点：当各组成部分在临床上并不等效时，可根据其临床重要性对结局进行排序。随后，可通过明确的汇总指标量化干预效应，例如胜率比。胜率比正日益普及，尤其是在心血管试验中[2]。Buyse 首先提出了这一方法[3]。广义成对比较框架通过比较治疗组之间所有可能的患者配对来开展分析。对于每一对患者，比较其结局，以判定其中一名患者的结果是较好（有利）、较差（不利），还是相近（中性）。随后，该方法汇总这些成对比较结果，以估计治疗效应。近年来，这一框架已扩展到多种情形，包括剂量–反应分析[4]、使用概率指数模型（PIM）的协变量调整[5,6]，以及变量选择策略[7]，，体现了其灵活性。然而，迄今为止，在临床研究中，广义成对比较主要用于个体随机试验。

在某些情况下，个体随机化并不可行，尤其是当不同组别的干预可能相互影响、即存在污染风险时。在这类情形下，出于实际、伦理或后勤方面的考虑，通常更倾向于采用整群随机化。这种方法随机分配的是参与者群体（或“集群”），而非个体，有助于避免污染、简化试验组织工作并提高参与者的参与度[8]。

然而，在整群试验中直接应用胜率比面临方法学挑战。胜率比的标准形式假设观测值相互独立，若忽略集群内结局的相关性，可能会低估方差并产生过于乐观的置信区间，从而影响统计推断的有效性。例如，Romiti 等人[9] 在整群随机环境中应用了广义成对比较框架，却未明确针对组内相关进行调整。他们的研究展示了当前分析工具在实践中的应用，也凸显了开发专为整群数据设计的方法的必要性。

当难以在所有集群中同时实施干预时，阶梯楔形整群随机试验（SW-CRT）提供了一种实用的替代方案。在这种设计中，集群会在连续的各个时期逐步从对照转为干预。这种分阶段推广在伦理方面具有优势——所有集群最终都能接受干预——同时也便于实施。此外，人们还提出了分批阶梯楔形设计等变体，以适应招募或干预推广延迟的情况，进一步提高了这种设计的实用性[10]。

然而，分析阶梯楔形整群随机试验的数据还面临其他挑战。首先，同一集群内的个体结局往往具有相关性，通常通过组内相关系数（ICC）量化，反映共同的环境或组织因素。其次，由于集群会在一段时间内被反复观测，同一集群内不同时期之间也可能存在持续相关性；这种集群效应的时间持续性可通过整群自相关系数（CAC）表征。此外，由于各集群在不同时间开始接受干预，基线风险的时间趋势可能会混杂治疗效应。因此，统计模型必须谨慎地同时校正聚类和时间，以避免估计产生偏倚[11,12]。此外，也已发布阶梯楔形整群随机试验的报告指南，以确保清晰描述此类设计的复杂特征[13]。

尽管广义成对比较（GPC）在临床研究中的应用日益广泛，但迄今为止，其应用仍主要局限于个体随机试验。迄今尚无方法学研究将 GPC 扩展至复杂随机设计（如经典整群随机试验或阶梯楔形变体），以明确考虑聚类或时间效应。本文通过将广义成对比较框架扩展至阶梯楔形场景，着手弥补这一方法学局限，使治疗效应分析能够适当考虑组内相关和长期时间趋势。

2 方法

2.1 阶梯楔形整群随机试验设计

阶梯楔形整群随机试验（SW-CRT）是一种设计：各集群按照随机序列，在不同的预先指定时间段从对照转为干预。该设计按离散时期推进，以$j = 1、\ldots、J$为索引，并在这些时期观察结局。集群以$k = 1、\ldots、K$为索引。在每个集群-时期组合（j，k）中，观察到的个体数量记为n_{jk}。阶梯楔形设计的一般结构如图 1 所示。

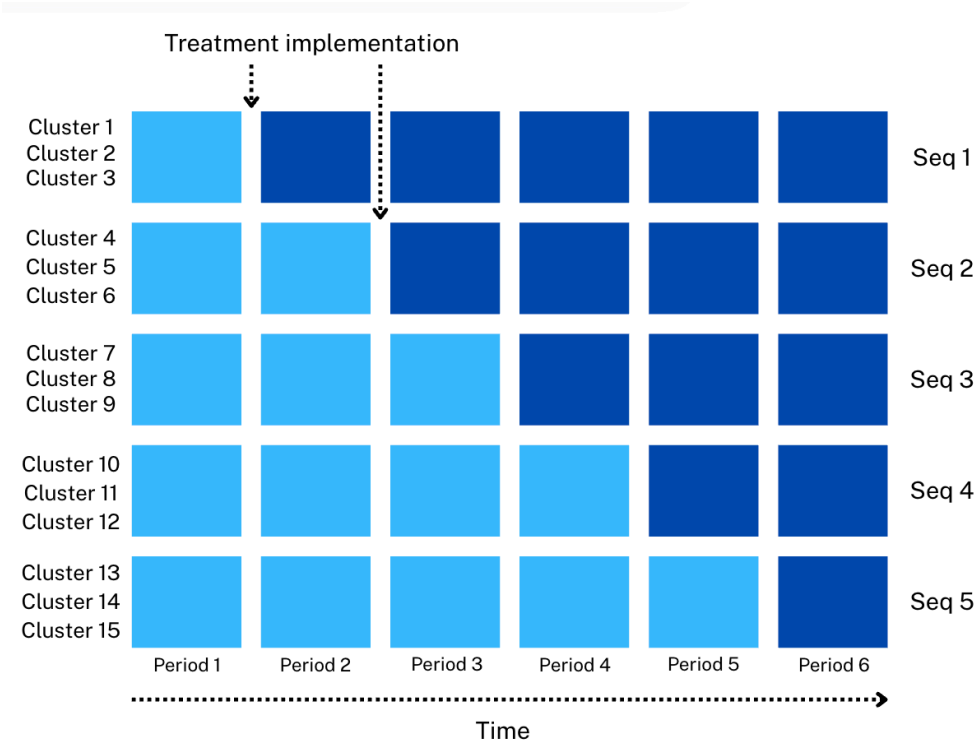

图 1 含 5 个序列（6 个时期）的阶梯楔形试验示例

每个集群都遵循指定的序列，该序列决定其开始接受干预的时期。研究结束时，所有集群均已接受干预，因此整个试验在两种条件下都有可用数据。因此，某个集群在j时期的干预状态完全由其序列分配和时期索引决定。

尽管本文着重讨论阶梯楔形设计，但本节介绍的符号表示和方法学进展自然也涵盖经典整群随机试验这一特殊情形。具体而言，平行组整群试验相当于只有一个转换时期的阶梯楔形设计，因此，所提出的框架也适用于阶梯楔形场景以外的情况。

阶梯楔形整群随机试验中存在两个主要的相关性来源。首先，同一集群内测量的结局往往会因临床实践、患者群体或组织因素相同而彼此相关。这种群组内相关性通常用组内相关系数（ICC）概括，该系数用于量化结局变异中可归因于群组间差异的比例。其次，由于会随时间反复观察集群，相关性可能会持续存在于

同一集群内的各个时期之间。在阶梯楔形设计中，这种时间持续性往往并不完美，并由整群自相关系数（CAC）描述，该系数反映集群特异性效应在连续各时期之间保持稳定的程度。

除相关结构外，阶梯楔形试验还会固有地受到长期时间趋势的影响，即结局的基线风险会随日历时间变化，而不受干预影响。由于在阶梯楔形整群随机试验中，干预暴露与时间系统性混杂，若未能考虑此类趋势，可能会导致治疗效应估计产生偏倚。这些特点凸显了采用统计方法同时调整聚类、时间相关和时间趋势的必要性，后续章节将对此进行讨论。

2.2 复合结局的广义成对比较

广义成对比较（GPC）提供了一个统一框架，可基于个体层面的结局评估治疗效应。该方法考察所有可能的个体配对，即分别从治疗组和对照组中各选一名个体，并依据预先设定的终点排序，判断哪名个体的结局更有利。

令Y_i表示个体i的结局。若考虑单个终点，则Y_i为标量。对于多重结局这一更一般的情况，我们定义一个向量

$$Y_i = \left(Y_i^{(1)}, \ldots, Y_i^{(M)}\right),$$

结局根据其临床相关性按层级排序：$Y^{(1)}{}_i$是最重要的终点，而$Y^{(M)}{}_i$是最不重要的终点，通常因其区分能力而选用。

治疗分配由指示变量X_{ijk}描述：若集群k在时期j中的个体i接受干预，则其值为 1；否则为 0。

对于分配至不同治疗组的两个个体i和i'，我们定义比较得分$W_{ii'}$。第一个出现差异的结局的索引为

$$m^* = \min\left\{m \in \{1, \ldots, M\} : Y_i^{(m)} \neq Y_{i'}^{(m)}\right\}.$$

