选择统计模型，将生物学交互作用评估为偏离效应加和性

David M. Thompson，Yan Daniel Zhao

摘要

Vanderweele 和 Knol 将生物学交互作用定义为“两种暴露在物理上相互作用，从而导致结局发生”的情形。生物学交互作用的一个标志是：因素共同作用产生的总效应，不同于这些因素独立作用时各自效应之和。

流行病学家构建统计模型来评估生物学交互作用。统计模型的形式决定了它是否适合检测效应加和性的偏离，或效应乘法性的偏离。学界普遍认为，应将生物学交互作用评估为偏离效应加和性。

本文比较了三种统计模型对一个数据示例的评估。该示例见于多本流行病学教科书，用于说明二项结局中的生物学交互作用。线性二项模型用概率差异来量化数据示例中偏离可加性的程度。该模型直接给出易于解释的估计值以及参数的 95% 置信区间，其中包括交互作用对比（IC）。对数二项模型和逻辑回归模型未检测到数据示例中存在偏离乘法性的情况。不过，这些模型的估计值有助于计算“交互作用所致相对超额风险”（RERI），该指标用于衡量相对风险尺度上偏离可加性的程度。

线性二项模型可直接评估效应加和性的偏离，结果易于解释，值得在研究和流行病学教学中得到更广泛的应用。针对该模型的局限性，也有相应的应对策略。

关键词：效应加和性和效应乘法性；生物学交互作用；统计交互作用；广义线性模型；交互作用对比（IC）；交互作用所致相对超额风险（RERI）

核心信息

- 流行病学界已形成共识，STROBE 声明也体现了这一点：应将生物学交互作用评估为偏离效应加和性的程度。

- 对数二项回归模型和逻辑回归模型在流行病学中被广泛用于评估生物学交互作用，尽管从统计形式上看，它们适用于检测效应乘法性的偏离。

- 对数二项回归模型和逻辑回归模型可以通过估计交互作用所致相对超额风险（RERI）等统计量，在相对风险尺度上量化可加性的偏离。然而，RERI 不能直接解释为超额风险，而且对 RERI 进行推断（估计和假设检验）较为复杂。

- 线性二项模型可使用现有的广义线性模型软件进行估计，并可直接估计交互作用对比（IC），其含义可解释为超额风险。

- 线性二项模型值得在研究和流行病学教学中得到更广泛的应用。

生物学交互作用与统计交互作用

在临床或人群健康研究中，人们常常关注与生物学交互作用相关的假设。Vanderweele 和 Knol [1 (第 54 页)] 将生物学交互作用定义为“两种暴露发生物理交互，从而导致结局”。Rothman [2 (第 171 页)]指出，“只要一种原因的效应取决于另一种原因是否存在，两种原因之间就会发生生物学交互作用”。

研究者构建统计模型来检测交互作用和效应修饰。Rothman[2 (p.169)] 指出，“在统计学中，‘交互作用’一词指偏离统计模型的既定形式。”模型的形式可以适用于检测偏离效应加和性的情况，也可以适用于检测偏离效应乘法性的情况。由于统计模型的形式会影响对统计交互作用的解释，Rothman [2 (p.170)]倾向于使用“效应测量修饰”一词，而非“交互作用”。

Rothman 将“生物学独立性”与效应加和性联系起来，并将“生物学交互作用”与偏离效应加和性联系起来。Rothman 问道：“为什么应将生物学交互作用评估为偏离效应加和性的情况” [2 （第178页）]？到2007年，STROBE 声明 认为，对 Rothman 这一反问的回应反映了这样一种“共识：采用绝对风险的加法尺度，比乘法尺度更为恰当 [，] 适用于公共卫生和临床决策” [3 （第817页）]。 STROBE 声明的作者提醒研究者，“在许多情况下，暴露相关的绝对风险比相对风险更受关注”，并建议他们“考虑

将相对风险估计值换算为具有实际意义时间段内的绝对风险” [3 （第825页）]。Vanderweele 和 Knol [1 （第37页）] 更明确地指出，“加性交互作用之所以重要（而非仅依赖乘性交互作用测量）”，一个原因是“它是更具相关性的公共卫生测量指标”。

效应的可加性与乘法性

本文与这一共识一致，但避免使用“加性交互作用”这一术语。相反，本文将这一概念与用于评估效应偏离效应加和性的统计模型联系起来。其中一种模型“风险差二项模型”[4], 直接以概率差的形式量化效应偏离效应加和性的程度，包括交互作用对比（IC）。该模型也称为“二项回归模型”[5, 6]。Richardson 等人[7], 将其作为边际结构模型的最后一步，并称之为“线性二项模型”，本文将沿用这一术语。

在线性二项模型中，检测到统计交互作用即构成效应偏离效应加和性的直接证据。对数二项模型和逻辑回归模型也可以间接评估可加性：将其相对风险或优势比估计值重新组合，可计算诸如“交互作用所致相对超额风险”（RERI）之类的统计量。

