超越点估计：迈向组内相关系数的恰当统计推断与报告

Yufeng Liu[1], Xiangfei Hong[2], * 和 Shanbao Tong[1], *

[1]上海交通大学生物医学工程学院，中国上海[2]上海交通大学医学院附属上海市精神卫生中心、上海市精神障碍重点实验室，中国上海

arXiv:2601.02765v1 [stat.ME] 6 Jan 2026

摘要

由于神经影像学研究以及许多其他生物医学研究的透明度和可重复性较差受到批评，使用组内相关系数（ICC）报告重测信度日益受到关注。因此，许多研究通过比较组内相关系数来评估其研究结果的信度，但往往未能检验组内相关系数之间是否存在统计学差异，也未报告置信区间。仅依赖点估计可能无法对总体层面的差异作出有效推断，并损害结论的信度。为解决这一问题，本研究系统回顾了2022年至2024年发表于*NeuroImage*的文章中组内相关系数的使用情况，并指出组内相关系数误报和误用的普遍程度。我们还提供了针对组内相关系数进行恰当统计推断的实用指南。对于该领域的从业者，我们推出了一款在线应用程序，可在使用组内相关系数时进行统计检验和样本量估计。我们重新计算了所回顾文章中报告的置信区间，并对组内相关系数值进行了正式检验，从而重新评估原有推断。结果表明，单纯依赖点估计可能导致不可靠甚至误导性的结论。具体而言，所回顾的11篇文章中，只有两篇基于组内相关系数作出了明确有效的统计推断，而另有两篇完全未能得出任何有效推断，这引发了对该领域研究结果可重复性的严重担忧。这些结果凸显了报告和解读组内相关系数时亟需采用严谨的推断框架。

关键词： 组内相关 系数、重测信度、神经影像 学

引言

重测信度用于评估在同一队列中反复施测时测量结果的稳定性[48]，。随着科学界可重复性危机日益受到关注，人们普遍担心当前已发表的许多、甚至大多数研究发现都是错误的，这使得重测信度也越来越受到重视 [25]。一项*Nature* 调查报告称，超过 70% 的研究人员未能重复同事的实验，超过一半的研究人员无法重复自己的实验[4]。可重复性有限会削弱公众信任、阻碍科学进步、浪费资源，并误导处于职业生涯早期的研究人员 [36]。这些问题在神经影像学研究中尤为突出，因为噪声数据和传统建模方法往往只能解释不到一半的观测变异，导致相当一部分变异无法得到解释 [11]。大量未解释的方差不仅会削弱神经影像学测量指标的信度，还会降低真实脑行为关联的效应量估计，并增加所需的样本量 [23]。因此，评估测量方法的信度对于确保所报告的发现稳健可信至关重要，而这一需求远不止于神经影像学领域。

电子邮箱：liu.yf@sjtu.edu.cn（Yufeng Liu）、xfhong@sjtu.edu.cn（Xiangfei Hong）、stong@sjtu.edu.cn（Shanbao Tong）。*共同通讯作者。

心理学和神经影像学领域的早期研究常使用 Pearson相关系数 (r) [8]报告重测信度。例如，静息态功能性近红外光谱数据的重测分析最初依赖 r 来量化不同测量时段之间的相似性[45], ；许多心理量表历来通过重复施测结果之间的 Pearson 相关来报告信度 [5, 32]。然而，Pearson 的r仅衡量线性关联或排序一致性，对不同测量时段之间的系统性差异并不敏感。例如，即使一组测量值是另一组平移或缩放后的结果，两组数据仍可能得到 $r = 1.0$ 。这意味着，存在系统偏差时，r 可能会大幅高估信度 [31]。

目前推荐且应用更广泛的方法是使用组内相关系数（ICC），它是根据同一受试者的重复测量所产生的方差分量计算得出的 [33, 48]。与 Pearson 的 r 不同，ICC 同时考虑了随机误差和系统性方差来源，因此能更全面、更易于解释地衡量重测信度[28]。

根据其正式定义，ICC 量化了总方差中可归因于受试者间差异的比例。因此，较高的 ICC 表明被试内变异较低、时间稳定性较高 [31]。McGraw 和 Wong [31] 提出的框架概述了针对不同实验设计的多种 ICC 模型；本文重点介绍应用最广泛的一种形式，该形式可在方差分析框架 [48]中表示。假设有 N 名受试者，每人测量 k 次，则受试者 j 的第 i次测量可用以下方差分析模型表示：

$$Y_{ij} = \mu + b_j + w_{ij} \tag{1}$$

其中， Y_{ij} 是观测分数， μ 是总体均值， b_j 是受试者 j 的随机效应（表示受试者 j的均值偏离 μ的程度），w_{ij} 是该次测量的残差误差。假定这些随机效应的方差分别为 $Var(b_j) = \sigma_b^2$ 和 $Var(w_{ij}) = \sigma_w^2$。在此模型下，ICC 定义为受试者间方差占总方差的比例：

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{2}$$

根据有限观测值计算得出的组内相关系数，只能描述当前样本。为了对总体层面作出有效推断，统计学家强调，应通过推断统计量化不确定性，例如在频率学派框架下使用置信区间或假设检验，或在贝叶斯框架下使用可信区间。

然而，在实践中，这些推断常常被忽视。许多研究仅依赖组内相关系数的点估计，就得出有关总体信度的结论或比较不同群体和研究，却不报告不确定性、不进行正式统计检验，也不报告置信区间。这类做法掩盖了固有的抽样变异，并增加了出现假阳性结果的风险[1, 10, 12, 29]。即使报告了置信区间，解读错误仍很常见，例如误以为置信区间重叠就必然意味着差异不显著。以这些方式误用和误解组内相关系数，不仅会夸大表面效应，还会损害科学结果的可重复性。

在本研究中，我们首先全面回顾了2022年至2024年发表于顶尖神经影像学期刊*NeuroImage*的文章，重点考察报告组内相关系数时仅提供点估计、未给出置信区间或统计检验的情况。在这些发现的基础上，我们提倡常规采用推断方法，例如置信区间、显著性检验和贝叶斯可信区间，以报告和推断组内相关系数。为支持这一做法，我们开发了一款开源网络应用程序，可用于：**(1)** 将单个组内相关系数与参照值进行评估；**(2)** 比较两个组内相关系数之间的差异；**(3)** 估计置信区间；以及**(4)** 估算基于组内相关系数的研究设计所需的样本量。借助该应用程序，我们系统地重新评估了上述*NeuroImage* 文章中报告的组内相关系数，以检验其原始结论在经过适当的推断统计分析后是否仍然成立。

方法

文献检索、数据收集与分析

我们系统地回顾了2022年至2024年发表于$NeuroImage$ 的研究文章。检索条件为：TI=(重测) OR AB=(重测) OR AK=(重测) A
在这些结果中，我们仅保留报告了原始组内相关系数计算结果的文章，排除了仅引用先前出版物中组内相关系数值的文章。文章筛选流程见图1(a)。

选定文章后，我们记录了各项研究中组内相关系数的报告和应用方式。具体而言，我们记录了组内相关系数值是仅以点估计的形式呈现，还是同时报告了推断统计量（例如置信区间、假设检验或贝叶斯可信区间）。随后，我们汇总了这些组内相关系数报告，以展示推断方法应用中常见的不足。