若存在这样的m^*，则：

$$W_{ii'} = \begin{cases} 1, & \text{if } Y_i^{(m^*)} \succ Y_{i'}^{(m^*)}, \text{(favorable pair)} \\ -1, & \text{if } Y_i^{(m^*)} \prec Y_{i'}^{(m^*)}, \text{(unfavorable pair)} \\ 0, & \text{if no such } m^* \text{ exists (neutral pair).} \end{cases}$$

基于此比较得分，可使用不同指标估计干预效应。令N_W和N_L分别表示治疗组个体与对照组个体之间所有成对比较中的有利配对总数和不利配对总数。第一个指标是胜率比，定义为

$$\mathrm{WR} = \frac{N_W}{N_L}.$$

数值大于 1 表明治疗组获得的有利配对多于不利配对，说明干预具有有益效果。

由于中性配对可能频繁出现，第二个指标获胜优势比通过将一半有利配对和一半不利配对分配给每个中性配对，对胜率比进行了修正。令N_T表示中性配对的数量。获胜优势比定义为

$$\mathrm{WO} = \frac{N_W + 0.5\,N_T}{N_L + 0.5\,N_T}.$$

尽管混合效应模型中的治疗效应以 log(WO)尺度表示，但该指标与净治疗获益（NTB）的关系使其具有简单直观的解释。NTB 定义为

$$\mathrm{NTB} = \frac{\mathrm{WO} - 1}{\mathrm{WO} + 1},$$

它对应于概率之差（$\mathbb{P}_T > \mathbb{P}_C$）−（$\mathbb{P}_C > \mathbb{P}_T$），即随机选取的一名接受治疗的个体，其结局优于一名对照个体的概率减去反向概率。该数值介于 –1 和 1 之间，可解释为概率尺度上的绝对治疗效应。对于小到中等的治疗效应，一阶近似为 $\log(\mathrm{WO}) \approx 2 \times \mathrm{NTB}$，这有助于解释以对数胜算比尺度表示的回归系数。

尽管这两种指标都符合广义成对比较框架，但在中性配对数量不可忽略的情况下，通常更倾向于使用胜算比。将中性配对纳入计算后，该指标能够提供更平滑、稳健的治疗效应估计，同时保留相同的可解释性：数值大于 1 表示治疗组具有优势。此外，胜算比与较为经典的干预效应估计指标相关。例如，在只有一个结局的特殊情况下，胜算比可解释为风险比的倒数[14]。因此，本文将主要考察胜算比这一效应指标。

2.3 广义成对比较中的协变量调整

在阶梯楔形整群随机试验中，潜在混杂可能来自多个方面：组内结局聚类、长期时间趋势，以及集群或时期层面的特征。本研究的一项主要动机是，若不针对这些潜在混杂来源进行调整，广义成对比较可能会产生有偏的治疗效应估计。为解决这些问题，我们用 X 表示需要考虑的协变量集合，其中可能同时包括个体层面和集群层面的变量。在广义成对比较框架中，可以采用不同方法将协变量纳入考量。

2.3.1 分层

最简单的调整策略是根据分类变量 $S = S(X)$ 对成对比较进行分层。例如，S 可以表示集群或试验时期。在每个层 s 内，仅考虑属于该层的个体之间的比较，并估计层特异性获胜优势比 WO_s。

这些层特异性 估计值随后通过对其对数加权合并为一个总体指标[15,16]：

$$\log(\mathrm{WO}) = \sum_s w_s \log(\mathrm{WO}_s),$$

权重为 w_s s，且满足 $\sum_s w_s = 1$。可以采用不同的加权方案，例如逆方差权重，或与各层大小成正比的权重。这种分层方法与经典分层检验相似，同时保留了效应指标基于获胜的解释。

2.3.2 基于概率指数模型的回归框架

为调整广义成对比较分析中潜在的连续型或分类协变量 X，研究者提出了概率指数模型（PIM），将广义成对比较纳入回归框架。其核心思想是，胜算比可以表示为概率指数的一种变换，概率指数定义如下：

$$\mathbb{P}\left(Y \succ Y_\star \mid X, X_\star\right) = \mathbb{P}\left(Y > Y_\star \mid X, X_\star\right) + 0.5\,\mathbb{P}\left(Y = Y_\star \mid X, X_\star\right),$$

也就是说，在给定各自协变量X 和 $X_\star$的情况下，随机选取的一个个体具有比另一个随机选取的个体更有利的结局的概率。

在概率指数模型中，该概率通过回归模型与协变量相关联：

$$g\left(\mathbb{P}\left(Y \succ Y_\star \mid X, X_\star\right)\right) = Z^T \beta,$$

其中，$g(\cdot)$是连接函数（逻辑斯蒂、概率单位、恒等）。向量Z 表示一个向量，其元素取决于X 和$X_\star$ ，例如$Z = X - X_\star$ 。参数向量 β 包含回归系数，用于描述各协变量如何影响有利比较的概率；其中，治疗系数直接对应调整后的对数获胜优势。对于阶梯楔形整群随机试验的分析，这种成对回归形式既能直接调整个体和整群层面协变量以及时间效应，又保留了成对比较原则。

可通过标准估计方程进行估计[17]。拟合模型可得出经调整的胜算比治疗效应估计值，并提供置信区间和 p 值。与分层相比，这种回归方法能够直接处理连续型和/或多个协变量。

2.4 基于广义成对比较分析阶梯楔形整群随机试验的方法评估

本节介绍采用广义成对比较（GPC）框架分析阶梯楔形整群随机试验（SW-CRT）的候选方法。每种方法都能估计获胜优势比（WO），同时以不同程度处理时间依赖性和数据整群所引入的混杂。

2.4.1 未调整时间趋势的方法

第一类评估方法有意忽略阶梯楔形设计的时间结构。该比较将试验视为平行组设计，并按照治疗状态汇总所有观察结果。

第一种方法称为**方法 a1：粗获胜优势**，它将所有接受干预的个体（不考虑时期或集群）与所有仍处于对照状态的个体进行比较。生成所有可能的治疗-对照配对，并将总体胜算比估计为

$$\mathrm{WO} = \frac{N_W + 0.5N_T}{N_L + 0.5N_T},$$

其中，N_W、N_L和N_T 分别表示有利配对、不利配对和中性配对的总数。该方法较为简单，但忽略了整群内相关和时间依赖效应。

一种改进方法称为**方法 a2：整群分层获胜优势**，通过分别计算每个集群内的获胜优势比来考虑聚类。在每个集群k中，将对照组个体与干预组个体进行比较，从而得到整群特异性获胜优势 WO_k。随后，将这些估计值汇总至各集群，以获得总体测量值。汇总在对数尺度上进行，以稳定方差；总体估计值为

$$\log(\mathrm{WO}) = \sum_{k=1}^{K} w_k \, \log(\mathrm{WO}_k),$$

其中，w_k 为满足 $\sum^K{}_{k=1} w_k = 1$的非负权重。

设m_k 和n_k 分别表示集群k中对照组和干预组的样本量。该集群内可形成的治疗-对照配对总数为

$$M_k = m_k \times n_k.$$

一种自然的选择是令w_k 与M_k成正比，即

$$w_k = \frac{M_k}{\sum_{\ell=1}^{K} M_\ell}.$$

此处，M_k 表示集群k中可用治疗–对照比较的总数，不论这些配对之后产生的是有利结果、不利结果还是平局结果。

该加权方案[3,16]确保贡献更多信息的集群对总体估计值具有更大的影响，同时避免使用逆方差权重可能导致的不稳定性（因为当某个集群内的所有比较产生相同结局时，方差有时可能为零）。

方法（a1）和（a2）易于实施且计算效率高。然而，由于这两种方法未对时间趋势进行调整，因此在阶梯楔形整群随机试验中可能存在偏倚，因为干预暴露与日历时间相混杂。

2.4.2 期间间的集群层面分析

为同时考虑聚类和时间动态，第二种方法在整群–期间对层面构建获胜优势比。具体而言，对于每个集群$k \in \{1,\ldots,K\}$以及每一对时期（j_1, j_2），其中$j_1 < j_2$，，在时期j_1期间于集群k中观察到的个体与时期j_2期间在同一集群中观察到的个体之间，进行所有可能的成对比较。在这些比较中，根据前文介绍的广义成对比较（GPC）框架，将结局分为有利、不利和中性。令$NW_{k,(j_1,j_2)}$、$NL_{k,(j_1,j_2)}$和$NT_{k,(j_1,j_2)}$分别表示这些集群内比较所得的有利、不利和中性配对数。

