标题：多版本处理理论下病原体菌株特异性效应的潜在变异

作者：Bronner P. Gonçalves[1]

所属机构：
[1] 健康与医学科学学院，萨里大学，吉尔福德，英国

通信作者：
健康与医学科学学院，萨里大学，30 Priestley Rd，吉尔福德 GU2 7YH，英国，
bronnergoncalves@gmail.com

利益冲突声明：作者声明不存在任何利益冲突。

摘要

制定感染相关的循证政策需要估计感染对健康的影响。然而，病原体变异可能使许多传染性病原体的研究变得复杂；所谓病原体变异，是指不良结局的发生取决于感染菌株。本文探讨在菌株特异性效应存在异质性且无法获得菌株组成信息时，如何解读感染对健康影响的流行病学研究。我们运用潜在结局和因果推断理论，分析存在多种治疗版本的情况，并指出，此类研究中常报告的量具有因果解释，而这种解释取决于感染菌株在人群中的频率。此外，与其他可能违反治疗变异无关性假设的情形一样，可迁移性还需要考虑非复合暴露情形所不需要的其他因素。本文讨论菌株特异性效应可能存在的异质性，有助于解读此类研究；基于上述原因，也凸显了病原体亚型数据的价值。

关键词：因果推断、一致性、潜在结局、传染病

多版本治疗理论下病原体菌株特异性效应的潜在变异

背景

对于许多病原体，感染对健康的影响可能因感染菌株而异。在流行病学中，菌株特异性效应存在差异的例子不胜枚举：有证据表明，感染不同SARS-CoV-2变异株后发生急性后期后遗症的风险存在异质性（1）；定植某些*B*组链球菌血清型与侵袭性感染风险较高有关（2）；疟疾的严重表现可能与毒力较强的寄生虫虫株有关（3）。

在此，我们考虑这样一种可能较为常见的情形：菌株特异性效应存在变异，但无法直接观测；也就是说，可以获得病原体是否存在的数据，但没有菌株信息（例如，因为检测试验无法区分不同菌株）。在讨论此类研究时，我们采用因果概念，旨在对这一具有菌株特异性效应潜在变异的情形作出连贯解释。具体而言，我们考察针对多种治疗版本开展因果分析时所形成的思想（4-6）；我们认为，这些思想有助于理解此类涉及复合暴露的研究，即版本变异无关性（相对于某一特定结局）不成立的暴露（当暴露变异无关性成立时，该暴露不是复合暴露）。

在继续讨论之前，我们先作两点说明。第一，尽管有多个含义不同的术语用于描述病原体亚型（变异株、血清型、基因型、克隆），但在全文中，我们会较为宽泛地使用“菌株”一词来指代它们（关于病原体变异可能分类方式的讨论，见（7）的引言）。还需注意的是，本文所考察的问题不同于不同菌株可能具有不同疫苗效应这一问题（7-9）。

记号

大多数因果推断分析都会采用一致性条件（详细讨论见（10）第 3 章），该条件将个体的观测结局与其实际接受的暴露水平下的潜在结局联系起来。现在，如果暴露

对应于感染某种特定病原体，且如果某些病原体菌株比其他菌株毒力更强，那么暴露不仅会有多个版本，而且由于毒力存在差异，暴露变异无关性也不成立。换句话说，如果我们将不同菌株引起的感染视为感染该病原体的不同方式，那么在致病性存在差异的情况下，一致性条件便不适用。

为讨论这一问题，我们考虑一项观察性研究，其中 A暴露表示通过一种无法区分菌株的方法检测到的病原体感染（1 = 有感染和0 = 无感染），而 Y 表示结局（例如，1 = 住院和0 = 随访期间未住院）。此外，令 K 表示病原体暴露的版本：无感染、$K = 0$，或感染菌株；在后一种情况下，K 取值于可能菌株的集合 Φ。因此，A 可视为 K 的粗化版本（4）。我们还将 Y^k定义为：若将 K 设为 k，则 Y 的潜在结局（11）。

在暴露变异会影响潜在结局的情境下，解释估计目标需要额外考量；VanderWeele 和 Hernán（4）对此进行了详细讨论。具体而言，在该研究中，他们利用涉及对 A 和 K进行干预的潜在结局，将暴露变异无关性定义为：对于所有个体、所有 a，以及代表暴露水平 a可能版本的所有 k 和 k'，均有 $Y^{a,\ k} = Y^{a,\ k'}$。在此，我们假设这一条件不成立，并将（4）中的框架应用于此情境。