本文也避免使用“乘法交互作用”这一术语，但将这一概念与用于评估效应乘法性偏离证据的统计模型联系起来。对数二项模型以相对风险（亦称风险比、患病率

比 [4,7] 或患病率比率）的形式估计效应。逻辑回归模型以优势和优势比的形式估计效应。在对概率或相应优势进行对数变换的对数二项模型和逻辑回归模型中，检出统计交互作用构成了效应乘法性偏离的直接证据。

二项结局的统计模型

线性二项模型、对数二项模型和逻辑回归模型都是广义线性模型的例子。它们都将结局视为源自二项分布。每个模型都采用由若干项相加构成的线性预测变量。从这个角度看，所有广义线性模型都可以视为“可加的”。因此，本文不使用“可加模型或乘法模型”这一说法，而是讨论评估效应可加性或乘法性的统计模型。

这三种模型都将二项结局与线性预测变量联系起来。它们的区别在于所采用的连接函数。线性二项模型使用恒等连接函数，对数二项模型使用对数连接，逻辑回归模型则使用logit连接函数。因此，线性二项模型直接作用于概率，而其他模型则对概率或其对应优势进行对数变换。由于每种模型估计的效应度量不同，因此它们检测数据集中的统计交互作用的能力也有所不同。

在回顾效应加和性的定义后，我们借助一个广为引用的生物学交互作用实例 [8]，比较这三种统计模型。线性二项模型在这些数据中检测到了统计交互作用。对数二项模型和逻辑回归模型则

评估相对风险或优势比的乘法性，未发现统计交互作用的证据。这些模型中不存在统计交互作用，并不意味着不存在生物学交互作用，而是说明效应没有偏离乘法性。

我们最后总结了三种模型在评估效应加和性方面的优势与局限。 尽管估计、检验和解释相对超额风险交互作用（RERI）存在一些复杂之处，它仍常用于流行病学研究，以量化偏离可加性的程度。相比之下，线性二项模型给出的效应估计值易于解释，其中包括交互作用对比。

定义效应加和性

考虑比较暴露于两个“危险因素”X 和 Z 中一个或两个因素与未暴露者的结局 Y 概率或“风险”。那么，p_{xz} 是一个概率，其下标表示在 X 和 Z 的不同“水平”下结局 Y 的概率或风险（表1）。

表1. 优　　两种暴露或危险因素（X和Z）不同水平下结局（Y）的概率

	Z=1 (“暴露于因素 Z”)	Z=0 (“未暴露于因素 Z”)
X=1 (“暴露于因素 X”)	p_{11}	p_{10}
X=0 (“未暴露于因素 X”)	p_{01}	p_{00}

Rothman [2 （第178页）] 指出，以下方程“将可加性确立为生物学独立性的定义”。

$$p_{11} - p_{00} = (p_{10} - p_{00}) + (p_{01} - p_{00}) \qquad \text{(Equation 1)}$$

根据Rothman公式，当两种暴露（X和Z）在生物学上相互独立，且其效应具有可加性时，它们共同且同时作用于Y所产生的效应（$p_{11} - p_{00}$）等于X的单独且独立效应（$p_{10} - p_{00}$）与Z的单独且独立效应（$p_{01} - p_{00}$）之和。当暴露共同且同时作用所产生的效应不同于其单独效应之和时，就存在效应可加性的偏离；Rothman认为，这表明存在生物学交互作用。

可加性也可以等价地定义为效应同质性。重新排列公式1中的各项，可得

$$p_{11} - p_{01} = p_{10} - p_{00}\ , \qquad \text{（公式2）}$$

$$p_{11} - p_{10} = p_{01} - p_{00}。 \qquad \text{（公式3）}$$

无论 Z ＝ 1 ($p_{11} - p_{01}$) 还是 Z ＝ 0($p_{10} - p_{00}$)，方程 2 表明 X 对 Y 的效应相同。效应同质性具有互反性。方程 3 表明，在 X 的所有水平上，Z 对 Y 的效应都相同，也就是说，无论 X=1 ($p_{11} - p_{10}$) 还是 X=0 ($p_{01} - p_{00}$)，效应均相同。当 X 和 Z 的效应具有可加性时，Y 与 X 的关联在 Z 的各个水平上具有同质性，而 Y 与 Z 的关联在 X 的各个水平上具有同质性。

使用概率（交互作用对比）或比值（相对超额风险交互作用）评估效应加和性

偏离效应加和性（或生物学独立性）的情况，无论是定义为联合效应与独立效应之间的不等，还是定义为效应之间的异质性，都可以通过交互作用对比（其各项为概率）和相对超额风险交互作用（其各项为相对风险）进行形式化评估。

对公式（1）中的各项重新排序，可得到交互作用对比 [9]:

$$p_{11} - p_{10} - p_{01} + p_{00} = 0 \qquad \text{（公式4）}$$

对公式4中的各项重新排序，并分别除以 p_{00} ，得到:

$$p_{11}/p_{00} - p_{01}/p_{00} - p_{10}/p_{00} + 1 = 0.$$

注意到这些概率比值即为相对风险（RR），可得:

$$RR_{11} - RR_{01} - RR_{10} + 1 = 0. \qquad \text{（公式5）}$$

Rothman [10] 将方程 5 左侧的量称为“交互作用所致相对超额风险”（RERI）。Rothman和Greenland [9] 则称其为“交互作用对比比值”（ICR）。