随后将集群k在期间（j_1,j_2）之间的胜算比定义为

$$\mathrm{WO}_{k,(j_1,j_2)} = \frac{N_{W_{k,(j_1,j_2)}} + 0.5\,N_{T_{k,(j_1,j_2)}}}{N_{L_{k,(j_1,j_2)}} + 0.5\,N_{T_{k,(j_1,j_2)}}}.$$

由于每个集群在不同期间对中会提供多个$\mathrm{WO}_{k,\ (j_1,j_2)}$估计值，因此随后必须对这些估计值进行合并，以获得总体干预效应，同时考虑时间趋势和整群间变异。为此，采用混合效应模型对整群–期间胜算比的对数$\log(\mathrm{WO}_{k,\ (j_1,j_2)})$进行建模，其中包含时间和干预的固定效应，以及用于捕捉集群和期间间残余异质性的随机效应。

我们首先考虑了一个基线设定，称为 **模型b1**，其中包括集群和期间对的随机截距。该模型既考虑了集群间未观测异质性，也考虑了涉及同一期间对的比较之间可能存在的相关性：

$$\log\left(\mathrm{WO}_{k,(j_1,j_2)}\right) = \alpha + u_k + v_{(j_1,j_2)} + \delta \times (X_{k,j_2} - X_{k,j_1}) + \gamma \times (j_2 - j_1) + \varepsilon_{k,(j_1,j_2)},$$

其中，$u_k \sim \mathcal{N}(0,\sigma_C^2)$是集群$k$的随机截距，而$v_{(j_1,j_2)} \sim \mathcal{N}(0,\sigma_P^2)$是另一个随机截距，用于捕捉特定于期间对（$j_1$,$j_2$）的变异。

尽管这一结构捕捉了集群间的基线异质性，但它假设所有集群的干预效应均为恒定值。为放宽这一假设，**模型b2** 在整群水平引入了治疗差异的随机斜率。这样，每个集群都可以有一个偏离总体治疗效应的自身偏差，从而考虑到情境或实施因素可能导致干预效应在不同集群间有所差异：

$$\log\left(\mathrm{WO}_{k,(j_1,j_2)}\right) = \alpha + u_k + v_{(j_1,j_2)} + (\delta + z_k) \times (X_{k,j_2} - X_{k,j_1}) + \gamma \times (j_2 - j_1) + \varepsilon_{k,(j_1,j_2)},$$

其中，$z_k \sim \mathcal{N}(0,\sigma_{Int}^2)$表示集群$k$的治疗效应随机偏差。

尽管模型 b2 允许治疗效应因集群而异，但它并未明确考虑阶梯楔形结构，即集群按照交叉转换序列进行组织。为反映这些序列间治疗效应的潜在差异，**模型 b3** 在序列水平引入了随机斜率。这一设定承认，属于同一序列的集群具有相似的暴露模式，因此对干预可能表现出相关响应：

$$\log\left(\mathrm{WO}_{k,(j_1,j_2)}\right) = \alpha + u_k + v_{(j_1,j_2)} + (\delta + w_{s(k)}) \times (X_{k,j_2} - X_{k,j_1}) + \gamma \times (j_2 - j_1) + \varepsilon_{k,(j_1,j_2)},$$

其中，$w_{s(k)} \sim \mathcal{N}(0,\sigma_S^2)$表示与集群$k$所属序列$s(k)$相关的治疗差异随机斜率。

基于这一思路，**模型 b4** 提出了一种更灵活的层级设定，将集群层级随机斜率嵌套于序列中。该模型既捕捉了不同序列间治疗效应的共同变异，也捕捉了同一序列内集群之间的残余异质性：

$$\log\left(\mathrm{WO}_{k,(j_1,j_2)}\right) = \alpha + u_k + v_{(j_1,j_2)} + (\delta + w_{s(k)} + z_k) \times (X_{k,j_2} - X_{k,j_1}) + \gamma \times (j_2 - j_1) + \varepsilon_{k,(j_1,j_2)},$$

其中，$w_{s(k)} \sim \mathcal{N}(0, \sigma_S^2)$ 为序列层级随机斜率，$z_k \sim \mathcal{N}(0, \sigma_{ClS}^2)$ 为嵌套于同一序列中的集群层级偏差。这种结构使模型能够捕捉序列之间的大尺度差异，以及集群之间更细微的异质性。

由于阶梯楔形整群随机试验中的集群数量通常有限，标准方差估计量和渐近近似可能导致推断出现偏倚。系统地应用了 Kenward–Roger 校正[18]，以纠正小样本偏倚和 I 类错误率。

2.4.3 个体层级分析

第三种策略是采用概率指数模型（PIM）框架，直接在个体成对比较层面分析数据。在这种设定下，每条观测对应于来自集群k、在时期j观察到的个体i，其结局为Y_{ijk}，治疗指示变量为X_{ijk}，观察时间为t_{ijk}。PIM 方法能够对给定各自协变量时，一名个体的结局优于另一名个体的概率进行建模，从而将获胜优势比的概念纳入回归框架[5]。

在最简单的设定中，即**模型 c1（标准概率指数模型）**，设（i、j、k）和（i'、j'、k'）表示从试验中抽取的两个个体。该模型定义为

$$g\left(\mathbb{P}\left(Y_{ijk} \succ Y_{i'j'k'} \,\middle|\, X_{ijk}, X_{i'j'k'}, t_{ijk}, t_{i'j'k'}\right)\right) = \alpha + \delta \times (X_{ijk} - X_{i'j'k'}) + \gamma \times (t_{ijk} - t_{i'j'k'}),$$

其中$g(\cdot)$为连接函数，通常为logit连接函数。

在这一表述中，δ 以获胜优势比的形式量化调整后的治疗效应，而 γ 则捕捉个体层面的线性时间趋势。考虑所有可能的个体配对，包括治疗-治疗配对、对照-对照配对以及治疗-对照配对。同一治疗组内的比较有助于估计时间效应，而跨组比较则同时为治疗效应和时间效应提供信息。由于利用了数据集中的所有可用信息，这一表述具有灵活性和高效性。

然而，标准概率指数模型估计假设观测值之间相互独立，或至少对轻度依赖具有稳健性。在阶梯楔形整群随机试验中，数据的层级结构会引发强烈的整群内相关，因为同一集群中的个体会受到共同情境因素和时间因素的影响。尽管概率指数模型框架通常对中等程度的相关结构具有稳健性，但忽略这种聚类可能导致低估方差并使I类错误率膨胀。

为解决这一问题，一种称为**模型 c2（整群限制概率指数模型）**的替代设定将比较限制在属于同一集群的个体之间，即$k = k'$。此时模型变为

$$g\left(\mathbb{P}\left(Y_{ijk} \succ Y_{i'j'k} \,\middle|\, X_{ijk}, X_{i'j'k}, t_{ijk}, t_{i'j'k}\right)\right) = \alpha + \delta \times (X_{ijk} - X_{i'j'k}) + \gamma \times (t_{ijk} - t_{i'j'k}).$$

通过将比较限制在集群内，该方法消除了簇间异质性，并确保估计的治疗效应完全基于集群内对比。这种结构与阶梯楔形整群随机试验的随机化机制直接吻合，因为治疗分配是在整群水平进行的，同时有助于避免跨集群比较可能导致的偏倚。

总体而言，两种概率指数模型设定都提供了经调整的治疗效应估计，并以获胜优势表示。标准概率指数模型（c1）利用所有可用的治疗-对照配对，这可能提高效率，但对群水平相关也可能更为敏感。相比之下，整群限制概率指数模型（c2）采取了更保守但稳健的方法，更好地遵循阶梯楔形设计的层级结构，并减轻群内依赖性的影响。

2.5 模拟研究

为评估所提出方法的统计性能，我们基于阶梯楔形整群随机设计开展了一项模拟研究。研究旨在评估不同分析方法在聚类程度、时间相关性和治疗效应各异的情况下表现如何，同时保持与典型阶梯楔形整群随机试验结构的一致性。

我们考虑了一种设计，其中$K = 45$个集群分布在5个序列中，各集群在$J = 6$个时期内依次从对照转为干预。每个集群的每个时期均观察$n = 10$个个体。这些数值的选择依据是，它们符合下文讨论的ETHER试验的特征。为每个k集群在第j个时期的i个体分配一个连续观测时间t_{ijk}，并在相应时期的时间区间内均匀抽取。