对于每个报告的组内相关系数，我们在可能的情况下提取以下信息：

1. 组内相关系数的点估计；2. 组内相关系数的置信区间或贝叶斯可信区间；3. 样本量；

4. 重测次数；

5. 作者对组内相关系数数值及其推断所得的结论、陈述或解释。

当组内相关系数仅以图形形式呈现（如柱状图）时，我们使用 Plot Digitizer（3.3.9 版，专业版；https://plotdigitizer.com/）对图形进行数字化处理，以获取相应的数值。根据提取的信息，我们重点关注两种常见的组内相关系数推断：

- **推断总体中的重测信度（单个组内相关系数评估）** ——将估计的组内相关系数与参照值进行比较，以推断其强度（例如，当组内相关系数 >0.75 时，得出某项测量在重测中具有“良好”信度的结论）。

- **推断两个总体之间重测信度的差异（两个组内相关系数比较）** ——比较不同测量、实验条件或研究中的组内相关系数，以推断哪一个的信度更高。

在这两种情形下，我们都审查了所报告的推断是否仅依赖点估计，还是有推断统计的支持，例如置信区间、正式的零假设显著性检验或贝叶斯可信区间。

OpenICC：用于组内相关系数推断和统计功效分析的基于网页的应用程序

为促进在实证研究中报告可复现的组内相关系数，我们开发了 OpenICC，这是一个开源的基于网页的应用程序，支持组内相关系数的统计推断，以及样本量和功效分析。该应用程序包含四个主要模块：

- **单个组内相关系数评估：**计算单个组内相关系数的置信区间，或针对参考值进行显著性检验，以推断重测信度的大小。

- **两个相关组内相关系数比较：**计算两个相关组内相关系数差异的置信区间，或对组内相关系数差异进行显著性检验。

- **两个独立组内相关系数比较：**计算两个独立组内相关系数差异的置信区间，或对组内相关系数差异进行显著性检验。

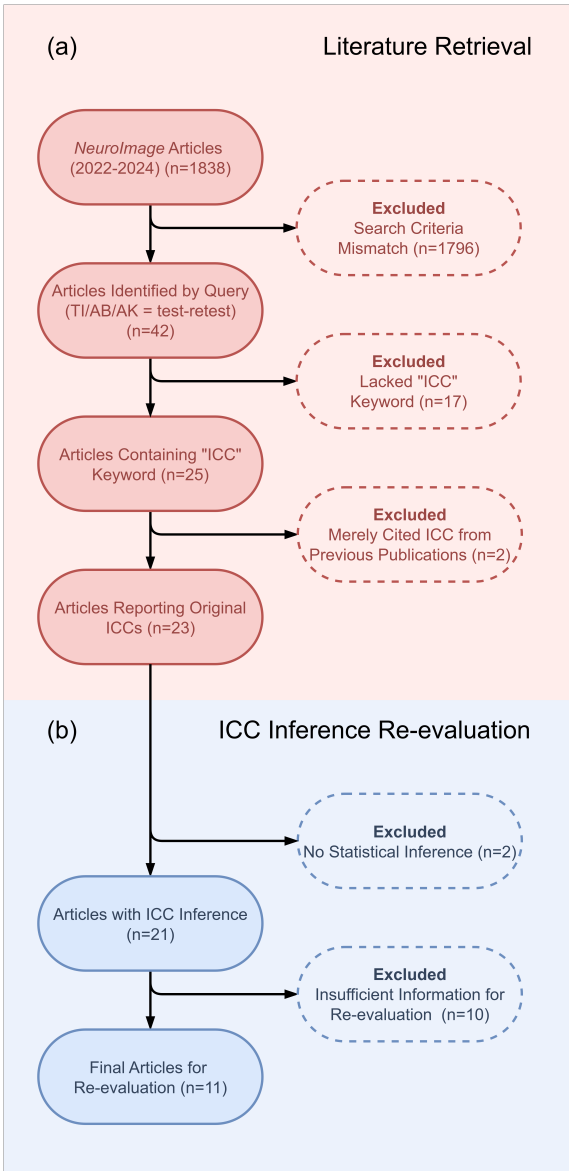

图 1：文献检索与组内相关系数重新评估流程图。(a) 阶段 1：文献检索。我们筛选出 23 篇*NeuroImage* 文章（2022–2024），这些文章报告了重测信度的原始组内相关系数估计值。(b) 阶段 2：组内相关系数推断重新评估。在这 23 篇文章中，有 10 篇提供了足够的数据，可用于重新评估其报告的组内相关系数推断。

- **样本量估计：**估计针对单个组内相关系数进行推断以及比较两个组内相关系数所需的样本量，适用于相关样本或独立样本。

后续章节将详细阐述各模块的统计原理。请注意，单个组内相关系数的样本量估计使用了 `ICC.Sample.Size` R 包 [39], ，该包实现了 Zou[47]提出的功效分析方法。为简洁起见，此处不详细列出完整公式。

(1) 单个组内相关系数评估

显著性检验。假设有 N 名参与者使用某一给定工具接受评估，共进行 k 次重复测量。目标是确定所得的重测信度（以组内相关系数衡量）是否达到令人满意的水平。为此，我们可以采用零假设显著性检验。以单侧检验组内相关系数是否大于某个参照值 ρ_0 为例，零假设和备择假设可表述为：

$$H_0 : ICC \leq \rho_0 \quad \text{vs.} \quad H_1 : ICC > \rho_0, \tag{3}$$

其中， ρ_0 是预先指定的基准值。根据 [15] 和方差分析（ANOVA），样本组内相关系数(r)可表示为（4）：

$$r = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (k-1)\text{MSW}}, \tag{4}$$

其中，MSB 和 MSW 分别是受试者间和受试者内的均方误差，或方差。进行假设检验时，使用针对组内相关系数的 Fisher z 变换，该变换不同于 Pearson相关系数的标准变换，用式（5）将 r 转换为 z ，转换后的值近似服从正态分布： [15, 20, 31],

$$Z_r = \frac{1}{2} \ln \frac{1 + (k-1)r}{1 - r}. \tag{5}$$

因此，在零假设下，抽样分布的期望均值为：

$$Z_{\rho_0} = \frac{1}{2} \ln \frac{1 + (k-1)\rho_0}{1 - \rho_0}, \tag{6}$$

其方差近似为：

$$V_{\text{single}} = \begin{cases} \dfrac{k}{2(k-1)(N-2)}, & k > 2, \\ \dfrac{1}{N - \frac{3}{2}}, & k = 2. \end{cases} \tag{7}$$

检验统计量 T_{Z_r} 定义如下：

$$T_{Z_r} = \frac{Z_r - Z_{\rho_0}}{\sqrt{V_{\text{single}}}}, \tag{8}$$

在 H_0 且 N较大时，其近似服从标准正态分布。如果 $T_{Z_r} > z_\alpha$，其中 z_α 是在显著性水平 α下标准化正态分布的单侧临界值，则我们拒绝 H_0 ，并推断总体组内相关系数(ρ)大于 ρ_0。