Hosmer 和 Lemeshow [11] 将其定义为“同时具有两种暴露的人群中，可归因于两者交互作用的疾病比例”。

方程 1（用于交互对比）和方程 5（用于相对超额风险交互作用）之间的代数等价性，证实了在概率尺度或相对风险尺度上评估效应加和性是合理的。交互对比和相对超额风险交互作用正式检验了以下假设：X 和 Z 对 Y 的效应具有可加性，或者等价地，X 与 Z 之间不存在交互作用。STROBE 声明 [3 (第 825 页)] 展示了如何使用 相对超额风险交互作用评估效应偏离加和性的情况。

数据示例：不同石棉和吸烟暴露工人的肺癌死亡率

Hammond 等人 [8] 比较了 17,800 名石棉工人和 73,763 名未暴露于石棉的工人发生二分结局——肺癌死亡——的风险。他们还记录了吸烟状态，因此参与者呈现出吸烟和石棉暴露的不同组合（表 2）。Hammond 的研究广泛用于流行病学教科书 [2 （第 168-180 页）， 12] 以说明生物学交互作用。

补充文件 1 展示了如何创建一个在特性上与已发表数据高度近似的数据集。为使该数据集的风险概率（以每10万人肺癌死亡数报告）反映已发表数据中的相应概率，我们假设石棉工人和未暴露工人对照组的吸烟率均为 0.28。

表 2. 有石棉暴露和/或吸烟者中的肺癌死亡人数（每 100,000 名工人）

吸烟	石棉暴露	
	石棉工人（n= 17800）	对照组（n=73763）
吸烟者	p_{11}=601.9	p_{10}= 121.1
不吸烟者	p_{01}= 54.6	p_{00}= 11.3

数据示例说明了效应加和性的偏离

如果石棉暴露和吸烟的效应具有相加性，那么同时经历这两种暴露的预期效应将等于两种暴露各自单独效应之和（公式1）。按照表1中引入的符号来定义 p_{xz}，其中 X 表示吸烟（1 = 吸烟者和0 = 不吸烟者），Z 表示石棉暴露（1=暴露和0=未暴露），估计风险概率为:

$\hat{p}_{11} - \hat{p}_{00} = 601.9 - 11.3 = 590.6$ 每10万人死亡数，其中的超额死亡可归因于两种暴露的联合效应。

$\hat{p}_{10} - \hat{p}_{00} = 121.0 - 11.3 = 109.7$ 每10万人中，单独归因于吸烟的超额死亡人数。

$\hat{p}_{01} - \hat{p}_{00} = 54.6 - 11.3 = 43.3$ 每10万人中，单独归因于石棉暴露的超额死亡人数。

可归因于双重暴露的肺癌死亡人数似乎超过了各项暴露单独效应之和。数据示例的交互作用对比：$p_{11} - p_{10} - p_{01} + p_{00}$ 表明，同时经历两种暴露者的肺癌死亡风险，比吸烟或石棉暴露所产生的单独风险之和高出约437.6例每10万人死亡数。针对数据示例计算的RERI，用于量化相对风险尺度上的效应加和性，$RR_{11} - RR_{01} - RR_{10} + 1 = [601.9/11.3] - [54.6/11.3] - [121.0/11.3] + 1 = 38.7$。

线性二项模型直接估计数据示例中的交互作用对比

线性二项模型[4,7] 直接根据概率及概率差异估计交互作用对比：

$$P(Y = 1) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ \qquad \text{（公式6）}$$

回顾一下，X 和 Z 在“暴露”时取值为 1，在“未暴露”时取值为 0，那么

$$\hat{p}_{00} = \beta_0$$

$$\hat{p}_{10} - \hat{p}_{00} = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

$$\hat{p}_{01} - \hat{p}_{00} = (\beta_0 + \beta_2) - \beta_0 = \beta_2$$

$$\hat{p}_{11} - \hat{p}_{00} = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - \beta_0 = \beta_1 + \beta_2 + \beta_3$$

将这些表达式代入定义效应加和性的公式1,

$$p_{11} - p_{00} = (p_{10} - p_{00}) + (p_{01} - p_{00})$$

$$\beta_1 + \beta_2 + \beta_3 = \beta_1 + \beta_2$$

在线性二项模型中，如果与乘积项或交互项相关的回归系数 β_3 等于零，则效应具有可加性。

将这些表达式代入公式4可知，模型对 β_3 的估计值直接估计交互作用对比：

$$p_{11} - p_{10} - p_{01} + p_{00} = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1) - (\beta_0 + \beta_2) + \beta_0 = \beta_3$$