个体二元结局Y_{ijk}根据成功概率为π_{ijk}的伯努利分布生成：

$$Y_{ijk} \sim \text{Bernoulli}(\pi_{ijk}),$$

且成功概率遵循如下形式的logistic模型

$$\text{logit}(\pi_{ijk}) = \log\left(\frac{p_0}{1-p_0}\right) + \delta \cdot X_{ijk} + \eta_k + \zeta_{jk} + \beta_t\, t_{ijk}.$$

此处，p_0表示对照条件下的基线概率，X_{ijk}为治疗指示变量（若个体接受干预，则$X_{ijk} = 1$，否则为0），β_t表示用于刻画长期趋势的连续时间效应。

引入随机效应η_k和ζ_{jk}以刻画群内依赖性。第一项$\eta_k \sim \mathcal{N}(0, \sigma^2_{\text{cluster}})$表示持续存在的集群层级偏差，它会在同一集群的所有个体之间诱发相关性。第二项$\zeta_{jk} \sim \mathcal{N}(0, \sigma^2_{\text{period}})$用于解释特定集群的时期间变异，从而在同一时期内观察到的个体之间引入额外的依赖性。两个随机效应假定相互独立。

在此设定下，同一集群中两个个体之间的总相关性可通过组内相关系数（ICC）表示：

$$\text{ICC} = \frac{\sigma^2_{\text{cluster}} + \sigma^2_{\text{period}}}{\sigma^2_{\text{cluster}} + \sigma^2_{\text{period}} + \pi^2/3},$$

其中$\pi^2/3$为逻辑斯蒂分布[19]的方差。因此，组内相关系数量化了可归因于共享集群层级效应的方差总体比例。

在阶梯楔形试验中，同一批集群会在多个时期内被反复观测，其相关结构也会随时间变化。这种时间依赖性可用整群自相关系数（CAC）量化，定义为给定集群内各时期之间保持不变的组内相关系数所占比例。因此，CAC控制着集群效应随时间的持续性：

$$\text{CAC} = \frac{\sigma^2_{\text{cluster}}}{\sigma^2_{\text{cluster}} + \sigma^2_{\text{period}}},$$

在模拟研究中，每个随机效应的方差均根据预先设定的组内相关系数和CAC值[20]按以下方式计算：

$$\sigma^2_{\text{cluster}} = \text{CAC} \times \text{ICC} \times \frac{\pi^2/3}{1-\text{ICC}}, \qquad \sigma^2_{\text{period}} = (1-\text{CAC}) \times \text{ICC} \times \frac{\pi^2/3}{1-\text{ICC}}.$$

这种参数化确保同一集群、同一时期内两个个体之间的相关性等于ICC，而同一集群但处于不同时期的两个个体之间的相关性等于$\text{CAC} \times \text{ICC}$。当$\text{CAC} = 1$时，

集群效应在时间上完全稳定，对应于标准的横断面 ICC。随着 CAC 降低，集群效应在不同时期之间的变异增大，反映出集群内时间相关性的逐渐减弱。

每种模拟情景均由一组独特的参数组合定义，这些参数控制基线风险、治疗效应、相关结构和时间趋势。具体而言，我们改变了基线事件概率 $p_0 \in \{0.01, 0.1, 0.2, 0.5\}$、组内相关系数 ICC $\in \{0, 0.05, 0.1, 0.3\}$、簇自相关 CAC $\in \{0.5, 0.75, 0.9, 1\}$ 以及时间效应 $\beta_t \in \{-0.05, -0.025, 0, 0.025, 0.05\}$。治疗效应通过对数优势比参数 $\delta \in \{0, 0.25, 0.5, 1, 1.5\}$ 设定。对于每种参数配置，我们模拟了 500 个相互独立的数据集，以确保能够可靠地估计经验表现。

随后，使用前文所述的所有候选方法对每个模拟数据集进行分析：(i) 未进行时间调整的获胜优势比（粗略分析和集群分层分析）；(ii) 使用线性混合效应模型进行集群层面分析；以及 (iii) 基于概率指数模型（PIM）的个体层面成对比较方法。

使用 R 中的 BuyseTest 软件包计算胜算比，使用 pim 软件包拟合概率指数模型，并使用 lme4 和 lmerTest 估计混合效应模型。对于每种方法，我们提取了估计的胜算比、其标准误以及相应的 p 值。使用 $\delta = 0$ 的情景评估 I 类错误，而非零配置则用于评估统计功效。

2.6 ETHER试验的检验效能分析

ETHER试验（*EmpoweringpaTiEnts forsHared dEcision in anticoagulant theRapy*，NCT06731244）是欧洲研究计划 MORPHEUS（HORIZON-HLTH-2022-11-01资助项目）的一部分。该试验旨在通过整合多组分风险预测评分和社会人类学量表（TDMI）的共同决策策略，改善非诱发性静脉血栓栓塞（VTE）患者的管理。TDMI 旨在临床咨询期间使用，为患者和临床医生提供个体化循证信息，说明长期抗凝治疗的获益与风险，同时考虑患者特征、临床病史和偏好。该试验评估 TDMI 是否能改善长期抗凝治疗中的决策共享、治疗安全性、效率和患者参与。

ETHER 试验采用阶梯楔形整群随机试验设计，纳入 45 个临床中心，分布于 5 个序列和 6 个时期。第一个时期持续 12 个月，随后依次经历五个为期 3 个月的时期，在此期间各集群按序从对照转为干预。

主 要终点是一个复合终点，由以下按优先顺序排列的标准组成： ia:

1. 全因死亡，2. 有症状静脉血栓栓塞症（VTE）复发，
3. 大出血或临床相关非大出血，4. 使用 PAM-13 评分测
量的患者激活程度[21]。

原始样本量依据方案中的假设确定：首先校准经典平行组随机试验中可检测到的治疗效应，然后应用适当的设计效应（缩放因子），以考虑阶梯楔形结构。由于这一流程并非专门针对基于GPC的估计量设计，本模拟研究在方案规定的固定样本量下，评估GPC方法的操作特性（检验效能和I类错误）。

模拟框架复现了 ETHER试验的阶梯楔形结构，其中 $K = 45$ 个集群在六个时间段内接受观察。对于每个集群-时期，模拟 $n = 10$ 名患者，每名患者的连续观察时间 t_{ijk} 均匀分布于相应时期内。

表 1 ETHER模拟中二分类结局所用的基线概率（p_0）和治疗效应（δ）。

标准	基线风险 p_0	治疗效应 δ（比值比）
全因死亡	0.10	–0.117 (0.89)
VTE复发	0.066	–0.674 (0.51)
大出血或临床相关非大出血	0.070	–0.816 (0.44)

对于三个二分类终点（死亡率、VTE复发和出血），按照第 2.5 节所述的 logistic模型，采用相同的模拟程序生成结局。根据方案假设设定各项指标的基线事件概率（p_0）和治疗效应（δ），并汇总于表 1。

对于连续型终点，PAM-13评分按如下方式模拟:

$$\mathrm{PAM}_{ijk} \sim \mathcal{N}(\mu_{ijk}, \sigma^2_{\mathrm{PAM}}), \qquad \mu_{ijk} = 62.6 + \delta_{\mathrm{PAM}} X_{ijk} + u_k + v_{jk},$$

其中 $\delta_{\mathrm{PAM}} = 5$ （预期均值增幅）以及 $\sigma_{\mathrm{PAM}} = 13$。[22]。随机效应$u_k$ 和v_{jk} 分别表示集群间和时期间变异。其各自方差根据预先设定的组内相关系数和整群自相关系数计算，方法与二分类模型相同，即将 $\pi^2/3$替换为 σ^2_{PAM}。模拟评分被截断至可接受范围 $[0, 100]$。在分层成对比较中，PAM绝对差值小于5.4分被视为平局，反映临床上可忽略的变异[22]。

对于每个模拟数据集，构建分层成对比较：根据上述结局的有序列表，对两个个体依次进行比较。按照广义成对比较框架，第一个非平局标准决定有利、不利或中性结果。该程序为每个集群和期间对生成有利、不利和平局比较的计数。

随后，使用第 2.4 节所述的分层混合效应模型（模型 b4）计算并建模整群–期间胜算。同时，使用聚类内概率指数模型（模型 c2）分析个体层面的比较。之所以选择这两个模型，是因为模拟研究表明，只有这两种方法在维持有竞争力的统计功效的同时，能够充分控制 I 类错误。随着纳入更多层级结局，依次应用这两个模型。

