arXiv:2304.05658v1 [stat.AP] 12 Apr 2023

预测新研究中的亚组治疗效应：在一家制药公司

Bj¨ornBornkamp* Silvia Zaoli* Michela Azzarito* Ruvie Martin†Carsten Philipp M¨uller‡ Conor Moloney§ Giulia Capestro* David Ohlssen†Mark Baillie*

摘要

本文介绍了一家大型制药公司围绕亚组识别这一主题开展数据挑战的动机、经历与经验。该数据挑战旨在探索未来临床试验中亚组识别的方法。为模拟真实场景，参与者可使用 4 项 III期临床试验的数据来确定一个亚组，并预测该亚组在未来研究中的治疗效应；挑战参与者无法获取这项未来研究的数据。共有 30 支团队报名参加挑战，参与者约 100 人，主要来自生物统计学部门。本文概述了举办这项挑战的动机、挑战规则及组织安排。最后，我们介绍挑战结果、参与者反馈以及从中获得的经验，并探讨如何将这些经验转化为统计实践。

关键词：共同任务框架；数据科学；亚组分析；亚组识别；机器学习

1 引言

在许多计算学科中，数据挑战是一种核心范式；定期举办竞赛，随着时间推移，推动了底层预测建模任务取得显著进步。例如，用于图像处理的 ImageNet 挑战 [9] 、生物医学领域的DREAM挑战 [30] 、用于蛋白质结构预测的 CASP 挑战 [19] ，或用于信息检索的 TREC 挑战 [36] 。事实上，如今生活的许多方面都以某种方式受到此类数据挑战结果的影响。Donoho（2017年）[11] 将数据挑战及其底层的共同任务框架（CTF） [23] 称为过去几十年来许多领域预测建模取得成功的“秘诀”。CTF 包含三个关键要素：（1）共同任务，（2）共享数据，以及（3）标准评估。共同任务、共享数据/信息和标准化评估，使团队能够围绕问题进行讨论和学习；挑战结束后，也能促进不同团队之间的交流，而这一切都基于

*全球药物开发，诺华制药股份公司，瑞士巴塞尔
†全球药物开发，诺华制药公司，新泽西州，美国
‡数据与数字化，Beyond Conception GmbH，阿尔滕多夫，瑞士§全球药物开发，诺华制药股份公司，都柏林，爱尔兰

建立在共同基础上。该框架包含一个具有激励作用的竞争元素，同时也包含一个供参与者使用的社区元素，他们共同解决同一个问题。

另一个有趣的方面是，它能够评估不同团队之间解决方案的广度。对于药物开发项目中的探索性分析，人们通常只能看到处理数据的单个人员（或团队）提出的一种解决方案。评估多个团队之间结论变异性、进而评估可靠性，往往会很有意义。

此外，从更大的组织角度来看，数据挑战是一种培训和学习方式，*i.e.*，既能提升参与者整体群体针对特定问题的技能组合，也能让整个组织从中学习。对参与者而言，这种学习相当有效，因为它属于“做中学”，同时又是在数据挑战的“安全空间”中进行的。

受这组益处的激励，我们决定在一家制药公司内部的定量科学家群体中，围绕随机临床试验中的亚组识别开展一项数据挑战。在药物开发环境中，基于临床试验数据进行亚组识别被视为最难正确处理的问题之一，并被称为“最难的问题” [29]。众所周知，这类分析容易得出基于统计分析的结论，而这些结论后来无法在后续试验中重现 [38, 2, 33, 28, 3,20, 1, 12]。造成这一现象的原因有多方面。其中之一是样本量不足：临床试验通常旨在对总体效应作出具有统计可靠性的陈述，而不是对亚组治疗效应作出这样的陈述。另一个问题是多重性/选择性推断：亚组治疗效应具有变异性，因此，在众多亚组中进行搜索自然会识别出结果随机偏低或随机偏高的亚组。所考察的亚组数量越多，这一问题就越严重。前述两项困难都不能主要仅通过采用某一种数据分析方法来解决；这还取决于如何规划和开展分析，以及如何解读其结果。这一点在亚组识别或亚组分析中尤其重要，因为这些分析通常并非预先定义，而是在事后进行的。这会带来灵活性，并导致统计分析存在许多选择，分析人员可能根据个人偏见、理论或解释/假设作出决定 [32], ，从而可能增加结果的变异性并降低质量。矛盾的是，因此，探索性研究反而需要更强的自律性和严谨性。此外，关于如何进行亚组识别或探索治疗效应异质性，目前形成的标准很少，而许多新方法仍在不断发表和提出。这从分析人员的角度进一步增加了大量选择。

尽管亚组识别或探索治疗效应异质性面临这些挑战，但为了做出充分知情的临床决策，人们自然希望且有科学上的必要去探索亚组治疗效应，这使分析人员陷入了两难境地。更好地了解治疗效应变异，也有助于项目团队更深入地了解药物的作用机制及其对应的疾病适应症。近期大量关于这一主题的论文和指导文件也表明，探索治疗效应变异十分重要 [24, 12, 18, 25]。

本文的结构如下。第 2 节将介绍挑战任务、相关规则以及所提供的数据。第 3 节将回顾并讨论挑战的结果，同时也会纳入参与者反馈。最后，我们将总结主要经验，并探讨如何将这些经验转化为统计实践。

2 亚组识别数据挑战

2.1 挑战任务

我们没有采用由组织者设定真实结果的模拟式挑战，而是希望模拟项目团队可能面临的真实亚组识别情境，其中“真实结果”并不为人所知。因此，我们通过搜寻一个内部药物研发项目来选定挑战案例：该项目已有一些临床试验可供使用，同时，同一适应症下针对同一种药物还有另一项正在进行的临床试验，可用于评分。这样，参与者便能作出真正的“前瞻性”预测，也可以对参与者提出的解决方案进行真正的“留出”评估。在许多方面，这反映了真实情形：基于现有数据找出具有良好治疗效应的亚组，然后在后续研究中对其进行评估。此外，我们还搜寻了一个拥有多项既往试验数据的项目，以便有机会识别出治疗效应更强的亚组。

为此，在银屑病关节炎（PsA）适应症中，Cosentyx（司库奇尤单抗，AIN457）被确定为合适的案例 [7]。Cosentyx是一种人源单克隆白细胞介素（IL）-17A拮抗剂，适用于治疗银屑病关节炎，以及中度至重度斑块状银屑病、强直性脊柱炎、非放射学中轴脊柱关节炎和两种幼年特发性关节炎亚型。银屑病关节炎是一种慢性炎症性疾病，会影响外周关节和中轴关节、附着点及皮肤，并且通常与身体功能受损和生活质量下降有关。炎症性关节炎的症状包括关节僵硬、疼痛和肿胀、活动受限以及身体力量下降。银屑病关节炎中使用的主要终点是美国风湿病学会（ACR）评分 [13], 即一种复合终点，结合了压痛关节数量、肿胀关节数量、患者对银屑病关节炎疼痛的评估、患者和医生对疾病活动度总体评估，以及患者对身体功能的评估。临床试验中的主要终点通常为ACR20或ACR50，这些二元终点表示，与基线相比，压痛关节和肿胀关节计数至少改善20%或50%，并且其他5项ACR组成部分中至少有3项至少改善20%或50%。

参与者获得了针对银屑病关节炎（PsA）的 Cosentyx 四项随机、双盲、安慰剂对照临床试验的匿名数据（见表1）。参与者的任务是处理这些数据，并定义一个治疗效应有所提高的亚组；该亚组将在其数据访问权限仍受限的 INVIGORATE 2试验中进行评估。根据 INVIGORATE 2中指定的主要结局，采用第16周的 ACR50 评估所识别亚组中的治疗效应。除了亚组定义外，我们还请参与者提供其对拟议亚组治疗效应 δ_{pred} 在风险差尺度上的预测，以及对该预测标准差的估计 σ_{pred}。这两个量都很重要，因为它们是规划“下一项试验”所必需的：它们有助于评估计划中新试验的成功概率或把握度保证 [26] 。