因此，线性二项模型对交互作用对比和 X*Z 交互作用的估计值是等价的。二者都能直接检验可加性；反对 $\beta_3 = 0$ 的假设的证据表明偏离了可加性。

补充文件 2 展示了使用 SAS PROC GENMOD [4,7] 构建线性二项模型的过程。该模型对每 100,000 名工人的死亡人数的点估计值见表 3，与表 2 报告的数值相同。表 3 还报告了模型对回归系数的估计值（及 95% 置信区间）。这些系数包括对不参与任何有中吸

表 3. 石棉暴露和/或吸烟者的肺癌死亡绝对风险（及风险差）（每10万名工人），由线性二项

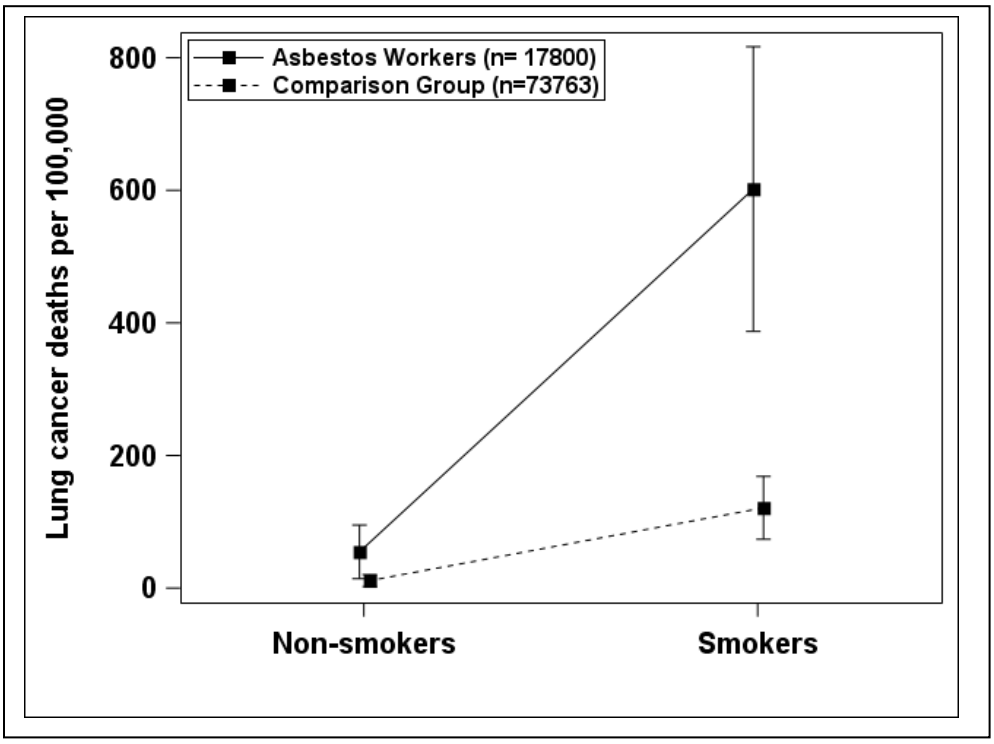
	吸烟	石棉	估计值	每10万人死亡数	估计值的95%置信区间	
					下限	上限
p_{11}	1 (是)	1（是）	0.006019	601.926	387.183	816.669
p_{10}	1 （是）	0（否）	0.001210	121.048	73.627	168.469
p_{01}	0 （否）	1（是）	0.000546	54.619	14.169	95.070
p_{00}	0 （否）	0（否）	0.000113	11.298	2.258	20.337
β_1	吸烟（$\hat{p}_{10} - \hat{p}_{00}$）		0.001098	109.750	61.475	158.025
β_2	石棉（$\hat{p}_{01} - \hat{p}_{00}$）		0.000433	43.322	1.873	84.770
β_3	吸烟*石棉		0.004376	437.557	213.768	661.345
交互对比 p_{11}-p_{10}-p_{01}+p_{00}			0.004376	437.557	213.768	661.345

线性二项模型对以下内容得出相同的推断 β_3，用于估计统计

吸烟与石棉暴露之间的交互作用，以及交互对比（估计值：每10万人死亡数为437.6；95%置信区间：213.8，661.3；P=0.00012702）。这些统计量所产生的P值一致，证实它们对吸烟和石棉暴露的效应具有可加性这一零假设进行了等效检验。

图 1展示了线性二项模型生成的估计值和置信区间，说明了在按石棉暴露划分的组别中，吸烟对肺癌死亡率的效应存在异质性。生成表 3和图 1的语法包含在补充文件 3中。

图 1. 石棉暴露与吸烟之间的生物学交互作用，表现为效应的非可加性或异质性



在数据示例中，二项对数模型和逻辑回归模型均未发现效应乘法性的偏离。

与线性二项模型不同，采用概率的对数变换（对数二项模型）或其对应优势（逻辑回归模型）的模型，可评估效应乘法性的偏离程度。效应乘法性的定义方式类似于效应加和性的定义。如果两个因素（A 和 Z）对某一结局（Y）的效应具有乘法性，则它们的联合

去性时，相对风险将符合以下

关系：$RR_{XZ} = RR_X \times RRz$，优势比将符合以下关系：$OR_{XZ} = OR_X \times ORz$。对数二项模型可估计并检验相对风险的乘法性。

$$\ln[P(Y = 1)] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ,$$

$$P(Y = 1) = \exp(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ).$$

由此可得：$RR_{XZ} = \exp(\beta_1 X + \beta_2 Z + \beta_3 XZ)$; $RR_X = \exp(\beta_1 X)$; $RR_Z = \exp(\beta_2 Z)$.