模拟情景考察了组内相关系数 $\in \{0.01, 0.03, 0.05, 0.10\}$、整群自相关系数 $\in \{0.5, 0.75, 0.9, 1.0\}$ 以及 $\beta_t \in \{-0.05, -0.025, 0, 0.025, 0.05\}$ 的各种组合。每种配置均重复模拟 500 次，以评估在 ETHER 方案规定的固定样本量下的 I 类错误率和统计功效。

3 结果

3.1 模拟研究

3.1.1 I类错误

我们首先在无治疗效应的原假设下评估了经验I类错误率（$\delta = 0$）。图2汇总了所有候选方法在组内相关系数、整群自相关系数和长期时间趋势 β_t 各种组合下的结果（$p_0 = 0.2$）。

当不存在时间趋势（$\beta_t = 0$）且集群间不存在相关性（ICC = 0）时，两种未调整方法（a1–a2）均维持了名义上的 5% 错误率。然而，整群间变异一旦增加，粗估计量（a1）的 I 类错误率便逐渐膨胀，拒绝率随 ICC 成比例上升。当时期间变异不存在或较小（CAC=1 或 0.9）时，聚类分层估计量（a2）维持了名义上的 5% 错误率，但在 CAC 值为 0.75 或 0.5 时则失效。存在任何长期趋势（$\beta_t \neq 0$）时，a1 和 a2 均无法提供有效推断，这证实了在阶梯楔形设计中忽略时间会导致结构性混杂。

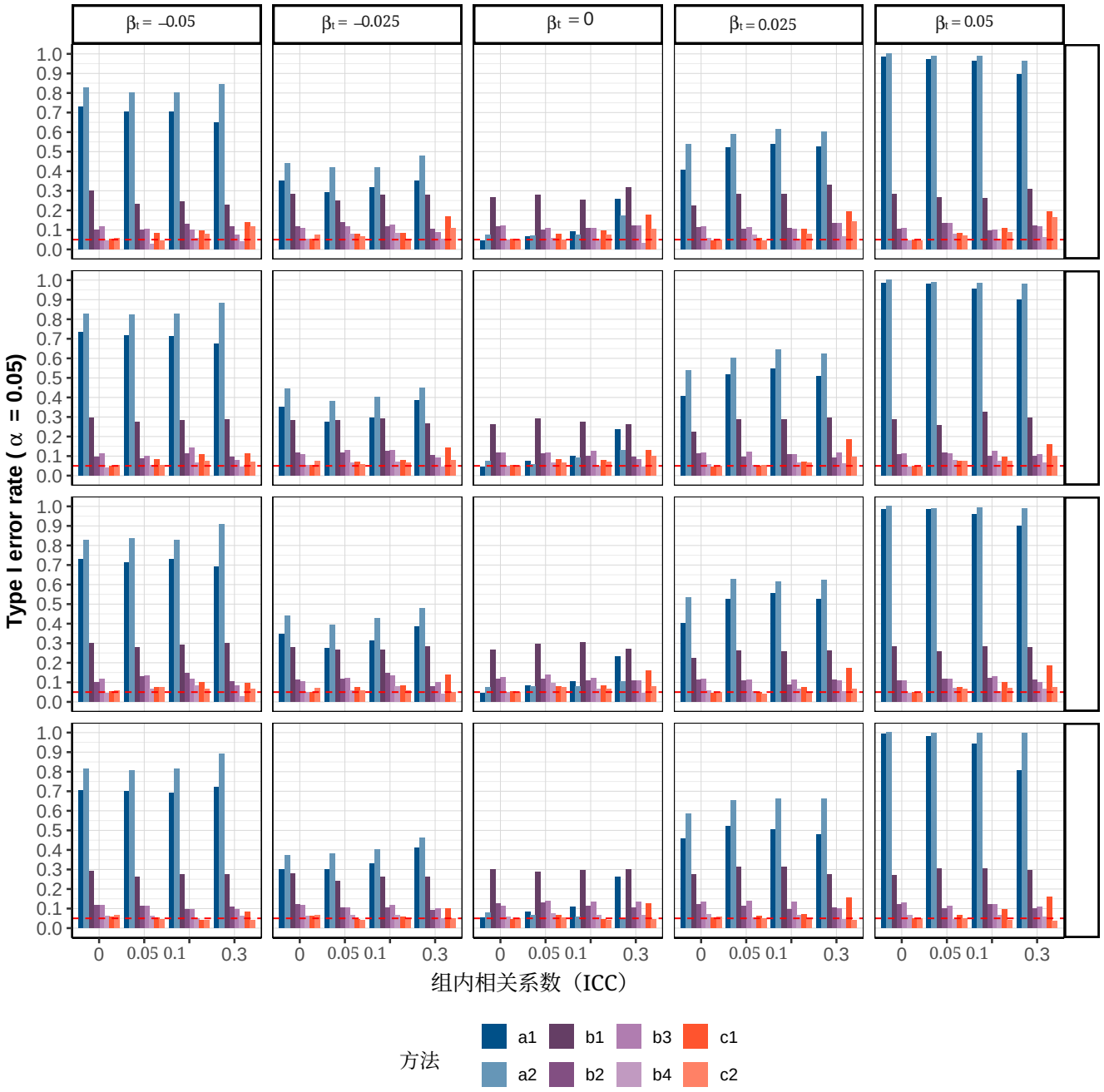

图 2 在无治疗效应的原假设（$\delta = 0$）下，当基线风险 $p_0 = 0.1$ 时，所有受评估方法的经验型I类错误率（$\alpha = 0.05$）。结果按组内相关系数（ICC）展示，并按整群自相关系数（CAC）和长期时间趋势（β_t）分层。比较了方法 a1–a2（未调整）、b1–b4（聚类层面混合模型）和 c1–c2（概率指数模型）。红色虚线表示名义5%水平。

在聚类层面混合效应模型中，最简单的模型形式（b1–b2）在几乎所有情景下都表现出显著膨胀，即使聚类程度较轻或时间趋势较弱，错误率也常超过 20–30%。纳入序列层级随机斜率（b3）减轻了这种膨胀，但未能将其消除。

只有同时纳入序列和集群层面随机斜率的完全分层设定（b4）始终控制住了I类错误，使其在所有 ICC、CAC 和 β_t 组合下都接近名义水平。

对于个体层面方法，当 ICC 值低于 0.1 时，标准概率指数模型（c1）的表现正确；当 ICC 值较高时，其 I类错误率显著增加。在时期之间变异不存在或较小的情况下（CAC=1或 0.9），整群限制概率指数模型（c2）通常在各种情景下都能保持正确的 I类错误率。对于时期之间变异较大的情况（CAC 值为 0.75 或 0.5），c2 往往出现中度膨胀，但错误率低于 10%。当集群之间和时期之间的变异都较大时（ICC=0.3 和 CAC $=0.5$），这种膨胀更加明显，I 类错误率达到 15%。

总体而言，图2清楚地表明，只有 b4 和 c2 这两种方法能够在阶梯楔形 GPC 分析中可靠地控制 I 类错误。

3.1.2 统计功效

接下来，我们评估了治疗效应非零（$\delta > 0$）时的统计功效，重点考察仅有的两种能够将 I 类错误控制在可接受范围内的方法：带有序列层级和集群层级随机斜率的分层混合效应模型（b4），以及集群限制型概率指数模型（c2）。

我们在逐渐增大的治疗效应参数（δ）取值下评估检验效能，并按 ICC、CAC 和长期时间趋势 β_t 分层。图3总结了这些参数所有组合的结果。为便于阅读，图中展示了具有代表性的 ICC–CAC 组合子集；涵盖所有相关结构和时间趋势情景的完整检验效能图集见补充材料。

在所有不存在线性趋势（$\beta_t = 0$）的情景中，对于小到中等的治疗效应，c2 始终比 b4 具有更高的检验效能。随着组内相关系数（ICC）升高以及整群自相关系数（CAC）降低，这一优势更加明显，表明当集群内依赖较强而期间间稳定性较弱时，c2 更具稳健性。对于较大的治疗效应（$\delta > 1$），两种方法的检验效能均趋于相近，说明 b4 与 c2 之间的差异主要与中等效应量有关，因为在这种情况下，将治疗效应与集群层面的异质性区分开来最具挑战性。

引入正向时间趋势（$\beta_t > 0$）后，两种方法之间的性能差距进一步扩大。时间趋势放大了 c2 的相对优势：在几乎所有 ICC 和 CAC 的组合下，c2 的检验效能都明显高于 b4。相同的定性模式仍然存在：随着 ICC 增大、CAC 降低，c2 的优势也随之增强；而对于较大的治疗效应，两种方法再次表现出相近的性能。这些结果表明，时间结构往往会强化 c2 的效率优势，尤其是在对集群层面模型而言较为困难的相关结构中。