2.2 共享数据

2.2.1 临床试验数据

所用临床试验的详细信息见表1。FUTURE 2-5研究的匿名化数据已在公司内部数据科学平台上向参与者开放。新试验INVIGORATE 2研究的数据一旦可用，仅能在受限监管数据平台上访问，参与者无法访问该平台。

所有5项研究的纳入和排除标准均相同，观察到的基线特征也非常相似。各项研究的一个主要区别在于Cosentyx的给药方式，见表1。标准

表1：数据挑战中纳入的临床研究汇总，包括各项试验中治疗方案的详细信息。缩写 NL 和 L 分别表示无负荷方案和负荷方案，s.c. 和 i.v. 分别表示皮下注射和静脉注射（详情见正文）。

研究	治疗	方案	样本量
FUTURE 2 (NCT01752634)	安慰剂（安慰剂）、75毫克、150毫克、300毫克	皮下注射，负荷方案	每组∼100
FUTURE 3 (NCT01989468)	安慰剂, 150mg, 300mg	皮下注射，负荷方案	每组 ∼135
FUTURE 4 (NCT02294227)	安慰剂，150mg（负荷方案），150mg（无负荷方案）	皮下注射、负荷方案 & 无负荷方案	每组 ∼106
FUTURE 5 (NCT02404350)	安慰剂，150mg（负荷方案），150mg（无负荷方案），300mg（负荷方案）	皮下注射，负荷方案和无负荷方案	∼990以 3:2:2:2 的比例
INVIGORATE 2 (NCT04209205)	安慰剂，6mg/kg + 3mg/kg	i.v.	每组∼190

该方案为“负荷”方案（在表1中缩写为“L”）；包括基线时及第1、2、3、4周注射，此后每四周注射一次。“NL”表示无负荷方案，即从基线开始每四周给药一次。FUTURE 2-5研究采用皮下注射（s.c.），而INVIGORATE 2研究采用静脉输注（i.v.）（基线时6mg/kg，此后每四周给予3mg/kg）。根据现有的临床经验以及对Cosentyx药代动力学的了解，预计该方案可产生与300mg负荷（L）方案相似的药代动力学特征和临床应答。INVIGORATE 2试验的试验方案中包含对静脉注射方案及其与皮下注射方案相似性的说明和论证，并已提供给参与者。

参与者获得了符合 CDISC ADaM 数据标准 [8] 的数据。数据包含来自四项研究（FUTURE 2、3、4 和 5）的 2,148 例患者。数据以 ADaM 格式分布于三个标准文件中：ADSL、ADBS 和 ADEFF。ADSL 包含 28 个基线协变量。这些协变量提供了一些基本信息，例如 AGE、性别、BMI；还包括病史信息，例如自首次银屑病关节炎确诊以来的时间、是否存在银屑病、是否存在多关节炎；以及与其他治疗和药物相关的协变量，例如皮质类固醇使用情况、TNF-α 抑制剂使用情况。ADBS 包含 71 个与实验室数值、临床测量、生活质量测量和疗效测量相关的基线协变量。ADEFF 包含 21 个疗效终点，包括 ACR50，涵盖 7 个时间点（第 1、2、3、4、8、12 和 16 周），具体取决于数据是否可用。这些终点包括 ACR 组成变量，例如 HAQ 残疾指数、CRP、医生对疾病活动度的总体评估、受试者对疾病活动度的总体评估、PsA 78 个关节压痛关节总评分，以及 PsA 76 个关节肿胀关节总评分。

向参与者提供了一个模板 Git代码仓库，其中包含访问数据集的脚本以及提交流程说明。团队应在该代码仓库中开展工作，并在提交文档中预先设定的位置填写相关信息，以此提交成果。该文档作为模板代码仓库的一部分提供。

2.2.2 更多信息来源

参与者获得了所有相关研究的试验方案、数据集规范，以及有关亚组识别的背景资料（*i.e.* 文献参考和示例代码）。我们确定了两位临床专家，以便解答参与者可能提出的关于该化合物、疾病或亚组相关性等方面的问题。参与者可以通过聊天或直接联系的方式提出问题。

2.3 标准评估

2.3.1 结果提交

Git代码仓库中的提交文件要求团队填写

(a) 亚组定义（最多 100 个字符，使用有效的 R 逻辑表达式，例如“CRP > 2.5 & AGE< 40”，或使用评分定义更复杂的亚组，例如“0.1*AGE -0.5*(MTXUSE $==$ 'Yes')) > 4.7”）；(b) 其 δ_{pred} 和 σ_{pred} 的取值；以及 (c) 用于推导该亚组的方法说明，包括临床/生物学依据（最多 300 词）。

只有当某个拟议亚组及其补集中的每组都至少包含 10 名接受治疗的受试者和 10 名对照受试者，且该亚组来自 INVIGORATE 2试验时，才会被视为有效。由于参与者无法获取 INVIGORATE 2试验的基线协变量，因此在最终提交前进行了两次测试提交：评估团队在 INVIGORATE 2试验中提出的亚组规模，并将结果反馈给团队。

2.3.2 评分规则

我们根据定量评分和定性评分对提交内容进行评分。定量评分旨在衡量所提出的亚组及治疗效果预测的质量；而定性评分则旨在奖励“良好的数据科学实践” [5]。

定量评分用于对所有团队进行排名，并为排名前三的团队颁发奖品。根据定性评分，还会从排名前五的团队中选出表现最佳的团队，并为其颁发额外奖品。

定量评分 Z 由两个评分的平均值计算得出。第一个评分评估了所提出的亚组在INVIGORATE2试验中的质量。如何对此进行最佳评估并不显而易见。直观而言，人们可能只想采用所提出的亚组在 INVIGORATE 2试验中的估计治疗效应大小。然而，这种方法存在问题，因为团队可能会刻意选择较小的亚组；由于变异性较大，这些亚组可能产生非常大的治疗效应（或非常小的治疗效应），使挑战结果容易受到偶然因素的影响。因此，我们改为使用亚组与治疗交互作用的检验统计量，来评估亚组中治疗效果修饰的证据。

为计算该评分，我们对INVIGO-RATE2患者第16周的ACR50应答拟合了以下逻辑回归模型：

$$logit(p_i) = \beta_0 + \beta_{trt} t_i + \beta_s s_i + \beta_{interaction} t_i s_i + x_i^{'} \beta \tag{1}$$

其中， p_i 为患者 i应答的概率， t_i 为指示治疗组别的二元变量（0 – 安慰剂，1 – Cosentyx）， s_i 为指示亚组归属的二元变量（1 – 亚组，0 – 非亚组）， x_i 为 INVIGORATE 2试验主要分析模型中也纳入的两个协变量（体重作为连续协变量，以及患者是否未接受过 TNF-α抑制剂治疗作为二元协变量）。

根据该模型，亚组中的治疗效应（在 logit 尺度上）为 $\beta_{trt} + \beta_{interaction}$，而互补组中的治疗效应为 β_{trt}。因此， $\beta_{interaction}$ 表示亚组中治疗效应的修饰。ˆ $\beta_{interaction} > 0$ 的值意味着亚组中的治疗效应大于互补组中的治疗效应。因此，我们将第一个评分定义为

$$S_1 = \frac{\hat{\beta}_{interaction}}{s.e.(\hat{\beta}_{interaction})}, \tag{2}$$

其中我们除以标准误，以避免结果过度受偶然因素驱动。为缓解模型拟合中完全分离相关的问题（除亚组的最低样本量要求之外），逻辑回归模型采用Firth方法进行拟合 [14]。在本分析中，主要伴发事件——治疗（及研究）中止——采用与主要研究分析相同的方式，通过复合策略进行处理，*i.e.* 中止治疗的患者被视为ACR50无应答者 [17]。