置信区间计算。 使用相同的 ICC 费舍尔 z 变换及其逆变换，我们还可以根据 ICC(r)的点估计量构建置信区间。在 $100(1-\alpha)$的置信水平下， Z_r 的 Z_ρ 区间估计为：

$$Z_{L,U} = Z_r \pm Z_{\alpha/2}\sqrt{V_{\text{single}}}. \tag{9}$$

对组内相关系数应用 Fisher z 逆变换，可得到原始组内相关系数尺度上的置信区间：

$$r_L = \frac{e^{2Z_L}-1}{e^{2Z_L}+k-1},\ r_U = \frac{e^{2Z_U}-1}{e^{2Z_U}+k-1}. \tag{10}$$

当仅知道 N、 k 和 r 的值，且无法获取原始数据时，Fisher z 组内相关系数变换尤为适用。当可以获取原始数据时，则可使用 `irr` [22]等 R 软件包实施更稳健的组内相关系数估计与检验程序。该软件包采用 Shrout 和 Fleiss[40] ，建立的方差分析框架，利用特定方差分析模型（例如单因素或双因素模型）得出的均方来计算不同形式的组内相关系数。

注：以上假设检验和置信区间估计均基于大样本假设。在 α 水平上拒绝零假设 $H_0 : ICC \le \rho_0$ ，等同于观察到参照值 ρ_0 位于 $(1-2\alpha)$ 置信区间下限之外。

（2）相关组内相关系数比较

显著性检验。考虑一组由 N 名参与者组成的群体，他们接受两种不同测量工具的评估，或在两种不同条件下接受评估。研究目的是推断两组测量值的重测信度之间是否存在显著差异，这通常用于比较同一队列受试者使用两种不同测量方法时的重测信度。假设每位参与者使用第一种测量工具测量 k_1 次，使用第二种测量工具测量 k_2 次。我们假设测量次数相等，即 $k_1 = k_2 = k$。令 ρ_1 和 ρ_2 分别表示第一种和第二种测量工具的总体组内相关系数。零假设和备择假设表示如下：

$$H_0 : \rho_1 = \rho_2 \ vs. \ H_1 : \rho_1 \neq \rho_2 \tag{11}$$

按照单个组内相关系数检验的相同步骤，我们可以对样本组内相关系数 r_1 和 r_2 [15] 使用 Fisher 的 z 变换（公式 (5)），从而得到变量 Z_{r_1} 和 Z_{r_2}。

检验统计量基于变换后的系数之间的差异。在零假设 $H_0 : \rho_1 = \rho_2 = \rho$ 下，预期差异 θ 为：

$$\theta = E(Z_{r_1} - Z_{r_2}) = \frac{1}{2}\ln\left\{\frac{1+(k-1)\rho}{1+(k-1)\rho}\right\} = 0 \tag{12}$$

该差异的理论方差 V 为：

$$\mathrm{V} = \mathrm{var}(Z_{r_1} - Z_{r_2}) = V_1 + V_2 - 2\,\mathrm{cov}(Z_{r_1}, Z_{r_2}) \tag{13}$$

其中，理论协方差项定义如下：

$$\mathrm{cov}(Z_{r_1}, Z_{r_2}) = \frac{k^2\rho_{12}^2}{2N\{1+(k-1)\rho_1\}\{1+(k-1)\rho_2\}} \tag{14}$$

这里， V_1 和 V_2 分别是 Z_{r_1} 和 Z_{r_2}的方差，如公式（7）所定义。 ρ_{12} 是两个测量工具所得测量值之间的总体组内相关系数。其样本估计值 r_{12}通过对一组完整的配对数据计算皮尔逊 r 得出。具体而言，对于N 位参与者中的每一位，其使用第一种测量工具获得的每个 k_1 测量值，都会与使用第二种测量工具获得的每个k_2 测量值配对，最终得到 Nk_1k_2 对数据用于计算。随后构造沃尔德检验统计量$T_{Z_r,\ \mathrm{dep}}$，如下所示：

$$T_{Z_r,\mathrm{dep}} = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\hat{\hat{V}}}} \tag{15}$$

^该统计量近似服从标准正态偏差分布。将 ρ_{12}、 ρ_1、 ρ_2分别替换为r_{12}、 r_1、 r_2 ，即可得到 V。当$|T_{Z_r,\ \mathrm{dep}}| > z_{\alpha/2}$时，在显著性水平 α 下拒绝零假设。在我们的验证计算中，由于无法获得原始数据的类间相关，r_{12} 被设为 0。这种简化可能导致更加保守的结论。

置信区间估计。 根据 [38], 提出的方法，计算了两个组内相关系数之差（$\rho_1 - \rho_2$）的 $100(1-\alpha)\%$ 置信区间，即 $[L, U]$。设 ρ_1 和 ρ_2 的置信区间分别为 (l_1, u_1) 和 (l_2, u_2)。下限 L为：

$$L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 - 2\widehat{\mathrm{corr}}(r_1, r_2)(r_1 - l_1)(u_2 - r_2) + (u_2 - r_2)^2} \tag{16}$$

上限 U为：

$$U = r_1 - r_2 + \sqrt{(u_1 - r_1)^2 - 2\widehat{\mathrm{corr}}(r_1, r_2)(u_1 - r_1)(r_2 - l_2) + (r_2 - l_2)^2} \tag{17}$$

其中

$$\widehat{\mathrm{corr}}(r_1, r_2) = \frac{k(k-1)}{[1 + (k-1)r_1][1 + (k-1)r_2]} r_{12}^2 \tag{18}$$

（3）独立组内相关系数比较

显著性检验。 考虑这样一种情形：有两个不同的队列，分别包含 N_1 和 N_2 名参与者。假设对第一组参与者进行 k_1 次测量，对第二组参与者进行 k_2 次测量。我们假设两组的参与者人数和测量次数相同，即 $N_1 = N_2 = N$ 和 $k_1 = k_2 = k$。目标是根据样本组内相关系数（r_1 和 r_2）比较总体组内相关系数（ρ_1 和 ρ_2）。在这种情况下，相关样本检验的公式可以简化。根据定义，总体类间相关 ρ_{12} 为0，因此理论协方差（式（14））为0。由此，检验统计量（式（15））可简化为：

$$T_{Z_r,\mathrm{indep}} = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\hat{V}_1 + \hat{V}_2}} \tag{19}$$

所得$T_{Z_r,\ \text{独立}}$ 值与标准正态分布中的$z_{\alpha/2}$ 临界值进行比较，以确定p值。

置信区间估计。 同样，将样本相关系数（式（18））设为 0 后，两个组内相关系数之差的 $(1-\alpha)100\%$ 置信区间可由（式（16）、（17））简化为：

$$L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2} \tag{20}$$

$$U = r_1 - r_2 + \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2} \tag{21}$$

（4）ICC比较的样本量

由于置信区间的宽度和显著性水平（即 p值）都对样本量高度敏感，因此确定适当的样本量是功效分析的基本组成部分。目标是确定所需参与者的最少人数，以便针对给定的效应量控制 I 类和 II 类错误。因此，样本量的确定是实验设计中一项常规且必不可少的步骤。

因此，在 OpenICC 应用程序中，我们提供了比较两个组内相关系数的样本量估计程序。具体方法略有不同，取决于这两个组内相关系数是相关的（源自同一队列中的受试者）还是独立的（分别来自两个不同的受试者组）。