如果相对风险之间不存在乘法性偏离，则：

$$\frac{\text{相对风险}_{XZ}}{\text{相对风险 z}} = \text{相对风险}_X$$

$$\exp(\beta_1 X + \beta_2 Z + \beta_3 XZ) = \exp(\beta_1 X)\exp(\beta_2 Z) = \exp(\beta_1 X + \beta_2 Z).$$

只有当 β_3（与乘积项 XZ 相关的回归系数）等于零时，这些等式才成立。同样，逻辑回归模型 $\ln[P(Y = 1)\ P(Y / = 0\ ] = \beta_0 + \beta_1) X + \beta_2 Z + \beta_3 XZ$ 用优势或优势比表示的效应乘法性进行评估。在这两个模型中，如果估计值或假设检验表明 β_3 不等于零，则构成效应乘法性偏离的证据。

应用于数据示例时，对数二项模型未发现吸烟与石棉暴露之间存在统计交互作用的证据（P=0.9637），以相对风险衡量时，各因素的

效应呈乘法关系且具有同质性。同样，逻辑回归模型未发现吸烟与石棉暴露之间存在统计交互作用（P=0.9581）。这表明以优势比衡量的效应不存在偏离乘法性的情况。图2和图3展示了对数二项模型和逻辑回归模型生成的估计值。补充文件 4 详细介绍了使用

图 2. 对数二项模型中的预测对数概率表明，效应乘法性没有偏离。

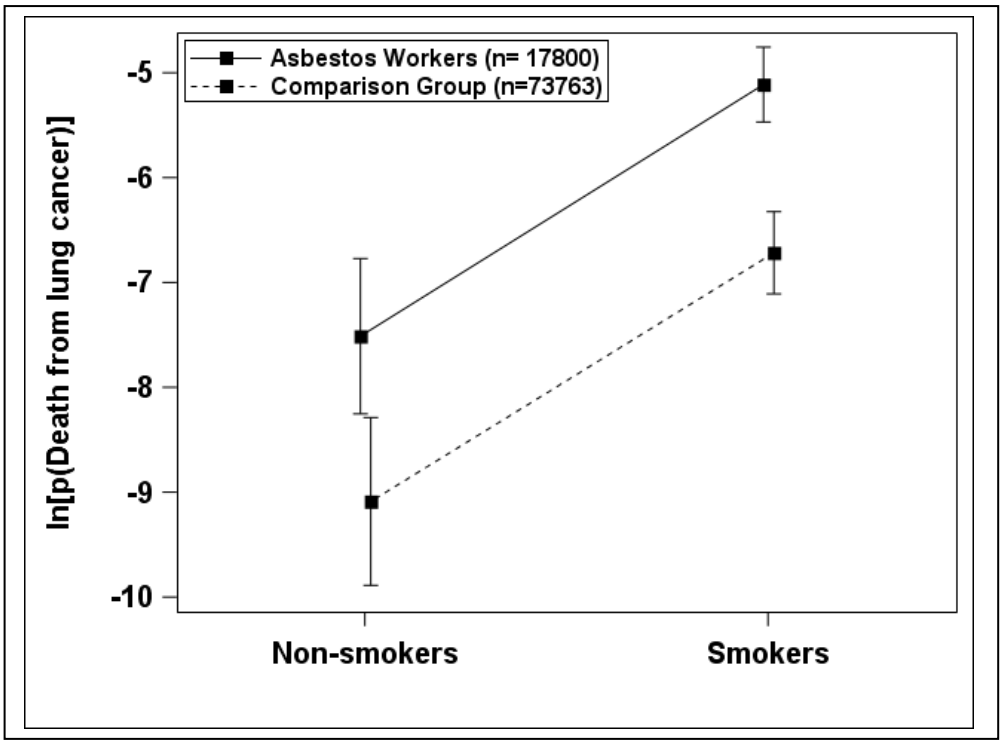
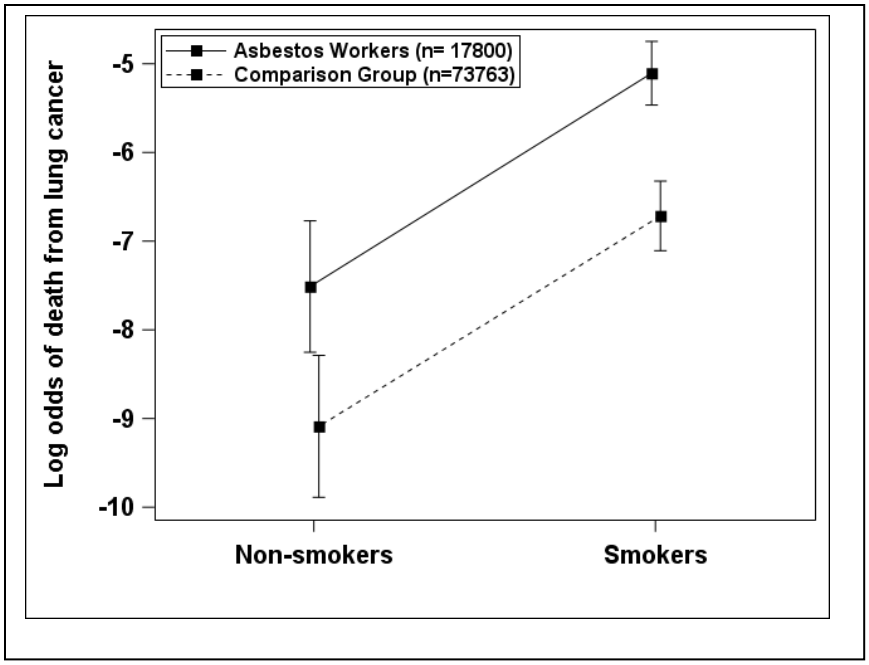