第二项评分用于评估对治疗效应（δ_{pred}, σ_{pred}）的预测，前提是已知亚组中风险差尺度上的治疗效应观测值$\hat{}\delta$。治疗效应估计值是基于上述逻辑回归模型 [16]进行标准化（将结果转换到风险差尺度）得到的。随后，该评分评估以 δ_{pred} 为中心、标准$\hat{}$偏差为 σ_{pred}的高斯分布在 δ处的对数似然：

$$S_2 = -0.5\log(2\pi) - \log(\sigma_{pred}) - 0.5\frac{(\hat{\delta} - \delta_{pred})^2}{\sigma_{pred}^2} \tag{3}$$

随后，使用团队间位置和尺度的稳健估计量对这两项评分 S_1 和 S_2 进行标准化，使其处于可比较的取值范围内。团队 j 的评分为

$$Z_{1,j} = \frac{S_{1,j} - \text{median}(S_{1,1}, ..., S_{1,k})}{1.4826\text{MAD}(S_{1,1}, ..., S_{1,k})} \tag{4}$$

$$Z_{2,j} = \frac{S_{2,j} - \text{median}(S_{2,1}, ..., S_{2,k})}{1.4826\text{MAD}(S_{2,1}, ..., S_{2,k})} \tag{5}$$

此处，$S_{1,j}$ 和$S_{2,j}$ 分别为团队 j 的评分， k 为团队总数。MAD是中位数绝对偏差的缩写。MAD采用标准缩放因子1.4826（这可确保对高斯分布数据的标准差进行一致估计）。使用稳健估计量的原因是，我们预计可能会出现极端结果，进而影响两个组成部分的加权。

用于对团队 j 进行排名的最终评分，是通过对两个评分取平均值得到的：

$$Z_j = \frac{Z_{1,j} + Z_{2,j}}{2} \tag{6}$$

用于评分的R函数 *i.e.* 用于计算 S_1 和 S_2，并提供给参与者以确保透明度，从而使他们有可能将其用于模型训练。

2.3.3 定性评估：同行评审

定性评分包括根据最佳实践对所有团队代码库进行同行评审。为规范评审流程，制定了两项标准：（i）科学与方法学——团队提交简短的方法学声明供评审，以帮助理解其理由和解读；（ii）项目文档和材料——制定这项标准是为了评估可重复性和知识转移。后一项标准包含若干具体要素，例如代码的清晰度（如是否采用良好的编码规范并添加注释、是否提供内容丰富的README文件、是否使用版本控制等），以及可重复性（例如，新用户能否依据通过项目代码库提供的文档，轻松独立地实施该方法并获得相同结果）。

第一项标准根据生物统计学和临床职能部门的 3 位高级副研究员对所提交的方法学声明（上文第 (c) 点）进行盲法评审后评定，这些副研究员均未参与数据挑战。第二项标准由组委会评定。定性评审仅针对定量评分排名前五的团队进行。

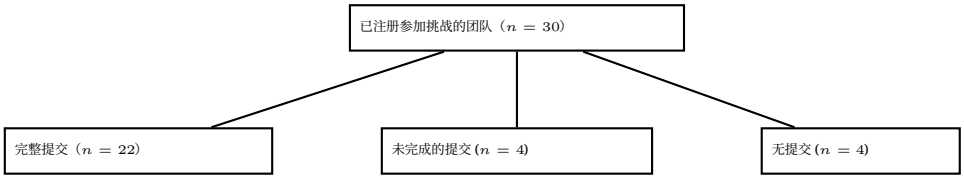


图 1：挑战参与情况总结。未修改 Git代码仓库中提交文件的团队计为“未提交”。在 4 个提交文件不完整的团队中，没有团队提供亚组理由，所有团队都提供了亚组定义，其中 3 个还提交了 δ_{pred} and σ_{pred}。我们在后续评估中排除了这些团队。

3 挑战结果

共有98名参与者以30个团队的形式报名参加挑战。在这30个团队中，有21个团队至少有一名成员从事生物统计学职能线工作，但也有来自真实世界证据、数据科学和人工智能以及业务分析等职能领域的团队。挑战于2021年12月13日以启动活动拉开帷幕，会上介绍了挑战及其规则。挑战期间还举办了一场活动，说明提交流程的详情。此外，还进行了两次测试提交，以便熟悉提交流程，并向参与者反馈他们在INVIGORATE 2试验中的亚组规模。最终提交于2022年1月24日完成。宣布获胜者后，举办了两场交流会，由获胜团队介绍其解决方案，并举行了一场包含组织者和参与者的圆桌讨论，探讨挑战的结果与经验教训。

如图 1 所示，在注册的 30 支团队中，我们收到了 26 份提交，其中 22 份符合要求，即提交文件填写完整。

表2展示了根据主要评估指标 Z 排名的各团队最终提交结果，还包括Z_1、 Z_2 以及在 INVIGORATE 2 数据上提出的亚组规模（N）。在表2中，我们简要总结了各团队采用的分析策略；组织团队根据各团队简短的方法描述进行判断，并在描述不清楚时有选择地检查其底层编程代码。即使团队将某种模型作为模型选择策略的一部分，但最终预测时并未选用该模型，我们仍将其记为已使用。“var. sel.”列评估团队是否在开始建模前进行了初步变量选择（依据数据或实质性背景知识）。“pen. regress.”列评估团队是否使用了惩罚回归（例如 LASSO）。“tree”列评估团队是否采用了单棵树方法，例如基于模型的划分（MOB） [39]。“forest/boost.”列评估团队是否使用了随机森林或提升技术。最后，“other”列涵盖团队采用的其他方法，例如标准回归技术、广义加性模型或其他非参数/非线性回归方法、关联规则学习以及图形的目视检查。“cross-val.”和“bootstrap”列评估团队是否以任何形式使用了交叉验证或自助法。

我们将更详细地介绍排名前五的团队所采用的策略

- 团队 1 通过缺失数据插补（删除缺失值超过 25% 的变量）、偏态数据转换和非正态数据标准化，对数据进行了预处理。随后，团队使用自助法重采样数据的训练集与测试集划分，对基于 LASSO 的虚拟孪生[15] 和基于模型的分区（MOB）[39] 进行比较，并根据多数表决（1000 次）选出最终模型。选定 MOB 方法后，据此确定最终亚组。

- 第2组根据与一名临床顾问的讨论和文献进行了协变量预先筛选。在此基础上，该组比较了MOB、LASSO（直接应用于亚组变量，其中连续变量的截断值基于中位数）和因果森林 [37] ，使用留一研究交叉验证，并决定根据 LASSO 方法的结果提交该亚组。

- 第3组使用了用于亚组选择的关联规则学习方法 Python 包 pysubgroup [22], ，并结合交叉验证方法。

- 第4组基于 S形 Emax 模型，采用贝叶斯剂量-反应建模来确定预测变量 [34]。随后，针对已识别的变量，采用自助法进行截断值选择，以确定一个亚组。

- 第5组选取3项研究作为训练数据，1项研究作为测试数据。首先，该组考虑了关于疾病的科学认知（合并用药、混杂因素、风险-获益）以及合并用药情况。经过一些数据整理（例如检查协变量共线性和缺失值模式）后，该组采用 SuperLearner[35] 方法，利用相关变量和若干机器学习模型，预测所有患者的治疗差异。除了 SuperLearner 内置的交叉验证外，他们还对整个估计过程进行了交叉验证。基于这些治疗差异，随后使用 Knockoff 算法选择预测变量。最后，采用广义加性模型确定连续变量的最优截断点，并在多变量模型中生成候选亚组。在候选亚组中，提交了测试数据上评分最高的亚组，并使用自助法估计标准误。