相关样本。 这种情形涉及一组 N 参与者，每位参与者均使用两种不同的测量工具进行评估。每种测量工具都会进行多次评估，因此可以分别计算其组内相关系数。样本量的计算依据是用于比较两个组内相关系数的检验统计量的性质。如小节“（2）相关组内相关系数比较”所述，该分析基于组内相关系数的 Fisher z 变换。设 Z_{r_1} 和 Z_{r_2} 为上述定义的变换后样本组内相关系数。它们的差异 $(Z_{r_1} - Z_{r_2})$近似服从正态分布：

$$(Z_{r_1} - Z_{r_2}) \sim \mathrm{N}(\theta, V), \tag{22}$$

其中， $\theta = E(Z_{r_1} - Z_{r_2})$ 是差异的总体均值， $V = \mathrm{var}(Z_{r_1} - Z_{r_2})$ 是其理论方差（定义见式（13））。若要以 $(1-\beta)$ 的目标统计功效，在显著性水平为 α时检测指定的差异$|\theta| = d$ ，则必须满足以下不等式：

$$d - Z_{\alpha/2}\sqrt{V} \geq Z_{\beta}\sqrt{V}, \tag{23}$$

其中， α 表示I类错误的概率， β 表示II类错误的概率。 $Z_{\alpha/2}$ 和 Z_{β} 分别是标准正态分布中对应于上尾概率 $\alpha/2$ 和 β的临界值。重新整理该不等式，即可求出估计量所需的方差：

$$V \leq \frac{d^2}{(Z_{\beta} + Z_{\alpha/2})^2}. \tag{24}$$

假设两种测量工具的测量值数量相等 $(k_1 = k_2 = k)$，则估计量的方差为：

$$V = \begin{cases} \dfrac{k}{(k-1)(N-2)} - \dfrac{k^2\rho_{12}^2}{N\{1+(k-1)\rho_1\}\{1+(k-1)\rho_2\}}, & k > 2, \\ \dfrac{2}{N-\frac{3}{2}} - \dfrac{k^2\rho_{12}^2}{N\{1+(k-1)\rho_1\}\{1+(k-1)\rho_2\}}, & k = 2. \end{cases} \tag{25}$$

对方差项进行简化，以便分离出 N。我们将分母$(N-2)$ （对于 $k > 2$）和 $(N - 3/2)$ （对于 $k = 2$）近似为 N；当样本量足够大时，这种近似渐近有效。此简化统一了不等式（25）中的不同情形，因此不等式可表示为：

$$\frac{k}{(k-1)N} - \frac{k^2\rho_{12}^2}{N\{1+(k-1)\rho_1\}\{1+(k-1)\rho_2\}} \leq \frac{d^2}{(Z_{\beta} + Z_{\alpha/2})^2} \tag{26}$$

通过重新排列各项以分离出 N，我们可以确定所需的最小样本量。所得公式如下：

$$N \geq (\frac{k}{(k-1)} - \frac{k^2\rho_{12}^2}{[1+(k-1)\rho_1][1+(k-1)\rho_2]}) \cdot \frac{(Z_\beta + Z_{\alpha/2})^2}{d^2} \tag{27}$$

独立样本。比较两个独立组的组内相关系数遵循类似的原则。关键差异在于，对于独立样本，类间相关 ρ_{12} 等于 0。在公式（27）中令 $\rho_{12} = 0$，协方差项便会消失，公式也因此大幅简化。所得公式给出每组所需的样本量（N），假设两组的样本量相等：

$$N \geq \left(\frac{k}{k-1}\right) \cdot \frac{(Z_\beta + Z_{\alpha/2})^2}{d^2} \tag{28}$$

重新评估*NeuroImage*论文中的组内相关系数推断

我们仅重新评估了信息完整、可供核验的组内相关系数推断。具体而言，重新计算组内相关系数置信区间及相应原假设显著性检验统计量所需的信息包括样本量、试验次数和具体的组内相关系数值。图1(b)展示了重新评估过程中筛选论文的流程。若一篇论文报告了多项基于组内相关系数的陈述，我们会核验其中最核心的内容（最多两项），通常是摘要部分中的陈述。如果摘要中未明确提及组内相关系数，我们就重新评估正文中依赖组内相关系数的关键推断陈述。

如果某项推断陈述基于特定的组内相关系数（或多个组内相关系数），我们便使用OpenICC重新计算置信区间并进行原假设显著性检验；否则，如果未指定，则核验所有相关的组内相关系数。补充材料（Excel文件，表S1）提供了所选推断陈述及待核验组内相关系数值的完整清单。假设检验根据推断类型进行设计：

1. 比较两个组内相关系数时采用了双侧检验；

2. 将单个组内相关系数与参照值进行比较时采用了单侧检验，参照值取自原始文章，或依据 [28] 中的信度水平建议：

   - <0.5：较差
   - 0.5-0.75：中等
   - 0.75-0.9：良好
   - >0.9：优秀

结果

重测研究中误用和误报组内相关系数的普遍程度

如图1(a)所示，检索过程最终得到23篇报告了原始组内相关系数计算结果的文章（图1(b)）。在排除两篇未涉及推断的文章后，剩下21篇文章包含推断性陈述或分析。其中，只有5篇报告了置信区间或可信区间。在其余16项未报告此类区间的研究中，有10项未进行任何统计显著性检验，而仅依靠组内相关系数的点估计进行推断。值得注意的是，其中一项研究甚至误用了样本量[19]。作者将1,395个感兴趣区（ROIs）视为独立样本，以计算重测信度的组内相关系数。然而，每位受试者的155个感兴趣区无论如何都完全不具备独立性（见[19]中的图7(c)和图8(c)）。不过，另外六项研究中的四项进行了组间检验，尽管其组内相关系数均源自同一数据集[3, 27, 30, 43]。在统计检验中将来自相关来源的组内相关系数视为独立观测所引发的问题，将在讨论部分进行讨论。

表 1：组内相关系数报告与推断概览

文章	报告置信区间/可信区间	组内相关系数推断	推断方法
Amemiya 等人，2022[2]	No	单一/差异	点估计
Aye 等人，2022[3]	No	差异	点估计/零假设显著性检验*
Bollack 等人， 2023[6]	Yes	单一/差异	CI*
Bordin 等人， 2023[7]	No	单一/差异	点估计
Cao 等人， 2023[9]	No	单一/差异	点估计/零假设显著性检验
De Lange 等， 2023[14]	No	单一/差异	点估计
Dular 等， 2024[16]	No	单一/差异	点估计
Dumais 等人，2022[17]	No	单个	点估计
Faber 等人，2022[18]	Yes	差异	点估计
Feng 等人， 2024[19]	No	单个	点估计
Flournoy 等人， 2024[21]	Yes	单个	CrI
Hu 等人，2022[24]	Yes	差异	零假设显著性检验*
Kai 等人，2022[26]	No	差异	点估计
Kennedy 等人，2022[27]	No	单一/差异	点估计/零假设显著性检验*
Lin 等人， 2024[30]	No	单一/差异	点估计/零假设显著性检验*
Oishi 等人,2023[34]	No	No	/
Pirastru 等人，2023 [35]	No	单一/差异	点估计
Radwan 等人,2022[37]	Yes	No	/
Tetereva 等人,2022[41]	Yes	单个	点估计
Tregidgo 等人， 2023[42]	No	单个	点估计
Wang 等人， 2024[43]	No	单一/差异	点估计/零假设显著性检验*
Xu等， 2023[44]	Yes	单个	点估计
Zhao等， 2023[46]	No	单一/差异	点估计
摘要	**是：31.58%**	**单一/差异：89.47%**	**正确** 零假设显著性检验/置信区间/可信区间：**25%**