值得注意的是，如果可能存在合并感染，则可以将其视为额外的暴露版本；Φ 这将包括感染每种菌株、双菌株合并感染、三菌株合并感染等。最后需要注意的是，虽然我们重点关注根据菌株信息定义的感染版本，但相关的暴露版本也可能与感染的其他方面相关，例如持续时间；举例来说，在一个简单的情景中，可能存在短期感染（相对于预先指定的持续时间而言）和长期感染。在实践中，感染持续时间信息可能难以收集，因为这需要频繁随访。

常被报告的经验量：可能的因果结构及其解释

图1的图A展示了一个因果图，表示上述变量之间可能存在的关系。如图所示，我们假设版本（K）先于暴露（A）（6），且给定K后，A 与Y d-分离；例如，这要求所有影响Y 的菌株都包含在Φ中。请注意，该图包含K 和Y的一个共同原因，以L 表示（例如，年龄可能影响病原体暴露和结局）；如上所述，A可视为K的粗化版本，因此我们仅展示版本-结局混杂因素L （另见（6）中的图5）。不过请注意，如果A与K 之间的关系不是确定性的，就可能存在处理-版本混杂因素，需要在图中表示出来。此外，由于图1表明只有K 会影响Y，因此下文将潜在结局定义为相对于K、 Y^k而言。

给定可交换性假设$Y^k \perp\!\!\!\perp K|L$，在存在多种感染版本的这一设定中，常被报告的量 $\sum_l E[Y|A=1,L=l]\Pr(L=l) - \sum_l E[Y|A=0,L=l]\Pr(L=l)$无需$K$的信息，可被解释（4, 12）为一项试验中将估计的比较：在由L定义的分层内，对每个臂，暴露版本（即菌株类型或无感染）根据各版本按A 和L 水平的相对频率被随机分配（见（4）中的命题8）：

$$\sum_l E[Y|A=1,L=l]\Pr(L=l) - \sum_l E[Y|A=0,L=l]\Pr(L=l)$$

$$= \sum_{l,k} E[Y^k|L=l]\Pr(K=k|A=1,L=l)\Pr(L=l) - \sum_{l,k} E[Y^k|L=l]\Pr(K=k|A=0,L=l)\Pr(L=l)$$

如果我们假设无感染只有一个版本，则该表达式可以写为：

$$\sum_l E[Y|A=1,L=l]\Pr(L=l) - \sum_l E[Y|A=0,L=l]\Pr(L=l)$$

$$= \sum_{l,k} E[Y^k|L=l]\Pr(K=k|A=1,L=l)\Pr(L=l) - \sum_l E[Y^0|L=l]\Pr(L=l)$$

$$(\because \Pr(K=0|A=0,L=l)=1)$$

$$=\sum_{l,k}\{E[Y^k|L=l]-E[Y^0|L=l]\}\Pr(K=k|A=1,L=l)\Pr(L=l)$$

因此，在 L 的各个水平内，该表达式对应于菌株特异性效应的平均值，并按由 $A=1$ and $L=l$ 定义的分层中各菌株的频率加权。举例来说，如果 Y 表示研究期间的住院，A 表示通过无法提供变异株信息的检测所发现的 SARS-CoV-2感染（K），而不同变异株对 Y 的影响可能不同（13）；L 表示年龄，那么，如果 L 足以控制版本-结局混杂，则上述经验量可以视为来自一项随机研究的效应估计：在给定 L 的条件下，被分配至感染组的参与者会以一定概率被分配至某一特定变异株，该概率由感染者中变异株的人群层面分布决定；随后将该组与对应于无感染的组进行比较。

在**图1**的图B和图C中，我们展示了该人群中存在三种流行菌株的病原体的两种假设情景。菌株1和菌株3的感染会对Y产生影响，但影响程度不同；菌株2则不会影响Y。图中还显示了感染者中的菌株组成。如图所示，尽管假设两种情景中的菌株特异性影响相似，但当菌株2的相对频率较低时，$\sum_l E[Y|A=1,L=l]\Pr(L=l)$ $-\sum_l E[Y|A=0,L=l]$ $L=l$ Pr( )的数值更高。