从表2或排名前五的团队所采用的策略来看，并未发现明确规律，无法判断团队工作流程中的某项特定要素是否能带来更好的结果。

这项任务为团队提供了许多选项和决策：在开始建模之前，是否应进行协变量预先筛选？如果要筛选，应如何进行？是基于初步数据分析（不使用结局变量）、 [4], 已经使用结局变量的简单单变量分析、与临床专家讨论，还是纳入科学文献？在建模阶段，可以采用不同的建模策略，但针对结局或治疗效应的模型并不能直接给出“亚组”，因此许多团队要么在这些模型的输入端额外处理（例如，对亚组指标使用 LASSO），要么对输出进行后处理（例如，基于 LASSO 使用虚拟孪生）。因此，即使团队采用了 LASSO 这类特定方法，具体做法也可能大不相同。最后，各团队以截然不同的方式实施自助法或交叉验证，以评估分析工作流程所提出亚组的质量。例如，有些团队在研究层面进行交叉验证，另一些团队则只是随机留出一部分数据，其规模与 INVIGORATE 2试验的数据量相当。

我们认为，亚组识别的性质（无法从模型输出中轻易获得；对于工作流程也尚未形成充分共识）增加了所采用分析方法的多样性。分析流程中的决策各不相同 [21], ，而不仅仅是统计模型的选择。即使团队采用了相同或相似的方法，也可能使用同一方法的不同变体，或以不同方式组合不同方法，最终识别出不同的亚组。因此，依据表2很难得出有力结论，判断某些特定“方法”是否优于其他方法。问题定义赋予团队自由度，使其能够制定超出单纯指定模型范围的分析策略。这说明，探索性问题的灵活性带来了许多选择。这项练习揭示了在制定分析策略并实施该策略时可供选择的诸多选项（其中许多往往并不明显）。

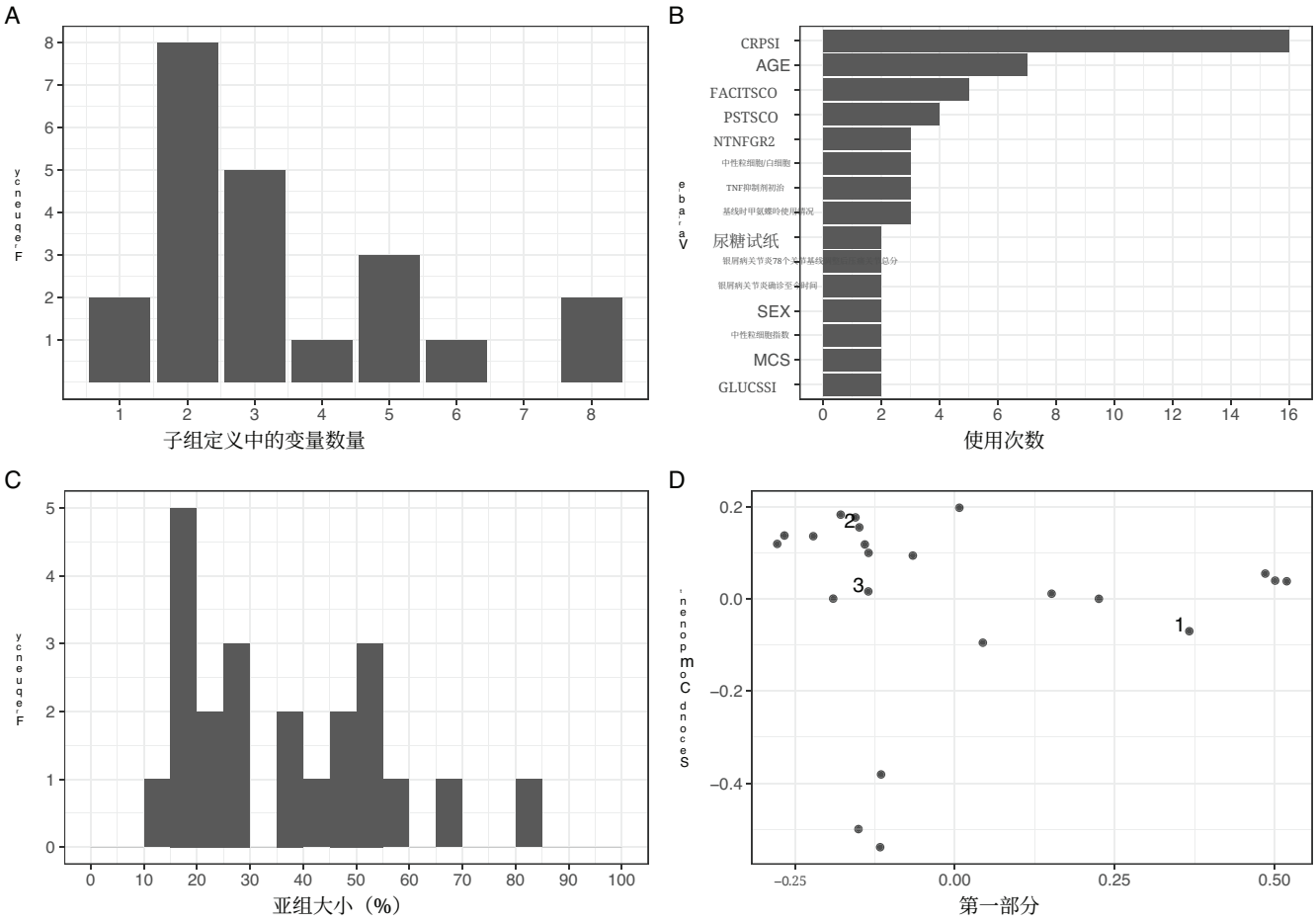


图2：参与者提交的亚组特征。A)亚组定义中使用的变量数量；B)每个变量在亚组定义中出现的频率（缩略词所代表变量见表3）。图中仅纳入至少使用两次的变量；C) 亚组规模直方图，以属于该亚组的全部患者百分比进行测量；D)根据亚组之间的成对Jaccard 距离计算得到的前两个主成分，排名前3的团队以相应编号标记。