注：报告置信区间/可信区间 表示是否提供了置信区间或可信区间。**组内相关系数推断** 说明是否对组内相关系数进行了推断分析；若有，则说明分析类型：单个 = 对单个组内相关系数的评估；差异 = 对两个组内相关系数的比较；单一/差异 =同时进行了这两种分析。**推断方法** 表示推断所依据的统计方法（点估计 = 仅基于点估计进行推断；零假设显著性检验 = 原假设显著性检验；置信区间 =使用置信区间；可信区间 =使用可信区间）。摘要 行报告了以下文章所占比例：(1) 报告了置信区间/可信区间；(2) 进行了任何推断分析；(3) 在进行推断统计的文章中使用了正确的推断方法（包括零假设显著性检验/置信区间/可信区间）。“推断方法”列中的星号（*）表示所采用的推断方法有误或存在缺陷，即使使用了零假设显著性检验、置信区间或可信区间。

此外，即使在报告了置信区间或可信区间的五篇文章中，也观察到了几种形式的 ICC 误报或误用，包括：

(1) **尽管有可用的置信区间，却仅依赖点估计进行推断**。[18,41, 44] 例如， [18] 仅根据点估计就对组内相关系数作出不当推断，未纳入现有区间信息。作者写道：

> ”$CerebNet$ 和 ACAPULCO^{rt} 在两个数据集中的组内相关系数介于 0.635 至 0.997 之间；在大多数情况下， $CerebNet$的结果更为一致（组内相关系数更高）。事实上，在 Kirby 数据的 27 个子结构中，有 24 个的组内相关系数在 $CerebNet$ 方面优于 ACAPULCO^{rt} ；在 OASIS1 数据集中，27 个子结构中也有 23 个如此，蚓部以及左、右半球皮层灰质的合并区域亦是如此。”

这种方法仅根据哪种方法得出的组内相关系数点估计“更高”来比较方法。忽略置信区间的这种推理未能考虑抽样变异，可能会将随机噪声误解为信度上的实质性差异。

(2) **误解置信区间重叠的含义**。一项研究错误地认为，置信区间重叠必然表明不存在统计显著差异[6]。具体而言，作者指出：

> “置信区间重叠表明，没有证据显示这些方法之间存在差异。”

这种解读具有误导性。一般而言，置信区间重叠并不意味着不存在统计显著差异。根据 [13], ，比较两个独立均值时，若两个 95% 置信区间的重叠部分小于平均误差范围的大约一半，通常表明结果在 $p \leq 0.05$ 水平上具有统计显著性。尽管这条准则针对的是均值，但其概念原则——即区间重叠不等同于假设检验——同样适用于组内相关系数。

(3) **错误地实施自助法程序**。在一个案例中，研究人员使用自助法程序比较两个组内相关系数，但采用了有缺陷的推断方法 [24]。具体而言，作者将自助法迭代视为独立观测，并使用双样本 t 检验对其进行比较，文中写道：

> “基于峰值和逐点分析的 ICC(A,1) 均值和置信区间通过自助法获得了 1600 次 [.....] 采用双样本 t 检验比较基于峰值和逐点分析的结果。”

这种方法在统计学上是无效的，因为自助法得出的 ICC 值并非独立样本，不应使用双样本 t 检验进行分析。关于这一错误的详细分析，以及比较组内相关系数的正确自助法程序，请参见讨论部分的“ICC比较中错误使用自助法重抽样”一节。

OpenICC 应用程序

OpenICC 是一个基于网页的应用程序，无需任何编程专业知识即可进行组内相关系数推断。该应用程序基于 R Shiny 框架构建，提供了简洁、用户友好的界面，并分为四个模块，每个模块均针对不同的分析功能而设计（图 2）。下文将详细介绍这些模块。OpenICC 可通过 https://nelab.shinyapps.io/OpenICC/ 公开访问。

单个组内相关系数评估 模块用于对单个组内相关系数进行假设检验。用户可以输入样本 ICC，并指定受试者人数（N）、试验次数（k），以及零假设和备择假设

。此外，用户还可以设定显著性水平（α），并选择单尾检验或双尾检验。根据这些参数，输出包括样本在零假设下的显著性水平（p值），以及总体参数的双侧置信区间。

两个组内相关系数比较 模块用于评估两个组内相关系数之间是否存在显著差异。该界面允许用户指定受试者数量（N）、重复测量次数（k）以及待比较的两个组内相关系数（ρ_1 和ρ_2）。随后，用户需要指明样本是相关的还是独立的；对于相关设计，还必须提供类间相关（ρ_{12}）。该模块支持单尾和双尾检验，以及用户自定义的显著性水平（α）。模块将返回p值以及差异的双侧置信区间（$\rho_1 - \rho_2$）。此外，该模块还提供可视化图形，展示p值和置信区间上下限如何随假定的类间相关变化，从而帮助用户评估其推断的稳健性。

样本量估计（单个） 用于计算在预先设定的精确度和统计功效水平下估计单个组内相关系数所需的最小样本量（N）。该模块借助R软件包`ICC.Sample.Size`[39]的功能，要求用户指定每位受试者的重复测量次数（k）、备择假设下预期的总体组内相关系数（ρ_1）以及零假设下的组内相关系数值（ρ_0）。用户还需选择检验类型（单尾或双尾）、显著性水平（α）和期望的统计功效（$1-\beta$）。随后，模块将返回达到指定设计参数所需的最小样本量（n）。

样本量估计（差异）用于计算检测两个组内相关系数之间统计显著差异所需的最小样本量。输入面板会提示用户指定关键设计参数，包括样本类型（相关或独立）、重复测量次数（k）以及预期的组内相关系数值（ρ_1 和 ρ_2）。对于相关样本，用户还必须提供预期的类间相关（ρ_{12}）。用户可通过选择单尾或双尾检验、设定显著性水平（α）以及指定所需的统计功效（$1-\beta$）来设定统计标准。随后，该模块将返回在指定设计条件下可靠检测组内相关系数差异所需的最小样本量（N）。

重测研究中已报告的ICC推断验证

我们使用 OpenICC 重新评估了 11 篇文章（共 21 篇）中的推断；这些文章提供了足够的信息以供验证。表2汇总了每项研究的推断类型（单个组内相关系数评估、两个组内相关系数比较或两者兼有）、重新计算的组内相关系数推断数量（重新计算的 ICC），以及每篇文章中原始陈述与重新计算结果相一致的组内相关系数推断数量（一致 ICC）。

