arXiv:2301.01823v1 [stat.ME] 4 Jan 2023

癌症流行病学中的个体脆弱性超额风险模型

预印本

Francisco Javier Rubio
统计科学系
伦敦大学学院
伦敦 , UK
f.j.rubio@ucl.ac.uk

Hein Putter
生物医学数据科学系
莱顿大学医学中心
莱顿 , 荷兰
h.putter@lumc.nl

Aurelien Belot癌症结局不平等网络 非传染性疾病
流行病学系 伦敦，英国
aurelien.belot@lshtm.ac.uk

2023年1月6日

摘要

未观测个体异质性是人群癌症生存研究中常见的挑战。这种异质性通常与模型设定错误以及未能记录真正相关的变量有关。我们在超额风险模型这一癌症流行病学的主要工具之一的背景下，研究未观测个体异质性的影响。我们提出一种个体超额风险脆弱性模型，以考虑个体异质性。这是将脆弱性建模拓展至相对生存框架。为便于推断所提模型的参数，我们选取能够产生边际风险和生存函数闭式表达的脆弱性分布。由此得到的模型具有直观的解释：脆弱性会使存活者中的健康个体得到选择。我们采用具有一般风险结构的灵活参数模型来建模超额风险，从而便于纳入时间依赖效应。我们通过一项模拟研究展示所提方法的性能。我们使用英格兰确诊的肺癌患者数据给出一个真实数据示例，并讨论未考虑未观测异质性对净生存估计的影响。该方法已在 R 包 IFNS 中实现。

关键词 超额风险；灵活；脆弱性；一般风险；净生存。

1 引言

癌症流行病学已成为许多国家和卫生组织的重点领域之一。该领域关注的结局包括癌症确诊后的死亡时间，以及癌症所致的生存。后者通常被视为癌症管理质量的替代指标。事件时间有多种分析方法。简而言之，总生存框架旨在研究与所有死因相关的生存函数；竞争风险框架则旨在分析死亡可能归因于不同死因时的生存时间。已知死因时，死因特异框架可用于研究死因特异性指标；而在死因不详或不可靠时（例如基于人群的癌症登记数据），相对生存框架可用于研究与癌症相关的风险。相对生存框架的核心思想是，将与癌症相关的风险和与其他死因相关的风险区分开来；这一方法无需知道死因，而是利用生命表中的预期死亡风险

数据。癌症流行病学中通常不采用总生存框架，因为这一指标并不能代表癌症相关的生存。此外，如果不同总体受到的其他死因相关风险死亡率差异很大，患者的生存可能会受到影响，从而使不同总体的总生存函数难以比较。尽管死因特异框架是分析癌症生存的有用方法，但大多数国家无法获得死因信息（例如死亡证明上的信息），或相关信息并不可靠，因此该框架在基于人群的癌症流行病学中的应用受到限制。基于这一原因，相对生存框架已成为国际比较癌症生存最常用的工具 [5]。相对生存框架假定，个体风险函数可分解为与其他死因相关的风险和与癌症相关的“超额”风险：

$$h(t;\mathbf{x}) = h_P(\text{age} + t; \text{year} + t, \mathbf{z}) + h_E(t;\mathbf{x}), \tag{1}$$

其中，$t > 0$ 是自诊断起测量的事件发生时间（通常以年为单位），h_P（年龄 $+t$；年份 $+t$，$\mathbf{z}$）是根据特征 $\mathbf{z}$对应的生命表得出的人群预期死亡风险，“年龄”表示癌症诊断时的年龄，“年份”表示诊断年份（因此，年龄 $+t$ 是时间 t时的年龄，年份 $+t$ 是时间 t时对应的年份）。人群预期死亡风险 h_P（年龄$+t$；年份$+t$，$\mathbf{z}$）被视为已知，而与癌症相关的超额风险 h_E（t；$\mathbf{x}$）是我们关注的量，并根据可用的患者特征 $\mathbf{x}$进行估计。
通常，$\mathbf{z} \subseteq \mathbf{x}$。另一个相关的关注量是净生存率，$S_N$（$t$；$\mathbf{x}$）$= \exp\{-\int_0^t h_E$（$r$；$\mathbf{x}$）$dr\}$，它是特征为 $\mathbf{x}$ 的患者所对应的超额风险函数 h_E（t；$\mathbf{x}$）的生存函数（关于这些问题的讨论，见 Perme 等人[28] 和 Rubio 等人[35]；癌症生存分析中使用的其他指标概述，见 Belot 等人[11] ）。癌症管理的政策制定和国际比较通常依据整个总体或关注亚组的净生存率，该净生存率按边际净生存率或平均净生存率计算