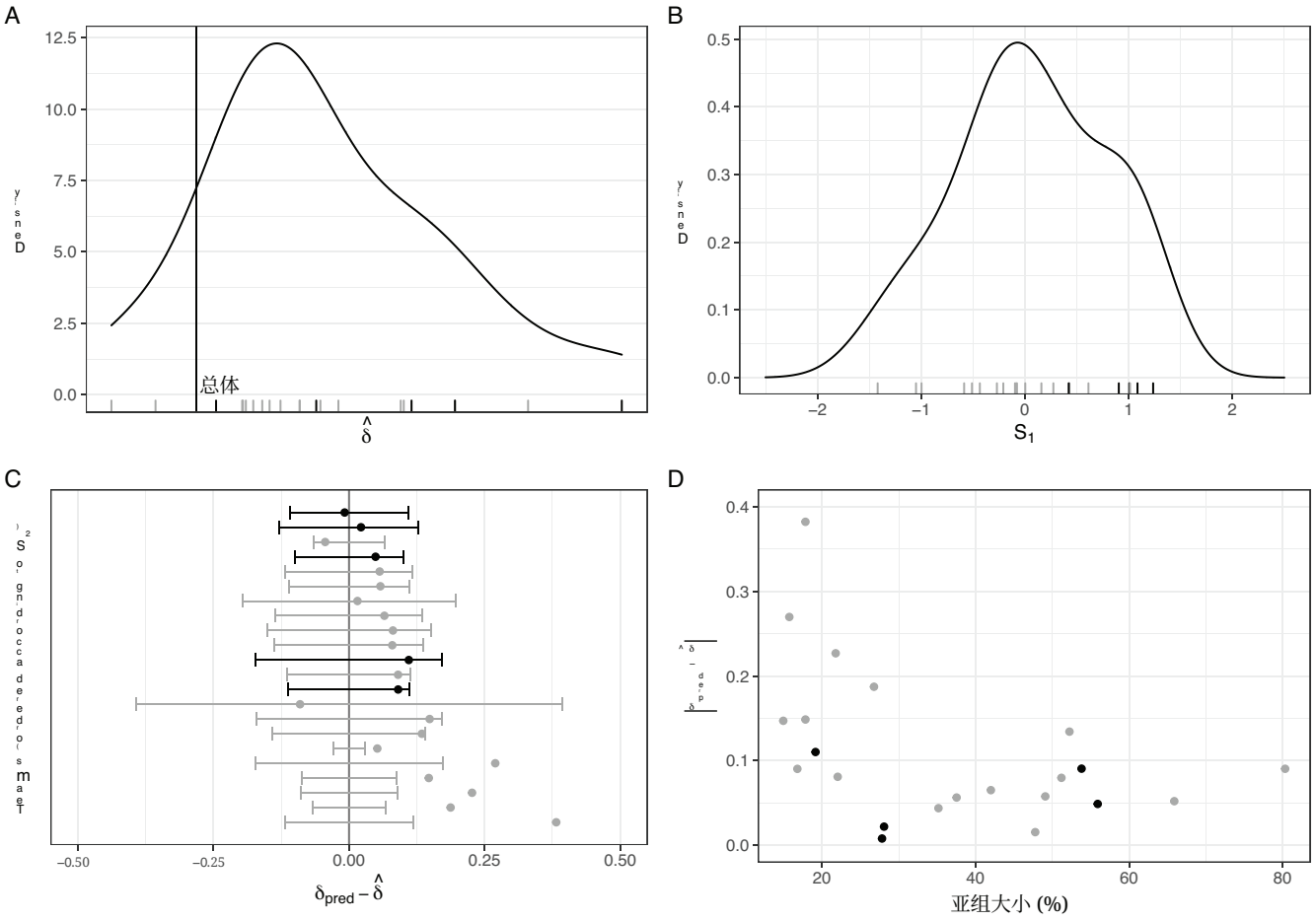
3.1 提交的亚组比较

不同团队提交的亚组也存在显著差异（见图2），包括亚组定义所用变量的种类和数量（图2A），以及亚组规模（图2C）；后者占试验患者的比例从10%到85%不等。

最常用的变量是反映炎症程度的C反应蛋白（CRP）、随机分组时患者的年龄（以年为单位）（AGE），以及衡量患者疲劳程度的指标（FACIT评分）（见图2B）。这些变量在团队汇总层面的出现情况与此前对FUTURE 2-5数据的分析一致[31], ，该分析中对这些结果的临床解读可供参考。

我们还通过计算各亚组之间的成对Jaccard距离，考察了所提交方案的相似性。Jaccard距离的定义为：1减去两个待比较亚组交集的基数除以这两个亚组并集的基数。我们进行了多维尺度分析，并展示了前两个主坐标（图2D）。

有趣的是，与所提出的亚组整体多样性相比，排名第一的团队（ Z 最高）提出的亚组与排名第二和第三的团队提出的亚组差异相当大。



ˆ 图 3：A) INVIGORATE 2 中所提出亚组内观察到的治疗效应 δ 的分布。垂直线标示试验中的总体治疗效应。x 轴上的刻度标示各个团队的结果，黑色刻度代表排名前五的团队；B) 评分 S_1 的分布。x 轴上的刻度标示各个团队的结果，黑色刻度代表排名前五的团队；C) 各团队对其 ˆ 亚组预测的治疗效应 δ_{pred} 与 INVIGORATE 2 中该亚组观察到的治疗效应 δ 之间的差异。图中还显示了基于 σ_{pred} 的 95% CI。团队按其 S_2 评分排序，排名前五的团队以黑色显示；D) 预测误差的绝对值与亚组大小的关系图，每个点代表一个团队，排名前五的团队以黑色显示。

3.2 提交亚组的治疗效应

在 INVIGORATE 2试验中，除了两个亚组外，所提交亚组在风险差尺度上的治疗效应均高于总体治疗效应（图 3A）。有趣的是，第一个评分 S_1 的分布仅略微偏向正值；如果亚组是随机选择的，该分布应以零为中心（图 3B）。其原因是 S_1 衡量的是 logit 尺度上的交互作用，而在本例中，这两种尺度（风险差与 logit 尺度）得出的结果并不完全一致。我们将此归因于 logit 尺度的非线性。在治疗下，应答概率的增加仅会使 S_1 小幅、次线性地增加，而安慰剂下应答概率的降低则会获得高得多的奖励。众所周知，交互作用的概念取决于效应测量的尺度，因此这一结果并不令人意外。事实上，既有经验观察，也有数学结果表明，与风险差尺度相比，logit 尺度能带来更均一的效应度量 [27, 10], 这也可能】【。

在此解释结果。有关风险差尺度与 logit 尺度之间差异的更多详情，以及如果我们使用风险差尺度上的评分，排名会如何变化的敏感性分析，请参阅附录 A。

参与者根据 FUTURE 2-5研究为其亚组提供的治疗效应预测值， δ_{pred}，几乎总是高于 INVIGORATE 2试验中同一亚组观察到的实际治疗效应，$\hat{}\delta$ (图 3C)。团队的平均高估幅度为 34%。我们将这一观察结果归因于“回归均值”效应。根据提交的标准差 σ_{pred} 得出的 95% 预测区间大多涵盖了观察值，但涵盖观察值的团队数少于预期（图 3C）。

对于较小的亚组，预测误差的绝对值更大（图 3D）。这一趋势可能是由于小样本中的变异性较大，导致“回归均值”效应更明显。从图 3D 还可以看到，整体排名前五的团队提出的亚组规模差异很大（黑点）：其中一个亚组仅占 20%，而其他团队的亚组约占 58%。

Z_1 和 Z_2 评分在低值区间呈一定正相关，即某团队在一项评分中的表现不佳，往往在另一项评分中的表现也不佳；但在高值区间，这种相关性便消失了（见附录中的图 4A）。这两项评分与最终评分 Z 的相关程度相近（Spearman 相关系数分别为 0.87 和 0.74），因此最终团队排名受到两者的影响大致相同。事实上，从图 3B 和 C 可以看出，按 Z 排名的排名前五的团队（以黑色突出显示），在 S_1 和 S_2方面的评分也很高。

为评估排名对评分定义的稳健性，我们测试了另一种团队排名方法：先分别根据 Z_1 和 Z_2计算两组排名，再对名次取平均值，得出最终评分。这种排名方法得出的前五名与原方法相近。

3.3 反馈调查：参与者的观点

结果公布后，研究团队向参与者发送了一份反馈问卷，共有65名参与者填写。问卷包含17个问题，包括用于评价满意度和组织情况的选择题，以及允许自由作答的问题，以收集更广泛的反馈，以及参与者在亚组识别方面的学习收获和建议。

总体而言，参与者对这项挑战的评价很高。在“你如何评价参与亚组挑战的整体体验？”这一问题中，评分范围为1（差）至5（很好），82%的应答给出了4分或5分，其中42%给出了5分。我们通过一道自由作答题询问大家喜欢这项挑战的哪些方面。常见的反馈包括：他们很喜欢了解一个新主题、亲身体验相关方法，并且很欣赏所提供的入门材料。参与者也对这项挑战贴近现实的设置给予了积极评价：他们有机会在真实的临床数据上测试方法，并解决一个没有已知答案的开放式问题，这与实际工作相似。作为挑战的一部分，参与者使用了公司内部新推出的数据科学平台，对许多人来说，这是他们首次有机会尝试该平台。他们很喜欢学习如何使用这个平台并探索其功能。一些参与者提到，他们很喜欢这项竞赛面向全球，涵盖公司所有主要办公地点，也借此机会结识了日常交际圈之外的同事。他们喜欢在竞争友好、氛围融洽的环境中，尝试与日常工作不同的新鲜事物。

我们还询问了大家学到了什么。个别参与者提到，他们很喜欢学习亚组识别方法（例如阅读论文、探索 R 包）。此外，他们也很珍惜这个

提升编程技能、发现新软件包并进一步熟悉有助于协作工作的良好数据科学实践的机会，例如使用 Git 进行版本控制。谈到亚组识别方面的收获时，参与者强调，领域知识对这项任务很重要，而且预测差异化治疗效应颇具挑战。最后，参与者也很喜欢进一步了解银屑病关节炎这一疾病领域。