总体而言，b4 和 c2 均展现出令人满意的统计性能。b4 在弱聚类和较强期间间自相关的情景下表现具有竞争力；在其他所有情况下，尤其是存在时间趋势和复杂相关结构时，c2 的效率始终更高。

3.2 使用候选方法分析 ETHER 试验的检验效能

在所有模拟情景中，检验效能随复合终点所纳入的标准数量，以及整群内相关（ICC）、簇自相关（CAC）和时间趋势 β_t 的变化而呈现系统性差异。图4总结了每种参数组合下原假设的拒绝率（$p < 0.05$）。

复合终点的丰富程度与由此产生的检验效能之间呈现出明确的单调关系。仅以死亡作为终点时，检验效能仍然很低（约为 5%–15%）。加入静脉血栓栓塞后，检验效能大幅提高（10%–50%）。将出血作为第三个组成部分纳入后，表现进一步提升，检验效能约为 25%–90%。最后，将四种结局合并（死亡 + 静脉血栓栓塞 + 出血 + PAM）时，所有情景下的检验效能均稳定高于 80%。正如预期，由于 PAM 这一组成部分具有很强的判别能力，它显著提高了治疗效应的可检测性，证实了更丰富的复合终点所带来的信息增益。

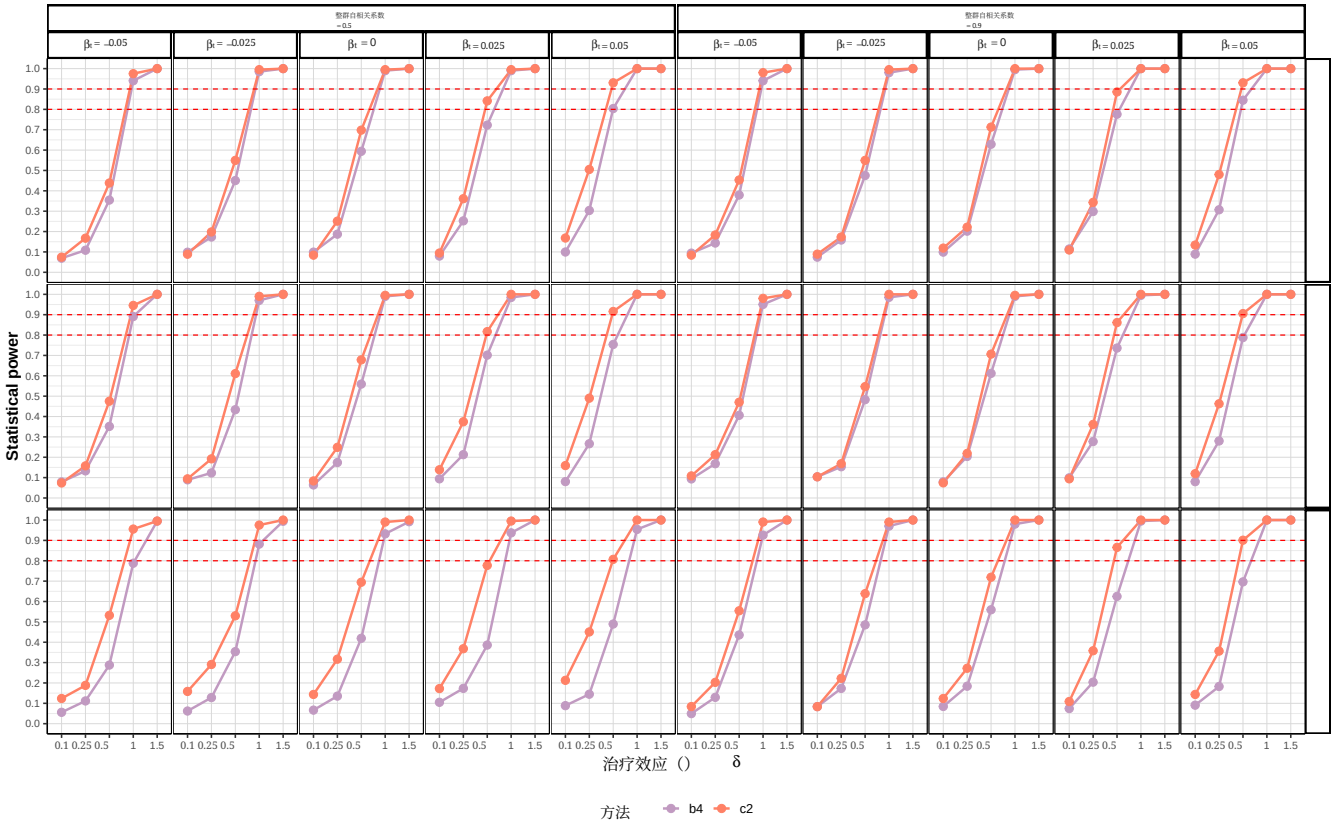

图 3 在所有非零治疗效应情景（$\delta > 0$）下，采用不同方法时的统计功效：表现最佳的两种方法分别为带有序列层级和集群层级随机斜率的分层混合效应模型（模型 b4）以及集群限制概率指数模型（模型 c2），基线风险为 $p_0 = 0.1$。对于每个治疗效应值（δ），并列显示代表这两种方法的两个柱状条。各面板按组内相关系数（ICC）、整群自相关系数（CAC）和长期时间趋势 β_t 分层。80%和90%处的水平虚线表示常用的功效基准。

在组内相关系数和整群自相关系数保持不变时，随着 β_t 从负值变为正值，检验效能稳步提高。由于正向的 β_t 与阶梯楔形设计中的治疗效应方向一致，较大的正向趋势自然有利于检测治疗效应。以 3 个终点的情况为例，当 $\beta_t = -0.05$ 时，检验效能平均约为 30%；当 $\beta_t = 0.05$ 时，则升至约 80%，这表明长期趋势对治疗效应估计有很大影响。

在固定 β_t 和 ICC 的情况下，提高 CAC 会带来适度但稳定的检验效能提升。增幅通常仅限于几个百分点，但在各终点集合和方法中均呈系统性出现，反映出集群内更强自相关结构所带来的益处。

关于方法性能，聚类内概率指数模型（c2）总体上优于混合效应比较模型（b4），但这种优势的幅度很大程度上取决于复合终点的丰富程度。当复合终点包含两个或三个成分（死亡 + VTE，或死亡 + VTE + 出血）时，两种方法之间的差异最为明显，此时 c2 达到了明显更高的检验效能。相比之下，当仅采用单一终点，或将全部四个结局合并（包括 PAM）时，两种方法的性能则更加接近。此外，随着时间趋势 β_t 越来越有利于治疗，c2 的优势逐渐减弱：在这些情形下，两种方法均受益于更强的长期趋势信号，从而缩小了它们之间的相对差距。总体而言，c2 仍保持着轻微但稳定的优势，不过其比较优势主要集中在中等复杂度的终点配置中。