我们将“一致率”定义为“一致 ICC”除以“重新计算的 ICC”，以反映重新计算的推断中与原始陈述相符的比例。

在纳入审查的研究中，共重新计算了 283 项推断，其中 123 项与原始结论一致，总体一致率为 43.46%。在研究层面，11 篇文章中仅有两篇的组内相关系数推断完全一致。相比之下，有两篇文章在其评估的所有陈述中均未得出有效推断。

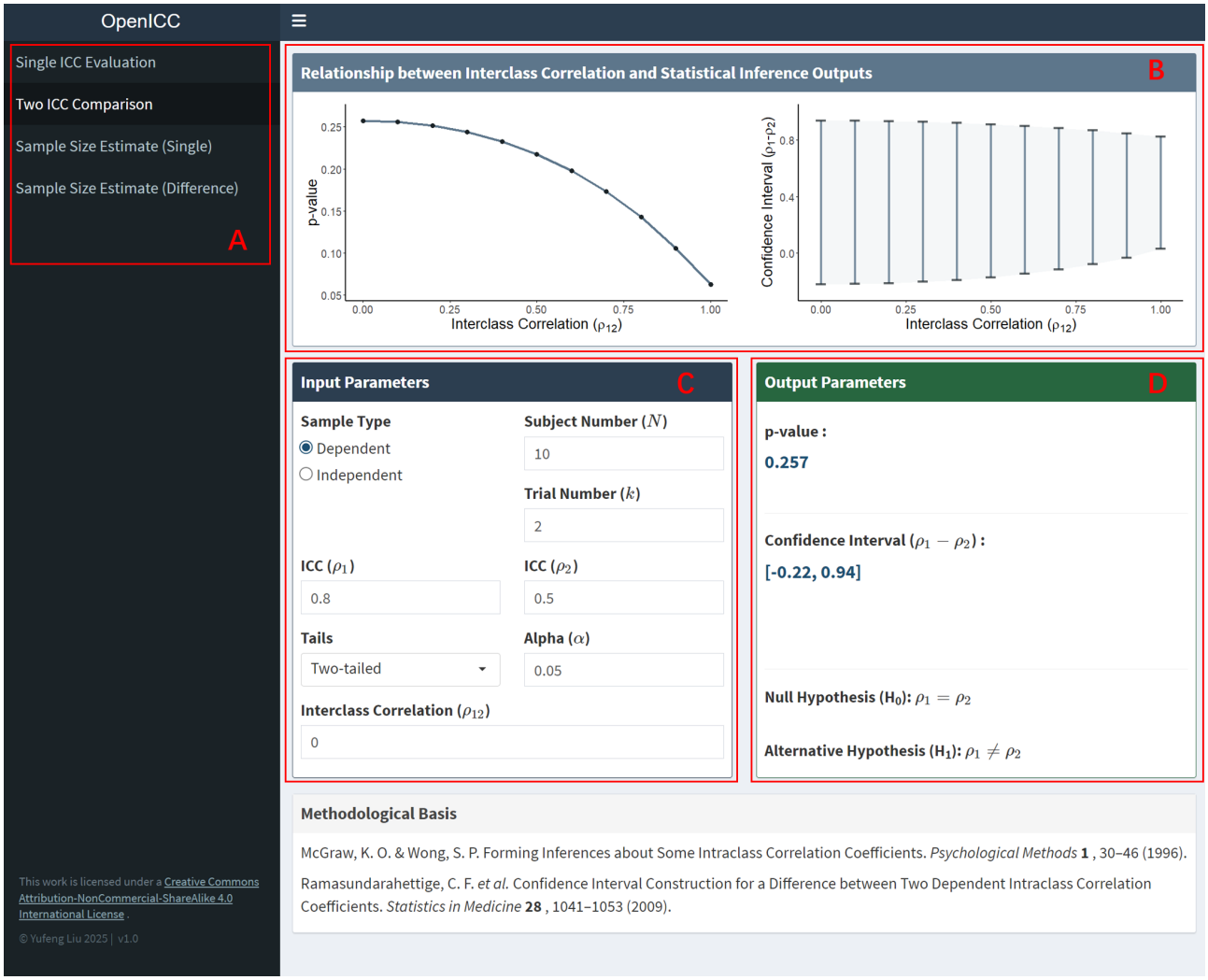

图 2：OpenICC 应用程序的用户界面，以“两个组内相关系数比较”模块为例进行说明。主布局由以下部分组成：(A) 导航侧边栏、(B) 可视化面板（如适用）、(C) 参数输入面板，以及 (D) 统计输出面板。

表2：*NeuroImage*中的 ICC 验证摘要

Paper	Inference Type	Recalculated ICC	Consistent ICC	Consistent Rate
Bollack et al., 2023[6]	Single	5	5	100.00%
Feng et al., 2024[19]	Single	3	3	100.00%
Pirastru et al., 2023[35]	Difference	30	20	66.67%
Xu et al., 2023[44]	Single	20	13	65.00%
Hu et al., 2022[24]	Difference	7	4	57.14%
Bordin et al., 2023[7]	Single	40	19	47.50%
Faber et al., 2022[18]	Difference	60	27	45.00%
Tregidgo et al., 2023[42]	Single	94	30	31.91%
Cao et al., 2023[9]	Difference	16	2	12.50%
Tetereva et al., 2022[41]	Single	2	0	0.00%
Zhao et al., 2023[46]	Single/Difference	6	0	0.00%
Summary	/	**283**	**123**	**43.46%**

注：推断类型 指明每项研究中所验证的推断类别。类型包括：单个 =单个组内相关系数的评估；差异 = 两个组内相关系数的比较；单一/差异 = 两种类型均包括在内。**重新计算的 ICC**表示每项研究中所验证的推断总数。**一致 ICC** 统计与原始陈述一致的重新计算组内相关系数。**一致率** 是经标准化的**一致 ICC**。**摘要** 行显示汇总总数。

讨论

违背　　独立性假设的区域间 ICC 比较　　s

这些研究中发现的一个关键方法学问题是违背独立性假设。正如结果部分所述，几篇文章进行了组间检验，尽管组内相关系数源自同一数据集 [3, 27, 30, 43]。例如，作者将区域间组内相关系数视为独立样本，并在 [3] 中陈述：

> “接下来，我们使用图 6 中展示的 57 个灰质 ROI，计算了 MT_{sat} 和 $R2^*$的区域 ICC。 [.....] 我们观察到，与皮层下灰质区域相比，MT_{sat} 在皮层区域中的 ICC 显著增加（z(56) = -2.093；$p = 0.036$；图 6）。”

虽然这种比较旨在突出区域差异，但其统计方法存在根本性缺陷。跨这 57 个区域计算均值和方差，意味着这些区域代表来自某一总体的随机、独立样本。然而，由于这些测量值来自同一批个体，它们具有固有的生理和解剖相关性。将这些相关区域视为独立区域，会人为夸大有效自由度，导致低估标准误。因此，这种对统计假设的违反会大幅增加 I类错误的风险，在不存在有意义差异的情况下产生虚假的统计显著性。