4 讨论与结论

我们围绕亚组识别这一主题开展了一项公司内部数据挑战。参与者利用现有临床试验数据，为即将开展的研究提出了一个亚组。

总体而言，几乎所有团队都成功找到了一个亚组，在这项新研究的风险差尺度上，该亚组的治疗效应估计值高于总体治疗效应。然而，大多数团队高估了其亚组在这项新研究中的治疗效应，我们将其归因于均值回归效应。

这项数据挑战带来了若干经验。最重要的第一条经验（反馈调查的应答也反映了这一点）是，识别治疗效应增加或降低的亚组“很困难”。即使在本次挑战的理想情形下——有四项临床试验可用于学习亚组定义——要找到能够在一项新的单独试验中重现的亚组结果，仍然很困难。几乎所有团队都能找到治疗效应大于观察到的总体治疗效应的亚组，但对大多数团队而言，其亚组的治疗效果预测都偏于乐观。预测质量取决于亚组规模（小亚组更难预测）。这一观察与《药物统计学》中的“经验之谈”一致，我们认为这是重要的实践经验。为此，我们在一个用于探索和测试的“安全空间”中开展了这项挑战，将其置于项目相关工作之外，同时尽可能贴近现实。此外，团队结果以半盲方式展示：只有排名前五的团队名称向公众公布，因此参与者不必担心在挑战中表现不佳可能造成的负面影响（否则可能导致他们不参与或不提交结果）。

关于亚组识别，并没有得到具体的经验，即某些分析方法在效果上优于其他方法。原因在于，亚组识别需要经过多个步骤才能得出一个亚组，因为大多数用于建模的分析方法并不会直接报告亚组。此外，大多数团队都将分析方法与偏倚校正或交叉验证技术等元方法结合使用，因此很难进行清晰的比较。所提供解决方案的广泛性表明，公司可能需要针对这一主题制定内部指导，以避免团队“重复造轮子”。在随后讨论挑战结果的会议中，参与者还提出了这样一个问题：重点是否真的应该放在治疗效应增加或减少的“亚组”上，而不是关注可能调节治疗效应的协变量。定义具体亚组这一额外步骤可能并非必要。此外，仅报告选定的亚组结果，可能无法很好地概括试验中观察到的治疗效应异质性。

关于挑战赛形式，我们的预期得到了证实（至少从参与者的反馈来看）：这类挑战有助于个人学习亚组识别、R、R Markdown、Git 等软件工具、新的科学计算平台，以及疾病适应症。参与者还提到，可以在团队中向其他经验更丰富的成员学习。活动结束后，非正式的学习以及正式的学习（作为结果讨论的一部分）也在各团队之间持续进行，排名前五的团队分享了他们的经验和学习成果：CTF，即共同任务，

数据和规则为参与者讨论问题和分析策略提供了一种“通用语”[6]。

此次挑战开展过程中有哪些不足？我们决定采用一个真实案例，并使用新研究中真实且未知的结果。这种做法的缺点是，我们无法确定“胜出”的亚组是否确实是最优的。新研究不过是“又一项研究”；在一项研究中考察亚组并据此评分时，亚组结果会存在随机性。在启动会上，我们也向参与者说明了这一局限。

在此次挑战中，所有参与者都必须提交一个亚组。挑战开始前，我们曾考虑过其他规则：团队无需提交亚组，但可以选择提交整个试验人群，并预测总体治疗效应。为了简化挑战规则，我们最终没有采用这一方案；不过事后看来，这仍是一个值得重新考虑的选项。

与真实案例不同的一种做法是开展模拟挑战，由组织团队掌控真实情况。在这样的挑战中，参与者可以提交算法，并根据模拟中的真实情况评估算法在重复抽样下的表现。在这种设置下，团队必须编写一套明确的亚组识别策略（这也会使解决方案更容易评估）。这种方法的缺点是，参与者实际上要做的就是猜测组织者认为合理的模拟真实情况。即便采用多种情景，在我们看来，这仍然是一个重要的不足。

在本次挑战中，我们使用了一项真实的“留出”研究，其主要结果尚未发表，且受到访问限制。挑战结束时，有几位参与者表示有兴趣自行探索留出研究的数据。遗憾的是，这并不可行，因为监管数据平台的访问限制仍然有效。此外，也没有可供分享给参与者的匿名化数据集。尽管这是本次挑战设置的一个缺点，但我们认为，使用一项结果未知的真实未来研究进行评分（从而更好地模拟现实）所带来的益处，足以弥补这一不足。

最后，我们为参与者提供了入门材料和代码。反馈表明，参与者很认可这些材料和代码，它们既帮助大家从方法论层面着手了解这一主题，也让大家可以直接使用代码。不过，这其中也存在需要权衡之处：许多团队将所提供的方法作为开展自身研究的起点，因此影响了参与者的解决方案。例如，令我们意外的是，在这项挑战中，使用传统标准统计方法（如基于回归的方法）的团队非常少。

ACKNOWLEDGEMENTS

我们感谢挑战的所有参与者，并特别致敬排名前五的团队成员：团队 1（团队“GO SUBGROUP GO!”）Mingmei Tian、Tao Li、Yongmin Liu、Pengpeng Wang；第2组（团队“Better Than Placebo”）Beilei He、Markus Lange、Jelena ˇCuklina、Abdelkader Seroutou；团队 3（团队“AI4Lifers”）Joana Marques Barros、Matthias Hueser、Hagen Muenkler；团队 4（团队“Rgonauts”）Isidoros Papaioannou、Marius Thomas、Zheng Li，以及团队 5（团队“The Counterfactual Knockoffs”）Sandra Lopez-Leon、Rima Izem、Matthias Kormaksson。我们还要感谢两位随时提供咨询的临床医生：Aimee Readie 和 Sarah Whelan。我们感谢整个 Cosentyx 团队对本次挑战的支持，尤其是 Xuan Zhu 和 Chengeng Tian。此外，我们还要感谢：Brian Porter、Janice Branson 和 Eric Gibson，他们参与了方法学声明的定性评审；Kostas Sechidis 和 Frank Bretz，他们协助测试并审阅了提供给参与者的材料（Git 模板仓库、挑战规则手册）；Jana St´arkov´a，她在挑战接近尾声时作为组织团队的一员提供了支持；最后还要感谢 Bj¨orn Holzhauer，感谢其围绕设计和

对挑战进行评分，并提出使用“真正的留出集”（即仍限制访问）研究进行评分的想法。

References

[1] M. Alosh、M. F. Huque、F. Bretz 和 R. B. D’Agostino Sr. 确证性临床试验亚组分析中的统计学考量教程。 *Statistics in medicine*，36(8):1334–1360，2017年。[2] S. F. Assmann、S. J. Pocock、L. E. Enos 和 L. E. Kasten。临床试验中的亚组分析及基线数据的其他（误）用法。 *The Lancet*，355(9209):1064–1069，2000年。[3] P. C. Austin、M. M. Mamdani、D. N. Juurlink 和 J. E. Hux。检验多个统计假设会产生虚假关联：一项关于星座与健康的研究。 *Journal of clinical epidemiology*，59(9):964–969，2006年。[4] M. Baillie、S. le Cessie、C. O. Schmidt、L. Lusa、M. Huebner，以及 STRATOS 倡议“初始数据分析”专题组。初始数据分析的十条简单规则。 *PLOS Computational Biology*，18(2):1–7，2022年2月。[5] M. Baillie、C. Moloney、C. P. Mueller、J. Dorn、J. Branson 和 D. Ohlssen。良好的数据科学实践：迈向药物开发实践规范。 *Statistics in Biopharmaceutical Research*，0(0):1–12，2022年。[6] M. Baker。作者档案：Gustavo Stolovitzky。 *Nat. Methods*，9(8):767，2012年8月。[7] H. A. Blair。司库奇尤单抗：银屑病关节炎治疗综述。 *Drugs*，81(4):483–494，2021年。[8] CDISC。ADaM 数据集和元数据标准，2021年。 https://www.cdisc.org/standards/foundational/adam。[9] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei。ImageNet：大型层次化图像数据库。载于 *2009 IEEE conference on computer vision and pattern recognition*，第248–255页。IEEE，2009年。[10] P. Ding 和 T. J. VanderWeele。不同效应尺度下齐性空间的微分几何。 *arXivpreprint arXiv:1510.08534*，2015年。[11] D. Donoho。数据科学的50年。 *Journal of Computational and Graphical Statistics*，26(4):745–766，2017年。[12] 欧洲药品管理局。确证性临床试验中亚组研究指南。*EMA/CHMP/539146*，2019年。[13] D. T. Felson、J. J. Anderson、M. Boers、C. Bombardier、M. Chernoff、B. Fried、D. Furst、C. Goldsmith、S. Kieszak、R. Lightfoot 等。美国风湿病学会为类风湿关节炎临床试验制定的疾病活动度测量指标初步核心指标集。 *Arthritis & Rheumatism: Official Journal of the AmericanCollege of Rheumatology*，36(6):729–740，1993年。[14] D. Firth。最大似然估计的偏倚缩减。 *Biometrika*，80(1):27–38，1993年。[15] J. C. Foster、J. M. Taylor 和 S. J. Ruberg。基于随机临床试验数据的亚组识别。*Statistics in medicine*，30(24):2867–2880，2011年。