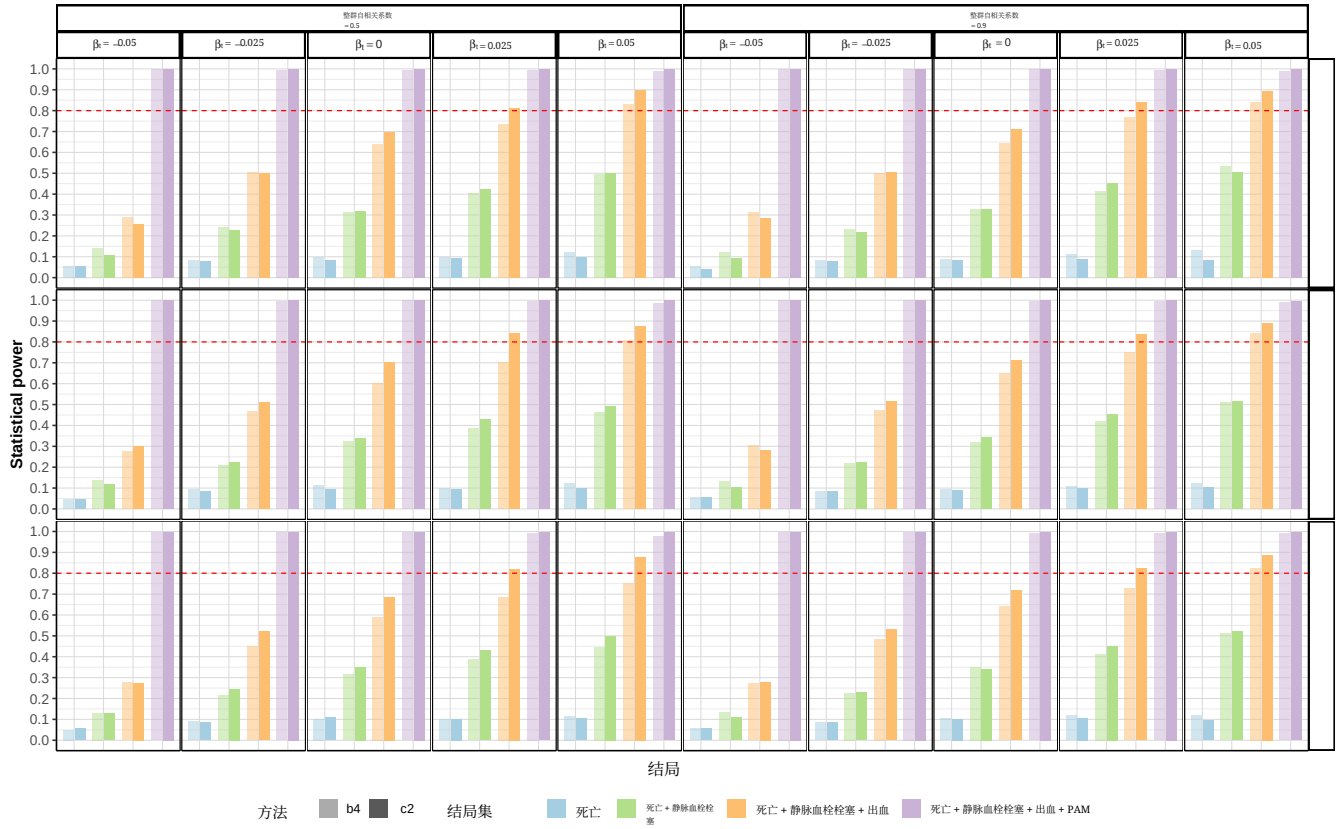

图 4 在不同模拟情景下，用于估计胜算比的各种方法的统计功效。柱形表示原假设被拒绝的重复模拟比例（$p < 0.05$）。结果按组内相关系数（行）以及与时间效应交叉的整群自相关系数 β_t（列）分层。柔和色彩用于区分连续的结局集，透明度则表示统计方法：**模型 c2** 和**模型 b4**。红色水平虚线表示 80% 的功效阈值。)

最后，利用从该模拟框架中获得的操作特性，我们评估了在方案规定的假设下 ETHER试验预期能够达到的统计功效。在 ETHER设计中，预计组内相关系数介于 0 和 0.01 之间，而整群自相关系数预计介于 0.75 和 1 之间。在这一相关结构范围内，将两种候选方法应用于完整复合终点时，均达到了非常高的检验效能。在 $(\text{ICC}, \text{CAC}) \in [0, 0.01] \times [0.75, 1]$, 的所有组合下，估计检验效能始终超过 90%，其中整群限制概率指数模型（c2）的数值接近 100%。这些结果表明，在原始 ETHER方案的假设下，并采用所提出的基于 GPC 的分析方法，该试验将拥有绰绰有余的检验效能来检测预先设定的治疗效应。特别是，纳入对治疗变化高度敏感的 PAM-13 成分，对于确保在真实试验条件下具有较高的判别能力和稳健的操作性能发挥了决定性作用。

4 讨论

本研究旨在探讨如何分析广义成对比较（GPC），使其能够恰当地反映阶梯楔形整群随机试验（SW-CRTs）固有的相关结构和时间特征。正如方法学文献所强调的，要在 SW-CRTs 中进行有效推断，统计方法必须同时调整聚类和长期时间趋势，以避免治疗效应估计出现偏倚[23]。我们的模拟研究结果表明，这种联合调整确实可行。具体而言，两种方法——基于分层混合效应形式的模型 b4，以及采用集群限制概率指数模型（PIM）的模型 c2——

被发现能够在考虑聚类、各时期的重复测量以及时间混杂的同时，持续提供有效推断。这些发现与既往研究一致；既往研究表明，概率指数模型可以通过基于回归的框架纳入协变量和结构化效应[17]。

在所评估的方法中，模型 b4 在所有模拟情景中均表现出优异的稳健性，包括簇自相关（整群自相关系数）较低的情形。其随机效应结构能够有效刻画集群间和时期间的异质性，使其成为适用于广泛阶梯楔形整群随机试验设计的特别可靠选择[24]。模型 c2 的表现同样良好，并且往往获得略高的统计功效，尤其是在组内相关（组内相关系数）较高的情景中。然而，这种效率提升是以相当大的计算负担为代价的。由于概率指数模型在个体成对比较层面运作，比较次数随簇-时期样本量呈二次增长，当纳入多个协变量时会导致难以承受的运行时间[17]。因此，在集群-时期样本量较小的设计中，概率指数模型可能仍然具有优势，因为配对总数自然受到限制。然而，对于较大的集群，混合效应模型显然更为实用。

模拟结果进一步凸显了基线事件概率 p_0 对检验效能和 I类错误的强烈影响。当 p_0 非常小（例如 0.01）时，这两个量均显著下降。这种表现是预期之中的，因为极为罕见的事件只能产生很少的有信息量的治疗–对照比较，从而降低检测真实治疗效应的能力，同时降低假阳性的概率。随着 p_0 增至更为适中的数值（如 0.1 或 0.2），有信息量的比较次数相应增加，从而使检验效能提高，并使 I 类错误率更接近其名义水平。

概率指数模型的小样本推断是另一个值得关注的问题。PIM估计依赖估计方程的大样本性质；当有效样本量有限时，方差估计量可能存在向下偏倚，从而导致 I 类错误率升高。方法学文献已记载这一现象，并提出了多种针对概率指数模型的小样本校正方法，包括偏倚校正的夹心方差估计量和基于重抽样的方法[25]。

本研究未进一步探讨这一问题，因为模拟情景是根据ETHER试验进行校准的；该试验每个时期包含相对较多的集群和个体，因此处于渐近状态。不过，在集群大小较小或有信息的比较较少的应用中，采用基于PIM的方法时，建议使用此类小样本校正。

概率指数模型的另一个重要局限是无法纳入随机效应，因为其估计依赖于无法适应层级随机效应结构的边际估计方程[17]。因此，概率指数模型无法显式建模簇自相关。尝试使用固定簇指标近似这种依赖结构未能奏效，且显著增加了计算成本。然而，在实践中，在大多数现实情景下，模型b4和c2的性能差异仍然不大，这表明两种方法均可用于阶梯楔形整群随机试验的分析。

我们的模拟主要聚焦于线性长期趋势，对应于相对简单的时间结构。在探索正弦时间模式时，结果大体保持相似，这可能是因为正弦函数围绕零振荡，因此不会随时间产生强烈的累积趋势。未来的研究应考虑更复杂的非线性时间结构，例如三次或基于样条的轨迹；若未充分校正，这些结构可能产生显著的特定时期偏差。

我们还探讨了一种扩展方法，即在PIM框架内直接使用时间的双变量样条函数对非线性长期趋势进行建模。其基本思路如下。与其通过线性对比（$t_{ijk} - t_{i'j'k}$）对时间进行调整，不如先使用维度为d的样条基$B(t) = \{B_1(t), \ldots, B_d(t)\}$（例如自然三次样条）对观察时间$t_{ijk}$和$t_{i'j'k}$进行展开。由于概率指数模型作用于个体的有序对，时间调整不能表示为时间的单变量函数，而必须表示为该对（t_{ijk},$t_{i'j'k}$）的平滑函数。这需要构建一个张量积样条基

$$\Phi(t_{ijk}, t_{i'j'k}) = \left\{ B_a(t_{ijk})\, B_b(t_{i'j'k}) \;:\; 1 \le a, b \le d \right\},$$

其中包含d^2 个基函数。每一对贡献协变量向量

$$Z_{ii'} = \left(1,\ X_{ijk} - X_{i'j'k},\ \Phi(t_{ijk}, t_{i'j'k})\right),$$

PIM估计量求解以下形式的修正估计方程

$$g\left(\mathbb{P}(Y_i \succ Y_i' \mid Z_{ii'})\right) = Z_{ii'}^{\top}\beta,$$

其中，$g(\cdot)$ 通常表示 logit连接函数。

尽管这种构造在捕捉非线性或不规则的时间动态方面具有相当大的灵活性，但其计算成本很高。样条参数的数量随样条维度（d^2）呈二次增长，所得设计矩阵也会迅速变得很大。实际上，即使是中等的样条维度（例如 $d = 3$），也会导致参数数量相当可观，尤其是在与PIM估计固有的大量成对比较相结合时。因此，基于样条的概率指数模型扩展仅适用于相对较小的集群-时期样本量，或是在仔细控制样条复杂度的情况下。