图 3. 预测对数优势表明，效应乘法性没有偏离

逻辑回归模型。



这些模型在检测统计交互作用方面的差异，并不会混淆数据是否体现生物学交互作用这一问题。相反，这些差异说明了以下两点的重要性：(1) 确定一种能够反映假设交互作用形式的效应度量（即概率或风险之间的差值或比值）；(2) 构建一个能够直接估计该效应

在统计模型中进行选择

选择能够生成相对超额风险交互作用估计值的对数二项回归模型或逻辑回归模型

在该数据示例中，对数二项模型和逻辑回归模型均未检测到统计交互作用。这些模型的形式适合检测效应乘法性的偏离

。尽管如此，在流行病学中，人们仍广泛使用这些模型，通过相对超额风险交互作用（RERI）等比值指标评估效应加和性的偏离。[3]。

尽管相对超额风险交互作用（RERI）得到广泛应用，但它也有一些缺点。由于 RERI 是由比值构成的，因此无法将其解释为同时归因于吸烟和石棉暴露的超额死亡人数。对于数据示例，计算得到的 RERI 为 38.7；与线性二项模型估计的交互对比（IC）——每 [illegible] 如何获得标准误有关，因为需要利用标准误构建 RERI 的置信区间，或检验与 RERI 相关的假设。Hosmer 和 Lemeshow [11], 提出了一种颇具影响力的方法，利用逻辑回归估计 [illegible] Andersson 等人 [13] 以及 Richardson 和 Kaufman[14], 提供了该方法的 SAS 语法；后者使用 SAS PROC NLMIXED 构建“线性优势比模型”。作为另一种方法，Richardson 和 Kaufmann [14] 建议采用自助法计算置信区间。对于数据示例，基于 500 个自助样本计算得到的 RERI 经验 95% 置信区间为 15.9 至 132.6。然而，由于 RERI 置信区间的上下限也是比值，因此在解释时面临与 RERI 估计值相同的困难。

选择直接估计交互作用对比的线性二项模型

逻辑回归广泛应用于流行病学研究二项结局，尽管其形式适合检测效应乘法性的偏离。该模型广受欢迎且经久不衰，一个重要原因是它采用了 logit 连接函数——二项响应的典则连接函数——因而具有理想的统计性质。其中一项优势是逻辑回归能够可靠地收敛并得到参数估计值。采用其他连接函数的模型可能会遇到收敛问题。Zou [15] 以及 Spiegelman 和 Herzmark [4] 讨论了对数二项模型的收敛问题，并建议在出现此类问题时采用修正泊松模型来解决。

使用非典型恒等连接函数的线性二项模型，在估计值上也可能出现不收敛。这一局限妨碍了该模型得到更广泛的认可，尽管它能够通过估计交互作用对比，直接评估效应加和性。为解决线性二项模型中的不收敛问题，Spiegelman 和 Herzmark [4] 主张对模型进行修正：保留恒等连接函数，但假定结局服从泊松分布。尽管这种方法能够收敛，但施加泊松假定会导致模型错误设定二项结局的方差。这种有意对结局假定分布进行的错误设定，会降低模型标准误以及基于这些标准误的假设检验和置信区间的效率。因此，Spiegelman 和 Herzmark [4] 建议计算即使在模型错误设定时仍保持稳健的标准误。Richardson 等人 [7] 也建议计算稳健标准误，但由于他们将其应用于加权数据，因此不主张对线性二项模型进行其他修正。补充文件 2展示了如何使用 SAS PROC GENMOD 纳入这些不同的建议。