[16] M. Ge、L. K. Durham、R. D. Meyer、W. Xie 和 N. Thomas。利用逻辑回归和加权风险差计算临床试验中经协变量调整后的比例差。 *Drug information journal: DIJ/DrugInformation Association*，45(4):481–493，2011年。[17] ICH协调指南。关于临床试验估计目标和敏感性分析的附录：临床试验统计学原则指南 E9(R1)。 *https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf*，2019年。[18] D. M. Kent、D. Van Klaveren、J. K. Paulus、R. D'Agostino、S. Goodman、R. Hayward、J. P. Ioannidis、B. Patrick-Lake、S. Morton、M. Pencina 等。治疗效应异质性的预测方法（PATH）声明：说明与详述。 *Annals of Internal Medicine*，172(1):W1–W25，2020年。PMID: 31711094。[19] A. Kryshtafovych、T. Schwede、M. Topf、K. Fidelis 和 J. Moult。蛋白质结构预测方法的关键评估（CASP）——第十三轮。 *Proteins: Structure, Function, and Bioinformatics*，87(12):1011–1020，2019年。[20] S. W. Lagakos 等。亚组分析的挑战——报告结果而不造成歪曲。 *New England Journalof Medicine*，354(16):1667，2006年。[21] J. T. Leek 和 R. D. Peng。统计学：P值只是冰山一角。 *Nature*，520(7549):612，2015年4月。[22] F. Lemmerich 和 M. Becker。pysubgroup：在 Python 中轻松进行亚组发现。载于 *Joint European conference on machine learning and knowledge discovery in databases*，第658–662页。施普林格，2018年。[23] M. Liberman。Fred Jelinek。 *Computational Linguistics*，36(4):595–599，2010年。[24] I. Lipkovich、A. Dmitrienko 和 R. B. D'Agostino。生物统计学教程：临床试验中的数据驱动亚组识别与分析。 *Statistics in Medicine*，36(1):136–196，2017年。[25] X. Nie 和 S. Wager。异质性治疗效应的准预言机估计。 *Biometrika*，108(2):299–319，2021年。[26] A. O'Hagan、J. W. Stevens 和 M. J. Campbell。临床试验设计中的把握度保证。 *Pharmaceutical Statistics:The Journal of Applied Statistics in the Pharmaceutical Industry*，4(3):187–201，2005年。[27] C. Poole、I. Shrier 和 T. J. VanderWeele。风险差真的是异质性更强的测量指标吗? *Epidemiology*，26(5):714–718，2015年。[28] P. M. Rothwell。随机对照试验中的亚组分析：重要性、适用情形与解读。 *The Lancet*，365(9454):176–186，2005年。[29] S. J. Ruberg。评估并传达患者亚群治疗效应的异质性：最棘手的问题。 *Pharmaceutical Statistics*，20(5):939–944，2021年。[30] J. Saez-Rodriguez、J. C. Costello、S. H. Friend、M. R. Kellen、L. Mangravite、P. Meyer、T. Norman 和 G. Stolovitzky。众包生物医学研究：借助社群推动创新。 *NatureReviews Genetics*，17(8):470–486，2016年。

[31] K. Sechidis、M. Kormaksson 和 D. Ohlssen。利用敲除变量识别受控预测性生物标志物。 *Statistics in Medicine*，40(25):5453–5473，2021年。[32] R. Silberzahn、E. L. Uhlmann、D. P. Martin、P. Anselmi、F. Aust、E. Awtrey、S. Bahnik、F. Bai、C. Bannard、E. Bonnier、R. Carlsson、F. Cheung、G. Christensen、R. Clay、M. A. Craig、A. D. Rosa、L. Dam、M. H. Evans、I. F. Cervantes、N. Fong、M. Gamez-Djokic、A. Glenz、S. Gordon-McKeon、T. J. Heaton、K. Hederos、M. Heene、A. J. H. Mohr、F. H¨ ogden、K. Hui、M. Johannesson、J. Kalodimos、E. Kaszubowski、D. M. Kennedy、R. Lei、T. A. Lindsay、S. Liverani、C. R. Madan、D. Molden、E. Molleman、R. D. Morey、L. B. Mulder、B. R. Nijstad、N. G. Pope、B. Pope、J. M. Prenoveau、F. Rink、E. Robusto、H. Roderique、A. Sandberg、E. Schl¨uter、F. D. Sch¨onbrodt、M. F. Sherman、S. A. Sommer、K. Sotak、S. Spain、C. Sp¨orlein、T. Stafford、L. Stefa nutti、S. Tauber、J. Ullrich、M. Vianello、E.-J. Wagenmakers、M. Witkowiak、S. Yoon 和 B. A. Nosek。众多分析者，同一数据集：以透明方式呈现分析选择的差异如何影响结果。 *Advances in Methods and Practices in Psychological Science*，1(3):337–356，2018年。[33] P. Sleight。辩论：临床试验中的亚组分析：看起来很有趣——但别相信它们！*Trials*，1(1):1–3，2000年。[34] M. Thomas、B. Bornkamp 和 K. Ickstadt。利用贝叶斯分层模型识别剂量探索试验中的治疗效应异质性。 *Pharmaceutical Statistics*，21(1):17–37，2022年。[35] M. J. Van der Laan、E. C. Polley 和 A. E. Hubbard。超级学习器。 *Statistical applications in genetics andmolecular biology*，6(1)，2007年。[36] E. M. Voorhees 和 D. K. Harman。 *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*。麻省理工学院出版社，2005年。[37] S. Wager 和 S. Athey。利用随机森林估计并推断异质性治疗效应。*Journal of the American Statistical Association*，113(523):1228–1242，2018年。[38] S. Yusuf、J. Wittes、J. Probstfield 和 H. A. Tyroler。随机临床试验中患者亚组治疗效应的分析与解读。 *JAMA*，266(1):93–98，1991年7月。[39] A. Zeileis、T. Hothorn 和 K. Hornik。基于模型的递归划分。 *Journal of Computational andGraphical Statistics*，17(2):492–514，2008年。

在风险差和 logit 尺度上优化治疗效应的差异

评分 S_1 衡量logit尺度上的治疗效应。在线性尺度上，治疗下应答的增加与安慰剂下应答的同等幅度下降对治疗效应的影响相同；而在logit 尺度上，治疗效应对安慰剂下的应答比对治疗下的应答更敏感。

事实上， S_1 与安慰剂下的应答概率呈负相关（见图4B），而与治疗下的应答概率的依赖性较弱（见图4C）。这可以用 logit函数的非线性来解释：该函数在较小值时增长更快。这意味着，与提高治疗下的应答相比，降低亚组在安慰剂下的应答，是获得较高 S_1 评分的更有效策略。