$$S_N(t) = \int S_N(t;\mathbf{x})\mathrm{d}\varphi(\mathbf{x}),$$

其中，φ 是协变量$\mathbf{x}$的分布函数。在人群研究中，对于感兴趣的人群，其协变量为$\mathbf{x}_1$，.....，$\mathbf{x}_m$，该量通常计算为

$$S_N(t) = \sum_{i=1}^{m} S_N(t;\mathbf{x}_i).$$

文献中已提出多种参数 [13, 16, 17, 25, 29, 34], 半参数 [36], 和非参数 [28] 方法，用于估计超额风险和净生存。

估计超额风险函数的挑战之一是，在人群层面可能无法获得所有相关协变量。未观测个体异质性（UIH）是指重要协变量未被记录或在样本中不可获得的情况，并且可能与模型设定错误相结合。不考虑 UIH 的影响可能十分显著。例如，在总生存分析中，Aalen[3] 和 Aalen 等人[4] 表明，忽略 UIH 会导致参数估计产生偏倚，并影响某些流行病学指标（如风险函数和发病率）的解释。由模型设定错误 [21] 或缺失协变量 [24] 产生的 UIH 也会观察到类似影响。此外，UIH 还会影响模型选择[32] 和因果效应的估计 [37], ，这凸显了对其加以考虑的重要性。

个体脆弱性模型是处理未观测异质性 [1, 8, 9, 15, 20, 23, 40, 41] 的一种易于实现的方法。脆弱性模型的基本思想是，将个体风险函数乘以一个服从参数分布的随机校正因子 G（关于这一点的更多详情，请参见第 2 节）。这样便可以处理偏离原始模型（不含脆弱性项）的非特异性情况，因为这些偏离可能是模型设定错误或缺少相关协变量所致。在相对生存框架下，Zahl [42] 研究了相关脆弱性模型的应用；这类模型包含两个随机脆弱性，分别影响与其他死因相关的风险函数和超额风险。然而，他们遇到了一些推断问题，因为其提出的方法增加了五个参数，除非对参数空间施加任意限制，否则这些参数无法同时估计。事实上，对两种竞争风险（分别与其他死因和癌症相关）进行联合建模是一项具有挑战性的问题，可能会导致不可识别模型，[38] 例如 Zahl [42]提出的模型。
Goeman 等人[19] 研究了在并发疾病风险模型和超额风险模型中结合使用脆弱性的方法。也就是说，他们假设将超额风险进一步分解为与并发疾病相关的风险（假设已知）和与癌症相关的风险。他们假设脆弱性项对并发疾病风险和超额风险的影响相同，并专注于使用具有分段基线风险且不含时间依赖效应的比例风险模型对超额风险进行建模。值得注意的是，假设采用简单模型（例如不含时间依赖效应）可能会导致脆弱性项的方差膨胀，因为该项会捕捉由缺失协变量或模型设定错误导致的未观测异质性。Rancoita 等人[30] 研究了共享脆弱性模型的应用；这类模型假设随机脆弱性同时影响与其他死因相关的风险和超额风险。Rubio 等人[35]研究了这种情况