Cheung [5] 通过提出同样使用恒等连接函数的修正最小二乘法（MLS）模型，解决线性二项模型中的不收敛问题。Cheung 的方法还计算稳健标准误。Cheung 的方法与之不同之处在于，它使用普通最小二乘法（OLS），而不是最大似然估计（MLE）。这样做避免了对结局的假定分布进行设定。这一策略解决了不收敛问题，但无法保证估计概率处于 0 到 1 的合理范围内。

结论

生物学交互作用通常被假设表现为效应的非加和性，而效应是以风险或概率之差进行量化的。将其应用于流行病学教学中广泛用于阐释生物学交互作用的数据示例时，线性二项模型能够检测到统计交互作用，而逻辑模型和对数二项模型则无法检测到。

这一结果印证了共识：通常应将生物学交互作用评估为偏离效应加和性的程度。RERI 等统计量广泛用于流行病学中，以评估相对风险尺度上的可加性。相比之下，线性二项模型得出的差概率差的估计值，包括交互作用对比，这些估计值可直接解释为超额风险。

广泛可用的广义线性模型软件使研究人员能够构建线性二项模型，并获得交互作用对比及其他效应的估计值和置信区间。该模型值得在研究中得到更广泛的应用，也值得在流行病学教学中更频繁使用。线性二项模型可能会遇到收敛问题，但已有策略可以应对这一局限。

资助

赵博士的部分工作获得了美国国立卫生研究院下属美国国立普通医学科学研究所提供的经费支持[资助

致谢：

作者感谢 Tabitha Garwe博士对稿件提出的重要意见。

利益冲突：无。

参考文献：

[1] VanderWeele TJ，Knol MJ。交互作用教程。《流行病学方法》2014；3：33-72。

[2] Rothman KJ。《流行病学：导论》。纽约：牛津大学出版社，2002年 .

[3] Vandenbroucke JP，von Elm E，Altman DG，等。加强流行病学观察性研究的 。《流行病学》2007年；18(6)：805-835。

[4] Spiegelman D，Hertzmark E。使用 SAS 轻松计算风险比或患病率比及其差值。《美国流行病学杂志》2005；162(3)：199-200。

[5] Cheung YB。用于估计风险差的修正最小二乘回归方法。《美国流行病学杂志》2007年；166(11)：1337-44。

[6] Bieler GS, Brown GG, Williams RL, Brogan DJ. 从复杂抽样调查数据估计模型调整风险、风险差

[7] Richardson DB，Kinlaw AC，MacLehose RF，Cole SR。用于风险比或患病率比及其差异的标准化二项模型。国际流行病学杂志 2015年；**44**（5）：1660-72。

[8] Hammond EC、Selikoff IJ、Seidman H。石棉暴露、吸烟与死亡率。《纽约科学院年鉴》1979年；

[9] Rothman KJ，Greenland S。《现代流行病学》。费城：Lippincott Williams and Wilkins,

[10] Rothman KJ。《现代流行病学》（第1版）。波士顿：Little, Brown and Company，1986 .

[11] Hosmer DW，Lemeshow S。交互作用的置信区间估计。《流行病学》1992；3(5):
452-456。

[12] Szklo M，Nieto FJ。《流行病学：超越基础》（第3版）。马萨诸塞州萨德伯里：Jones and
Bartlett，2004。

[13] Andersson T，Alfredsson L，Källberg H，Zdravkovic S，Ahlbom，A。生物学交互作用测量指

[14] Richardson DB, Kaufman JS. 交互作用所致相对超额风险及其相关置信界限的估计，《美国流
0.

[15] Zou G. 针对二元数据前瞻性研究的修正泊松回归方法。《美国流行病学杂志》
2004;159(7):702-706.

联系信息

David M. Thompson，俄克拉何马大学健康科学中心生物统计学与流行病学系，俄克拉何马城，
OK 73104（电子邮件：dave-thompson@ouhsc.edu）。

Yan Daniel Zhao，俄克拉何马大学健康科学中心生物统计学与流行病学系，俄克拉何马城，
OK 73104（电子邮件：daniel-zhao@ouhsc.edu）。

补充文件 1. 数据示例

用于创建一个近似于 Hammond 等人发表的数据集的 SAS 语法 [8]。结局为每10万名工人中的肺癌死亡人数。假定石棉工人和未暴露工人对照组的吸烟率均为 0.28。

```
/*数据示例*/ data one; array asbn (2) (73763 17800); /*已发表研究中的 n 值*/ array rate
2.6 58.4 601.6); /*已发表研究中每 10 万人的肺癌死亡人数*/ smokeprev=0.28; /*假定
gcadeath=0 to 1; mult=rate [2
)) *
d; end; keep asbestos
; run; proc sort data=one (keep=asbestos smk lungcadeath count)
ending smk descending lungcadeath; run; /*包含个体
put; end; run;
```

```
proc 格式;    value smkf 1="吸烟者" 0="不吸烟者";    value gpf    1="石棉
=73763) ";    value death
原因"; run;    /* 表 2. 接触
人) wi
```

```
th 暴露
接触石棉和/或吸烟*/ proc freq data=two order=data;    weight count;
h / nocol nopct outpct out=three;
```

```
格式 smk smkf. 石棉 gpf. lungcadeath death.; run;
```

```
数据 four;    集 three;
              ct row*1000;
```

proc report nowd data=four; where lungcadeath=1; columns smk asbestos,
roup "吸烟" format=smkf. order=data; define asbestos /
efine perhunthou / analysis " format 6.2;