若要对来自同一队列的两组区域（例如皮层区域与皮层下区域）的组内相关系数值分布进行统计比较，建议采用受试者层面自助法程序。与

独立样本 t 检验不同，这种非参数方法能够考虑共享受试者群体所导致的 ICC 估计值之间的依赖性。该程序包括：

1. 从原始数据集中有放回地重抽样 N 名受试者，以生成一个自助法样本；

2. 基于这个新样本，重新计算所有皮层区域和皮层下区域的组内相关系数
；3. 计算本次迭代中皮层区域与皮层下区域组内相关系数的平均差；4. 重复这一过程（例如 $B = 1000$ 次），以构建差异的经验抽样分布。

随后，通过检查该差异分布的95% 置信区间是否包含零来进行统计推断。该方法严格检验一组区域的信度是否系统性地高于另一组，同时保留数据的底层相关结构。

ICC比较中自助法重抽样的错误用法

如结果部分所述，一项研究采用双样本 t 检验比较通过自助法生成的两组组内相关系数值分布，并将自助法迭代次数视为样本量 [24]。这种做法在方法学上存在三个方面的缺陷：

1. 依赖性：从同一数据集得出的自助法估计值并非独立样本；2. 正态性：组内相关系数的抽样分布通常呈偏态，违反了 t 检验的正态性假设；3. 统计功效膨胀：最关键的是，这种方法将自助法迭代次数（例如，$B = 1600$）视为样本量（N）。

比较两个相关的组内相关系数时，恰当的自助法程序是生成两者差异的抽样分布，并依据相应的置信区间进行统计推断。令 $(\theta = \rho_1 - \rho_2)$ 表示两个组内相关系数之间的总体差异。根据 (n) 名受试者的观测数据，我们首先计算样本统计量 (r_1) 和 (r_2)，得到观测差异 $(\theta = r_1 - r_2)$。对于每位受试者 (i)，令 $(\mathbf{X}i)$ 表示重复测量的完整向量，定义为 $(\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ik_1}, X_{i,k_1+1}, X_{i,k_1+2}, \dots, X_{i,k_1+k_2})$，其中 (k_1) 和 (k_2) 分别为两种方法的重复测量次数。

因此，正确的重抽样策略是将受试者作为完整单位进行重抽样——保留（$\mathbf{X}i$）的联合结构——以构建（θ）的经验分布。自助法算法如下：

1. 从原始数据集 $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ 中有放回地重抽样 n 受试者，生成自助法样本$\{\mathbf{X}_1^*, \dots, \mathbf{X}_n^*\}$。

2. 计算该自助法样本的组内相关系数，以得到r_1^* 和r_2^*。

3. 计算差异统计量：$\theta^* = r_1^* - r_2^*$。4. 重复

此过程 B 次，以生成差异的自助法分布：$\theta_1^*, \theta_2^*, \dots, \theta_B^*$。

随后通过根据该分布构建置信区间（CI）进行统计推断。使用百分位数法，令 θ_q^* 表示自助法分布的第 q 个分位数。差异 θ 的 $100(1-\alpha)\%$置信区间为：

$$\mathrm{CI} = [\hat{\theta}^*_{\alpha/2}, \quad \hat{\theta}^*_{1-\alpha/2}] \tag{29}$$

如果该区间不包含零，则认为两个组内相关系数之间的差异在 α 水平上具有统计显著性。

依赖于大小的不确定性在ICC比较中的体现

由于组内相关系数定义在有界尺度上，并且是根据方差分量估算的，其抽样分布通常是不对称的。因此，组内相关系数估计值的不确定性在其取值范围内并非均匀分布：接近上限的估计值往往不确定性较小、置信区间较窄，而处于中间范围的组内相关系数值通常表现出更大的变异性。因此，仅凭点估计无法体现不确定性如何随组内相关系数尺度变化。这一点具有重要意义：两个组内相关系数点估计之间相同的绝对差异，根据其大小不同，可能对应着显著不同程度的统计证据。

使用 OpenICC 中的两个组内相关系数比较模块，考虑一项包含 $N = 28$ 和 $k = 2$ 的研究（假设$r_{12} = 0$）。两个较高组内相关系数之间相差 0.10（例如，0.95 与 0.85）时，具有统计显著性，其 ICC差值的置信区间不包含零（ $p = 0.04$，95% 置信区间： $[0.01, 0.25]$）。然而，两个中等组内相关系数之间相同的数值差异（例如，0.75 与 0.65）不具有统计显著性，其置信区间包含零（$p = 0.47$，95% 置信区间： $[-0.18, 0.40]$）。

这进一步表明，仅凭点估计不足以刻画总体层面的差异。要确定观察到的数值差异是否反映了总体组内相关系数之间的真实差异，必须报告置信区间，或等价地，对组内相关系数差异进行正式的显著性检验。

样本量与重测次数之间的权衡

许多重测研究仅依赖一次重测。然而，增加重测次数可以大幅减少所需的样本量。当 r_1 和 r_2 位于 $[0, 1]$ 且其他参数保持不变时，公式（27）（28）中的不等式约束意味着，随着重测次数 k 增加，所需的最小样本量会单调减少。因此，在显著性水平（α）和统计功效固定的情况下，纳入额外的重测可以通过减少所需参与者数量来提高设计效率。

为说明这一效应，我们使用 OpenICC 的 样本量估计（差异） 模块，并设置以下参数： $r_1 = 0.8$ 、 $r_2 = 0.6$、$r_{12} = 0$ 、 $\alpha = 0.05$，以及统计功效 $= 0.8$。在这些条件下，进行两次重测（$k = 2$）的设计需要 96 名参与者。将重测次数增加到三次（$k = 3$）后，所需样本量减少至 64 名参与者；进一步增加到四次重测（$k = 4$）后，所需样本量降至 54 名参与者。

重要的是，增加 k 所带来的收益会逐渐递减。在此示例中，将 k 从 2 增至 3，可使所需样本量减少 32；而将 k 从 3 增至 4，仅能减少 10 名受试者。此外，每增加一次重测，都会加重参与者的负担，并可能提高流失（退出）的风险。因此，在考虑通过增加重测次数来减少样本量时，研究人员应权衡效率提升与这些实际成本。

相同的重测次数假设与适用范围

在 OpenICC 中，我们假设两种方法的重复测量次数相同（$k_1 = k_2$）。尽管理论上可以对重测次数不同的组内相关系数进行统计比较，[15], 但此类分析需要估计 $\theta = E(Z_{r_1} - Z_{r_2})$，而这又需要获取原始受试者层面数据。

结论

本研究系统考察了神经影像学研究中组内相关系数（组内相关系数）在重测信度分析中的使用情况，并发现发表于$NeuroImage$的文章中普遍存在非标准做法。采用标准化推断方法进行重新分析后发现，许多基于组内相关系数的

结论在省略置信区间和正式统计检验时可能并不可靠。为解决这些问题，我们提供了明确的方法学指南、标准化推断程序，以及一个可公开获取的用于基于组内相关系数的假设检验和样本量估计的工具箱。尽管我们的实证综述聚焦于*NeuroImage*，但本文提出的方法学问题与建议广泛适用于神经科学和生物医学研究中的重测研究。通过促进透明、严谨且标准化的组内相关系数推断，本研究旨在提高神经影像学研究中可靠性评估的可重复性和可信度。

致谢

本研究获科技部2030脑科学与类脑研究项目（2022ZD0208500）和国家自然科学基金委员会（82472097）资助。X Hong 获上海市科学技术委员会（23Y11900500）资助。