讨论

以上表达式为具有潜在异质性菌株特异性效应的研究提供了一种与因果推断理论相一致的解释。具体而言，若假设无感染只有一个版本，那么 $\sum_l E[Y|A=1,L=l]\Pr(L=l)$ $-\sum_l E[Y|A=0,L=l]$ $L=l$ Pr( ) 可被视为以感染者中 L 的分布及菌株组成为权重，对 l 所定义分层中的 l 层级菌株特异性效应（相对于无感染）取平均。此外，与（6）类似，作者指出，将在一个人群中估计的复合暴露效应应用于另一个人群较为复杂，因为这取决于暴露版本分布，即使菌株特异性效应，

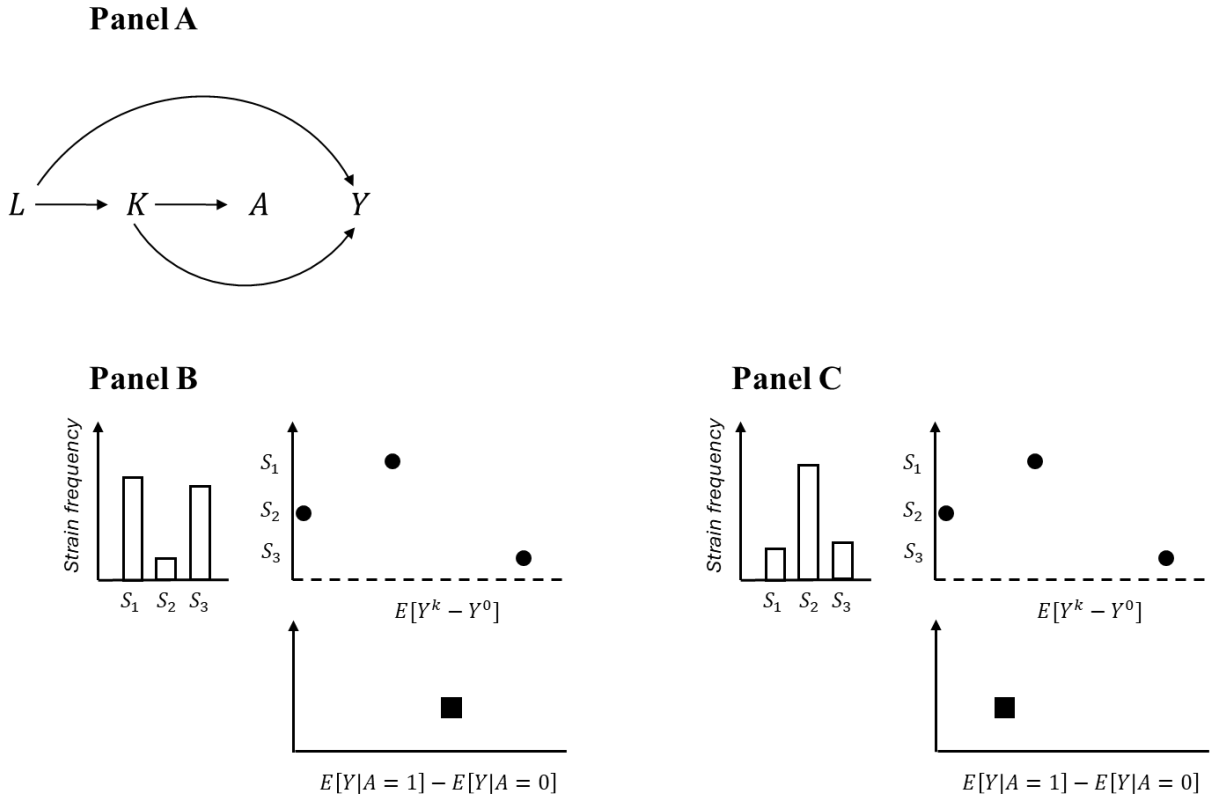
$E[Y^k - Y^0|L = l]$，在两个人群之间具有可迁移性，$\sum_l E[Y|A = 1, L = l]$ $L =$Pr(l) $-\sum_l E[Y|A = 0, L = l]$ $L = l$ Pr( ) 仍可能不同，除非在 L 的各个水平内，两个人群中的版本 k 分布相似。因此，当目标是将平均感染效应迁移至不同人群，或估计菌株特异性效应时，就需要菌株组成数据。值得注意的是，如果不存在可能导致不同结局的多种感染版本，那么 A 将属于非复合暴露；在一致性和条件可交换性成立的情况下，$\sum_l E[Y|A = 1, L = l]$ $L = l$ $-\sum$ $E[Y|A = 0, L = l]$Pr( ) Pr($_l$ $L = l$) 就可解释为 A 对 Y 的效应。