在标准的阶梯楔形分析中，时间通常被视为分类因子，这反映了研究时期持续时间相同。然而，在ETHER试验中，第一个时期为12个月，而后续时期仅持续3个月，因此存在明显的不均衡。这一特点表明，与简单的分类指标相比，将时间建模为连续变量或采用基于样条的表示更为合适。从这一角度看，概率指数模型可能提供额外的灵活性，因为连续时间差可以直接纳入成对比较层面。这一优势在生存分析场景中可能更加显著，因为事件发生的时间包含特别丰富的信息。

本研究尚未涵盖若干方面，值得进一步探讨。首先，实用性阶梯楔形试验中常见的簇大小不等，可能与集群层面的相关结构相互作用，并影响获胜优势估计的稳定性。其次，复合终点多个组成部分之间的相关性也可能影响广义成对比较的方差估计[3]，值得开展专门的方法学研究。第三，尽管我们的模拟证实，模型 b4 中采用的时期对比较表现良好，但其他方法——例如通过基于样条的调整联合建模所有时期对比——可能进一步提高存在复杂非线性时间趋势时的效率。

总体而言，本研究为阶梯楔形设计中广义成对比较的分析提供了实用指导。混合效应建模（模型 b4）总体上是一种可靠且可扩展的方法，而基于PIM的方法（模型 c2）在某些条件下可带来略高的效率，但代价是计算成本大幅增加，且无法纳入随机效应。在大多数现实情境中，两种方法得出的结论相近，因此选择时应考虑计算可行性、簇-时期样本量以及预期时间趋势的复杂程度。

资金/支持
本研究由欧盟资助，资助项目编号为 HORIZON- HLTH-2022-TOOL-11-01。本文所表达的观点和意见仅代表作者本人，并不一定反映欧盟或 HORIZON-HLTH-2022- TOOL-11-01 的观点。欧盟及资助机构均不对此承担责任。

利益冲突
作者声明不存在潜在利益冲突。

参考文献。

1 Dong G、Li D、Ballerstedt S、Vandemeulebroecke M。采用广义解析解计算胜率比，以分析复合终点并考虑各组成部分的临床重要性顺序。*Pharmaceutical Statistics.* 2016 1763 2 15 5 430 437 10 1002 ; ( ): – . doi: . /pst. . Pocock SJ、Ariti CA、Collier TJ、Wang D。胜率比：一种基于临床优先级分析临床试验复合终点的新方法。*European Heart Journal.* 2012 352 3 33 2 176 182 10 1093 ; ( ): – . doi: . /eurheartj/ehr . Buyse M。两样本问题中优先结局的广义成对比较。*Statistics in Medicine.* 2010;29(30):3245–3257. doi: 10.1002/sim.3923 4. Johns H、Campbell B、Bernhardt J、Churilov L。趋势的广义成对比较：对胜率比和胜算比的扩展，适用于剂量-反应和预后变量分析。*Statistical Methods in Medical Research.* 2023;32(3):609–625. doi: 10.1177/09622802221146306

5. Thas O、Vansteelandt S、Molenberghs G、Aerts M。概率指数模型。*Journal of the Royal Statistical Society: Series B (Statistical Methodology).*2012;74(4):623–671. doi: 10.1111/j.1467-9868.2011.01020.x 6. Song J、Verbeeck J、Huang B 等。胜算比：统计推断与回归。*Journal of Biopharmaceutical Statistics.* 2023;33(2):140– 150. doi: 10.1080/10543406.2022.2089156 7. Mao L。用于变量选择和风险预测的正则化胜率比回归及其在心血管试验中的应用。*BMC Medical Research Methodology.* 2025;25(1):1–14. doi: 10.1186/s12874-025-02554-w 8. Giraudeau B、Weijer C、Eldridge SM、Hemming K、Taljaard M。为何以及何时应采用整群随机化？*Journal of Epidemiology and Population Health.* 2024 202197 9 72 202197 10 1016 2024 ; : . doi: . /j.jeph. . . Romiti GF、Guo Y、Corica B、Proietti M、Zhang H、Lip GYH。房颤移动健康技术整合护理：来自mAFA-II随机临床试验的胜率比分析。*Thrombosis and Haemostasis.* 2023 1769612 10 123 11 1042 1048 10 1055 0043 ; ( ): – . doi: . /s- - . Kasza J、Bowden R、Hooper R、Forbes AB。分批阶梯楔形设计：一种可抵御集群招募延迟的设计。*Statistics in Medicine.* 2022 9438 11 41 18 3627 3641 10 1002 ; ( ): – . doi: . /sim. . Billot L、Copas A、Leyrat C、Forbes A、Turner EL。应如何分析整群随机试验？*Journal of Epidemiology and Population Health.* 2024 202196 12 72 202196 10 1016 2024 ; : . doi: . /j.jeph. . . Li F、Hughes JP、Hemming K、Taljaard M、Melnick ER、Heagerty PJ。用于阶梯楔形整群随机试验设计与分析的混合效应模型：概述。*Statistical Methods in Medical Research.* 2021 0962280220932962 13 30 2 612 639 10 1177 ; ( ): – . doi: . / . Hemming K、Taljaard M、McKenzie JE 等。阶梯楔形整群随机试验的报告：扩展CONSORT 2010声明，并附解释与阐述。*BMJ.* 2018 1614 14 363 1614 10 1136 ; :k . doi: . /bmj.k . Oakes D。关于包含多种事件类型的临床试验中的胜率比统计量。*Biometrika.* 2016 026 15 103 3 742 745 10 1093 2023 22 4 748 ; ( ): – . doi: . /biomet/asw . Dong G、Hoaglin DC、Huang B 等。分层胜出统计量（胜率比、胜算比和净获益）。*Pharmaceutical Statistics.* ; ( ): – 756 2293 16 10 1002. doi: . /pst. . Buyse M、Verbeeck J、Saad ED、De Backer M、Deltuvaite-Thomas V、Molenberghs G.，编。广义成对比较手册：以患者为中心的分析方法。Chapman and Hall/CRC。第1版，2025。17. Thas O、Neve JD、Clement L、Ottoy JP。概率指数模型。*Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2012;74(4):623–671. doi: 10.1111/j.1467-9868.2011.01020.x 18. Kenward MG、Roger JH。基于限制性最大似然的固定效应小样本推断。*Biometrics.* 1997;53(3):983–997. doi: 10.2307/2533558 19. Morgan KE、Turner RM、Eldridge SM、Dias S、Wiles N。为二元结局的整群随机交叉试验选择恰当的分析方法。*Statistics in Medicine.* 2017 7137 20 36 22 3514 3532 10 1002 ; ( ): – . doi: . /sim. . Hooper R、Bourke L 等。阶梯楔形及其他纵向整群随机试验的样本量计算。*Statistics in Medicine.* 2016 7028 21 35 26 4718 4728 10 1002 ; ( ): – . doi: . /sim. . Hibbard JH、Stockard J、Mahoney ER、Tusler M。患者激活度量表（PAM）的开发：患者和消费者激活的概念化与测量。*Health Services Research.* 2004;39(4p1):1005–1026. doi: 10.1111/j.1475-6773.2004.00269.x 22. Wilkinson TJ、Nair D、Smith AC 等。患者激活度量表的最小临床重要差异。*Kidney International Reports.* 2025 044 23 10 1016 2025 04 . doi: . /j.ekir. . . . Hemming K、Haines TP、Chilton PJ、Girling AJ、Lilford RJ。阶梯楔形整群随机试验：研究依据、设计、分析与报告。*BMJ.* 2015 391 24 350 391 10 1136 ; :h . doi: . /bmj.h . Hussey MA、Hughes JP。阶梯楔形整群随机试验的设计与分析。*Contemporary Clinical Trials.* 2007;28(2):182–191. doi: 10 007 25 1016 2006 05 . / j.cct. . . . Jaspers S、Verbeeck J、Thas O。小样本中的协变量调整广义成对比较。*Statistics in Medicine.* 2024;43(21):4027– 4042. Epub 2024 Jul 4doi: 10.1002/sim.10140

补充信息

补充支 ˇ´持信息可在出版商网站上的文章在线版本中找到 te.

本文引用格式：Taylor M.、Lauritzen P、Erath C 和 Mittal R. 关于简化基于 â ˇAŸ增量重映射â ˇA´Z 的输运方案。计算物理学杂志。2021;00(00):1–18。

附录