补充文件 2. 线性二项模型的 SAS 语法

以下语法展示了如何使用 SAS PROC GENMOD 构建线性二项模型 [4,7] 。MODEL语句指定了纳入线性预测变量的自变量：smk（吸烟状态）、asbestos（石棉暴露状态）以及 smk*asbestos（吸烟状态与石棉暴露之间的交互作用）。MODEL语句中的选项指定结局（肺癌）服从二项分布，并通过恒等连接函数将其直接关联到线性预测变量。LSMEANS语句估计每种暴露组合对应的每 100,000 名工人的死

MATE语句列出定义交互作用对比（交互对比，IC）的四个系数（1 −1 −1 1）：

$$(1)\, p_{11} + (-1)p_{10} + (-1)p_{01} + (1)p_{00} = 0$$

```
/*线性二项模型*/
proc genmod data=long descending; class smk (ref=first) 石棉 (ref=first); model lungcadeath =  smk 石棉 smk*石
                                                                                  / cl; ods output lsmeans=lsmeans estimates=估计值
                                                                                  交互对比" smk*石棉 1 -1 -1 1; run;
```

```
/*包含 REPEATED语句的语法，该语句启动广义估计方程估计稳健标准误，这是 Richardson 等人
                                                                                  escending;
class smk (ref=first) 石棉 (ref=first) id; model
                                                                                  d; lsmeans smk*石
                                                                                  1 -1 -1 1;
run;
/*Spiegelman 和 Herzmark 倡导的线性二项模型修改版，适用于收敛失败的情况*/
proc genmod data=long descending; class smk (ref=first) 石棉 (ref=
                                                                                  ntity
                                                                                  s
                                                                                  棉 1 -1 -1 1;
run;
```

补充文件 3. 表 3 和图 1 的 SAS 语法

以下语法使用线性二项模型（补充框 S2）输出的数据集，生成表 3 和图 1。

```
/* 表 3. 采用线性二项模型估算的石棉暴露和/或吸烟者的肺癌死亡绝对风险（及风险差）（每 100,000

数据 死亡率; 设置 最小二乘均值; 死
                    00; ucl=上限
                        )00;

run;
过程 打印 无观测值 数据=死亡率; 变量 吸烟 石棉 估计值 死

/*交互作用对比（交互对比，IC）的估计值*/
数据 交互对比;
    设置 估计值; 交互对比=
                                    =

run;
过程 打印 无观测 数据=交互对比; 变量 标签 meanestimate 交互对比 交互对比_

run;

/*回归系数的估计值，可解释为超额死亡人数*/

数据 beta2;
    设置 betas (其中=(df=1)); 超额死亡人数=

run;
过程 打印 无观测 数据=beta2; 变量 parameter estimate excessdeaths

/* 图 1. 石棉暴露与吸烟之间的生物学交互作用，以效应的非加和性或异质性为例说明*/

过程 模板; 定义 样式 styles.mystyle; 父级=
                                    ; 样式
                                    =1;
                                色 线
```

ods html 样式=styles.mystyle；过程 sgplot 数据=死亡率；序列 y=死亡率 x=吸烟 / 分组=石棉 名
聚类 聚
类宽度 . 标记 标记属性（符号方形填充 大小=10）： 高低 x=吸烟 高=ucl 低=lcl / 分组=石棉 分组
聚类 聚类宽度 . 类型线
= = =1 =
条 线条属性（图案） 低端帽须线 高端帽 须线； x轴 值 （0 1） 标签 " " 值属性 （大小 14 权重
属性=（大
e" / 标题="" 位置=内部 向下=2 位
html 关闭；

补充文件 4. 对数 二项模型和逻辑回归模型

图 2 中对数二项模型及其估计值的呈现。

proc genmod data=long descending; class smk (ref=first) 石

ods html style=styles.mystyle; proc sgplot data=lsmeans; series y=estimate x=smk / group=

cluster clusterwidth . 标记

石棉 groupdisplay cluster

s=(0 1)

'ln [p(肺癌死亡)]" labelattrs=(size

'1); format smk smkf. 石棉 gpf.;

down=2 position=topleft valueattrs=(size=12

运行; ods html

图 3 中逻辑回归模型及其估计值的展示。

proc genmod data=long descending; class smk (ref=first) 石棉

=

ods html style=styles.mystyle; proc sgplot data=lsmeans; series y=estimate x=smk / group=

cluster clusterwidth .

group=asbestos groupdisplay

二的对数优势" labelattrs=

old); format smk smkf. asbestos gpf.;

```
run;
ods html close;
```