参考文献

1. Altman, D. G. 为什么需要置信区间。《World Journal of Surgery》29，554–556。（2005年）。2. Amemiya, S. 等。使用静息态功能磁共振成像指标测量的信度及对血流动力学改变的敏感性：与123I-IMP单光子发射计算机断层成像的比较。NeuroImage 263，119654。（2022年）。3. Aye, N. 等。脑微观结构多参数图（MPM）的重测信度。NeuroImage 256，119249。（2022年）。4. Baker, M. 1,500名科学家揭示研究可重复性的真相。Nature 533，452–454。（2016年）。5. Beck, A. T. 抑郁测量量表。《Archives of General Psychiatry》4，561。（1961年）。6. Bollack, A. 等。评估用于纵向正电子发射断层成像研究的新型数据驱动淀粉样蛋白 β 沉积指标。NeuroImage 280，120313。（2023年）。7. Bordin, V. 等。定量磁化率图的最佳回波时间：基于基底神经节和皮层下脑核团的重测研究。NeuroImage 278，120272。（2023年）。8. Bruton, A. 等。信度：它是什么，以及如何测量？《物理治疗》86，94–99。（2000年）。9. Cao, H. 等。相位编码方向对人类功能连接组重测信度的影响。NeuroImage 277，120238。（2023年）。10. Casella, G. & Berger, R. 统计推断，第2版。（2024年）。11. Chen, G. 等。组内相关：神经影像学的改进建模方法及应用。Human Brain Mapping 39，1187–1206（2018年）。12. Cumming, G. 新统计学：为什么以及如何开展。Psychological Science 25，7–29。（2014年）。13. Cumming, G. & Finch, S. 目测推断：置信区间以及如何阅读数据图。American Psychologist 60，170–180。（2005年）。14. DeLange, S. C. 等。使用CATO——连接分析工具箱进行结构和功能连接重建。NeuroImage 273，120108。（2023年）。15. Donner, A. & Zou, G. 检验相关组内相关系数的相等性。Journal of the Royal Statistical Society: Series D (The Statistician) 51，367–379。（2002年）。16. Dular, L. & Špiclin, Ž. BASE：脑龄标准化评估。NeuroImage 285，120469。（2024年）。17. Dumais, F. 等。eICAB：一种用于威利斯环多类别分割与分析的新型深度学习流程。NeuroImage 260，119425。（2022年）。18. Faber, J. 等。CerebNet：一种用于小脑详细子分割的快速可靠深度学习流程。NeuroImage 264，119703。（2022年）。

19. Feng, W. 等。使用双翻转角星形堆叠、压缩多梯度回波采集在两分钟内完成全脑多参数映射。NeuroImage 297，120689。（2024年）。20. Fisher, R. A. 研究工作者的统计方法。第5版（1934年）。21. Flournoy, J. C. 等。采用精准神经科学方法估计情绪加工过程中神经反应的信度：对任务态功能磁共振成像的启示。NeuroImage 285，120503。（2024年）。22. Gamer, M. 等。Irr：各种评定者间信度和一致性系数，2019年。23. Gell, M. 等。测量噪声如何限制脑行为预测的准确性。自然·通讯 15，10678。（2024年）。24. Hu, Z. 等。超越组效应的事件相关电位个体信度新视角：多感官研究与计算建模。NeuroImage 250，118937。（2022年）。25. Ioannidis, J. P. A. 为什么大多数已发表研究发现都是错误的。公共科学图书馆医学 2，e124。（ 2005 年）。26. Kai, J. 等。使用体内弥散磁共振成像绘制皮层下连接组：可行性与信度。NeuroImage 262，119553。（2022年）。27. Kennedy, J. T. 等。ABCD任务态功能磁共振成像数据的信度与稳定性挑战。NeuroImage 252，119046。（2022年）。28. Koo, T. K. & Li, M. Y. 信度研究中组内相关系数的选择与报告指南。《脊椎按摩疗法医学杂志》15，155–163。（2016年）。29. Lee, D. K. P值的备择假设：置信区间与效应量。《韩国麻醉学杂志》69，555–562。（2016年）。30. Lin, L. 等。基于体素的纹理相似性网络揭示个体差异，并与生物本体相关联。NeuroImage 297，120688。（2024年）。31. McGraw, K. O. & Wong, S. P. 对某些组内相关系数进行推断。Psychological Methods 1，30–46（1996年）。32. Mohammadi, F. 等。粗大动作发展测试第三版波斯语版本的心理测量学特性评估。《运动学习与发展杂志》7，106–121。（2019年）。33. Müller, R. & Büttner, P. 对组内相关系数的批判性讨论。Statistics in Medicine 13，2465–2476。（1994年）。34. Oishi, H. 等。活体人脑外侧膝状体各亚区的大分子组织体积映射。NeuroImage 265，119777。（2023年）。35. Pirastru, A. 等。情绪效价和刺激习惯化对情绪生成过程中功能磁共振成像信号信度的影响。NeuroImage 284，120457。（2023年）。36. Pusztai, L. 等。研究可重复性与临床前验证：问题与解决方案。《自然综述：临床肿瘤学》10，720–724。（2013年）。37. Radwan, A. M. 等。基于弥散磁共振成像约束球面反卷积的白质解剖图谱、其变异性及可重复性。NeuroImage 254，119029。（2022年）。38. Ramasundarahettige, C. F. 等。两个相关组内相关系数之差的置信区间构建。Statistics in Medicine 28，1041–1053（2009年）。39. Rathbone, A. 等。ICC.Sample.Size：组内相关系数的样本量和统计功效计算，2015年。40. Shrout, P. E. & Fleiss, J. L. 组内相关：用于评估评定者信度。心理学通报 86，420–428。（1979年）。41. Tetereva, A. 等。捕捉脑与认知的关系：整合跨任务的基于任务的功能磁共振成像可显著提升预测和重测信度。NeuroImage 263，119588。（2022年）。

42. Tregidgo, H. F. 等。 使用弥散磁共振成像和改进的组织学图谱对丘脑核团进行精确的贝叶斯分割。 *NeuroImage* **274,** 120129。（2023年）。43. Wang, Y. 等。 利用来自多个形态学特征的联合概率分布绘制形态学皮层网络。*NeuroImage* **296,** 120673。（2024年）。44. Xu, F. 等。 用于检测脑血流正常变异的三维速度选择性动脉自旋标记的重测信度。*NeuroImage* **271,** 120039。（2023年）。45. Zhang, H. 等。 基于功能性近红外光谱对独立成分分析衍生的静息态功能连接进行重测评估。*NeuroImage* **55,** 607–615。（2011年）。46. Zhao, C. 等。 在7特斯拉下进行全脑无畸变三维伪连续动脉自旋标记。*NeuroImage* **277,** 120251。（2023年）。47. Zou, G. Y. 以精度和保证度估计组内相关系数的样本量公式。*Statistics in Medicine* **31,** 3972–3981。（2012年）。48. Zuo, X.-N. & Xing, X.-X. 人脑功能连接组学静息态FMRI测量值的重测信度：系统神经科学视角。*Neuroscience & Biobehavioral Reviews* **45,** 100–118。（2014年）。