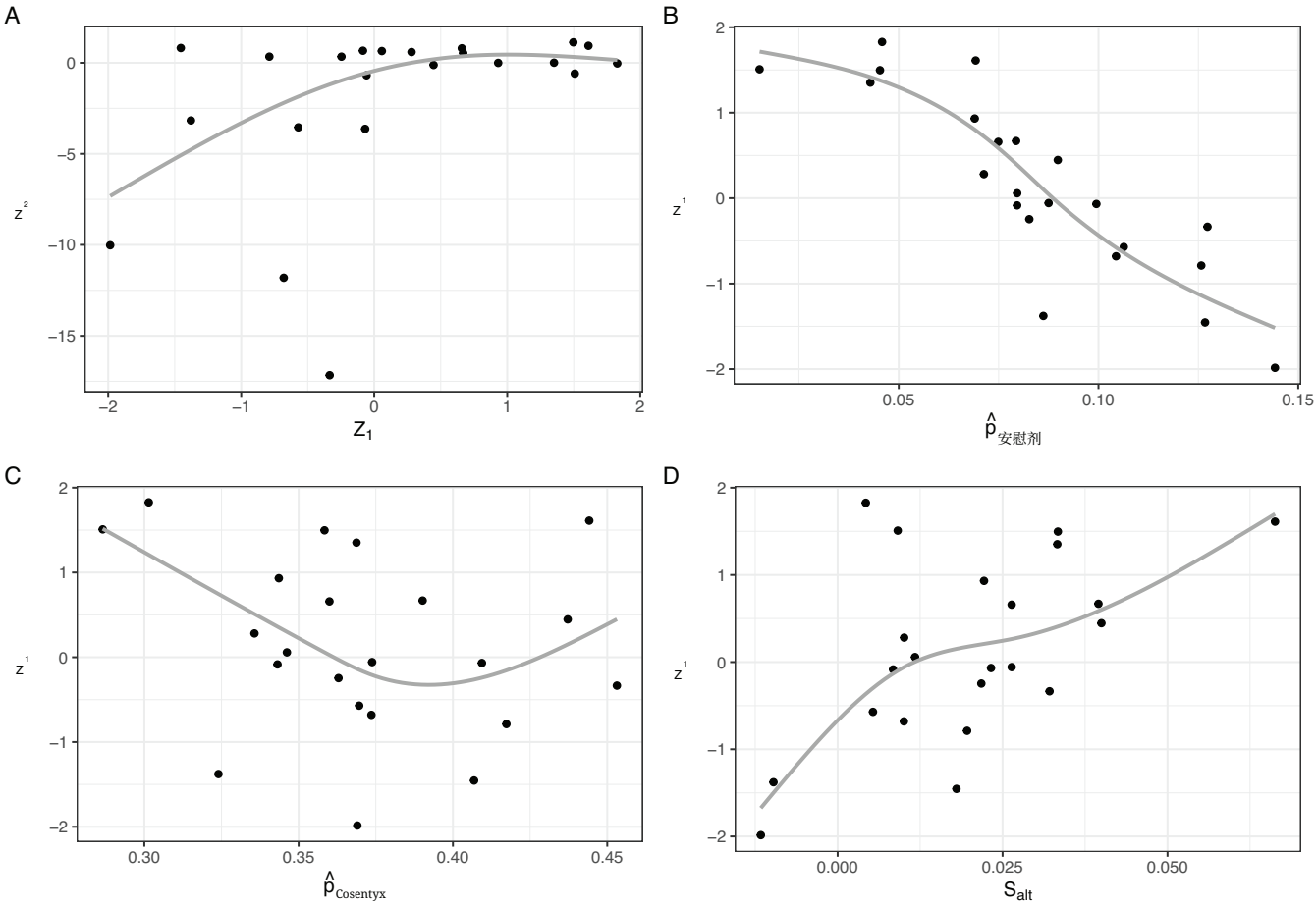


图 4：A）两个评分 Z_1 和 Z_2的散点图；B）亚组中评分 Z_1 与安慰剂下的应答概率的关系图；C）亚组中评分 Z_1 与治疗下的应答概率的关系图；D）风险差尺度上的评分 Z_1 与替代评分 S_{alt} 的关系图，适用于 $\alpha = 0.5$。所有面板中，每个点代表一个团队，灰色平滑线为自由度为 3 的三次回归样条。

为了检验排名对治疗效应采用风险差尺度或 logit 尺度进行度量这一选择的稳健性，我们进行了敏感性分析，将原始排名与采用风险差尺度上的替代定义计算的 S_1 评分所得排名进行比较。我们定义 $S_{alt} = \hat{} \hat{} f_{sub}^{\alpha}(\delta_{sub} - \delta_{overall})$，其中 f_{sub} 是亚组中的患者人数除以总体样本量，$\hat{} \hat{} \delta_{sub}$ 和 $\delta_{overall}$ 分别是亚组中和所有患者中观察到的治疗效应，α是一个参数。引入 f_{sub}^{α} 这一项是为了避免因偶然而获得较高治疗效应的极小亚组得到奖励。当 $\alpha = 0$取 0.5 和 1 时，排名略有变化，但总体上相似。图 4D 也显示了这一点：原始 S_1 评分与替代评分（当 $\alpha = 0.5$时）总体相关，但存在一些离群值。

表2：按评估标准排列的最终挑战排名。该表还展示了团队使用的统计方法（主要根据其简短的方法描述整理）以及亚组规模（N）。列名分别代表变量选择（变量选择）、惩罚回归（惩罚回归）、基于树的方法（树）、随机森林/提升（随机森林/提升）和交叉验证（交叉验证）。

Team	变量选择	惩罚回归。	tree	森林/提升	其他	交叉验证。	自助法	N	Z_1	Z_2	Z
1	x	x	x			x		106	1.50	1.13	1.31
2	x	x	x	x		x	x	107	1.61	0.94	1.27
3					x			205	1.83	-0.03	0.90
4		x					x	213	0.66	0.80	0.73
5		x		x	x	x	x	73	1.35	0.01	0.68
6	x	x	x			x		160	0.67	0.53	0.60
7		x	x	x		x		306	0.93	-0.01	0.46
8	x	x	x			x	x	68	1.51	-0.59	0.46
9	x				x			182	0.28	0.59	0.44
10	x	x	x	x	x	x	x	187	0.06	0.65	0.35
11	x				x			143	-0.08	0.66	0.29
12			x	x		x		64	0.45	-0.12	0.16
13		x		x	x	x		195	-0.25	0.34	0.05
14	x	x	x				x	84	-0.79	0.34	-0.22
15		x	x	x		x	x	134	-1.45	0.81	-0.32
16	x		x	x		x		199	-0.06	-0.68	-0.37
17			x			x	x	57	-0.07	-3.63	-1.85
18			x		x			60	-0.57	-3.54	-2.06
19	x	x	x	x	x		x	251	-1.38	-3.17	-2.27
20			x	x		x	x	83	-1.99	-10.02	-6.00
21	x		x					102	-0.68	-11.81	-6.25
22		x	x			x		68	-0.33	-17.16	-8.75

表3：图2B中所示变量的变量描述

缩略词	变量描述
CRPSI	C反应蛋白（mg/L）
AGE	年龄（年）
FACITSCO	疲劳总评分
PSTSCO	PASI总评分
NTNFGR2	既往抗肿瘤坏死因子治疗次数
中性粒细胞/白细胞	中性粒细胞/白细胞 (%)
TNF抑制剂初治	TNF-α抑制剂初治
基线时甲氨蝶呤使用情况	基线时甲氨蝶呤使用情况
尿糖试纸	尿糖试纸
银屑病关节炎78个关节基线调整后压痛关节总分	银屑病关节炎78个关节基线调整后压痛关节总分
银屑病关节炎确诊至今时间	自首次确诊银屑病关节炎以来的时间（年）
SEX	Sex
中性粒细胞指数	中性粒细胞（绝对值）（10E9/L）
MCS	SF-36心理健康总分
GLUCSSI	空腹血清葡萄糖（mmol/L）