我们讨论的另一项启示是，随着菌株组成发生时间变化（例如，在 SARS-CoV-2变异株出现后的时期），无论人群层面免疫力是否发生变化，上文讨论的常见报告指标都可能随时间变化，因为对于某些 k，Pr($K = k|A = 1, L = l$) 会发生变化。

与（4）中的论点相关，需要注意的是，治疗无多版本假设是否被违反取决于所考察的结局。例如，对于“随访期间发热”这一结局，感染下的潜在结局可能不受菌株影响；然而，如果结局是住院，那么病原体暴露差异可能与潜在结局相关，因此上文的讨论也适用。

总体而言，针对存在多种治疗版本的研究所采用的因果推断方法，为对菌株特异性效应存在潜在异质性的分析提供了一种连贯的解释。特别是，该方法表明，在存在多种治疗版本的情况下，可以在稍有不同的解释下获得研究人群中的平均效应；这种解释承认版本在人群层面的分布，但不需要菌株数据。

图1。存在多种病原体暴露版本的流行病学情境。图A与（12）中的图1相似，展示了一张因果图，其中 K 表示感染菌株（或者对于 $K = 0$，表示不存在任何病原体菌株），A 表示总体感染状态，L表示混杂因素，而 Y表示感兴趣的结局。图B和图C展示了菌株的人群层面分布不同（即 $\Pr(K = k|A = 1)$）对数量$E[Y|A = 1] - E[Y|A = 0]$ 的影响（对于图B和图C，我们假设不存在混杂因素，因此这些项未进行分层）。在这些图中，我们假设存在三种菌株，分别用 S_1、 S_2和 S_3表示。最后请注意，（12）中图4B所示的因果结构是在心理社会构念的背景下提出的，与此处讨论的情境有相似之处。



参考文献

1. Xie Y, Choi T, Al-Aly Z. 德尔塔变异株出现前、德尔塔变异株和奥密克戎变异株时期的SARS-CoV-2感染急性后期后遗症。《新英格兰医学杂志》. 2024年;391(6):515-25. 2. Seedat F, Brown CS, Stinton C, Patterson J, Geppert J, Free man K, et al. 与早发型B组链球菌感染相关的细菌载量和分子标志物：系统综述与荟萃分析。《儿科传染病杂志》. 2018年;37(12):e306-e14. 3. Gupta S, Hill AV, Kwiatkowski D, Greenwood AM, Greenwood BM, Day KP. 恶性疟原虫疟疾中的寄生虫毒力与疾病模式。《美国国家科学院院刊》. 1994年;91(9):3715-9. 4. VanderWeele TJ, Hernan MA. 多种治疗版本下的因果推断。《因果推断杂志》. 2013年;1(1):1-20. 5. Vanderweele TJ. 多种中介变量版本的中介分析。《流行病学》. 2012年;23(3):454-63. 6. Hernan MA, VanderWeele TJ. 复合治疗与因果推断的可迁移性。《流行病学》. 2011年;22(3):368-77. 7. Gilbert P, Self S, Rao M, Naficy A, Clemens J. 筛查分析：利用疫苗试验数据评估疫苗效力如何随病原体基因型和表型变异而变化的方法。《临床流行病学杂志》. 2001年;54(1):68-85. 8. Perenyi G, Stensrud M. 变异株特异性治疗效应及其在疫苗研究中的应用。《生物统计学》. 2025年;81(2). 9. Gilbert PB, Self SG, Ashby MA. 评估针对不同人类免疫缺陷病毒型别的差异性疫苗保护的统计方法。《生物统计学》. 1998年;54(3):799-814. 10. Hernán MA, Robins JM. 《因果推断：假如》。2025年7月10日版本：博卡拉顿：Chapman & Hall/CRC ； 2020 年。 11. Little RJ, Rubin DB. 通过潜在结局研究临床和流行病学研究中的因果效应：概念与分析方法。《公共卫生年度评论》. 2000年;21:121-45. 12. VanderWeele TJ. 构造测量与因果推断：迈向心理社会构念的新测量模型。《流行病学》. 2022年;33(1):141-51. 13. Bager P, Wohlfahrt J, Bhatt S, Stegger M, Legarth R, Moller CH, et al. 丹麦感染SARS-CoV-2奥密克戎变异株与德尔塔变异株相关的住院风险：一项观察性队列研究。《柳叶刀·传染病》. 2022年;22(7):967-76.