即对与其他死因相关的风险进行估计时可能存在偏差，这是使用分层不足的生命表所导致的。他们提出通过引入一个仅影响与其他死因相关风险的随机脆弱性来校正这种偏差。据我们所知，文献中尚未研究能够处理超额风险模型中未观测异质性的特定脆弱性模型，也未研究未观测异质性对净生存估计的影响；这正是本研究的动机。
本文提出了一种脆弱性超额风险模型，用于在超额风险函数中考虑潜在的未观测异质性（UIH）。也就是说，我们假设 UIH 会影响超额风险，这可能是缺失协变量或模型设定错误的结果。为减轻对模型设定错误的担忧，我们使用具有丰富风险结构的灵活参数模型来拟合超额风险 [34], ，该模型既可以纳入影响时间尺度的协变量，也可以纳入作用于风险尺度的协变量。这类模型包含多种常用的风险结构（如比例风险和加速失效时间）作为特殊情形。我们推导了该模型的一些性质，并针对脆弱性分布的几种选择刻画了边际风险和生存函数。我们对这些函数给出了直观解释，并讨论了所提出脆弱性模型的可识别性。所得模型具有闭式表达，从而能够使用最大似然法以易于处理的方式实现模型，并估计其参数。

本文结构如下。第2节介绍脆弱性模型，并推导边际风险和生存函数，这些函数是写出似然函数所必需的，同时讨论这些量的解释。我们考虑若干种脆弱性分布，它们可使边际风险和生存函数具有闭式表达。第3节介绍用于超额风险的灵活参数模型，并讨论参数的解释。第4节给出似然函数的表达式，并讨论参数的点估计和区间估计。我们还讨论如何选择包含或不包含脆弱性的模型。第5节介绍一项模拟研究，展示所提出模型在具有实际意义的情景中的表现，并说明在存在未观测异质性（由于遗漏了一个协变量）的情况下，个体脆弱性模型恢复边际净生存的能力。第6节通过英格兰肺癌患者的数据展示一项应用。第7节总结研究结果，并对在相对生存框架中使用脆弱性模型进行一般性讨论。本研究的方法已通过 R 包 IFNS 实现，可从 https://github.com/FJRubio67/IFNS 获取。

2 超额风险脆弱性模型

本节中，我们考虑一种基于风险分解（1）的脆弱性模型。该模型将脆弱性模型 [9] 扩展到了相对生存框架。我们计算边际风险和生存函数，稍后将用它们构建似然函数，并讨论所得表达式的解释。我们还讨论了几种脆弱性分布的选择，这些选择可使风险函数和生存函数具有闭式表达。

考虑以下脆弱性模型，其中我们引入一个脆弱性（随机效应） $\lambda \sim G$，它与（1）中的超额风险函数相乘，并允许其在个体之间变化。具体而言，考虑条件风险模型

$$\begin{aligned} \tilde{h}(t \mid \lambda; \mathbf{x}) &= h_P(\text{age}+t; \text{year}+t, \mathbf{z}) + \lambda h_E(t; \mathbf{x}), \\ \lambda &\sim G. \end{aligned} \tag{2}$$

其中， G 是一个绝对连续的累积分布函数，其支持集为 $\mathbb{R}_+$ 且均值为 1。在补充材料中，我们进一步介绍了所提出模型在第3节所述风险结构下的可识别性细节；在脆弱性分布均值为 1 的假设下，该模型的可识别性得到保证。条件生存函数为

$$\tilde{S}(t \mid \lambda; \mathbf{x}) = \exp\{-[H_P(\text{age}+t; \text{year}+t, \mathbf{z}) - H_P(\text{age}; \text{year}, \mathbf{z})]\} \exp\left[-\lambda H_E(t; \mathbf{x})\right], \tag{3}$$

其中， $H_P(\cdot)$ 和 $H_E(\cdot)$ 分别表示与 $h_P(\cdot)$ 和 $h_E(\cdot)$ 相关的累积风险。然后，对脆弱性 λ 关于分布 G 进行积分消去（见补充材料）后，对于特定协变量向量$\mathbf{x}$，亚组特异性边际生存函数可写为

$$\tilde{S}(t; \mathbf{x}) = \exp\{-H_P(\text{age}+t; \text{year}+t, \mathbf{z}) + H_P(\text{age}; \text{year}, \mathbf{z})\} \mathcal{L}_G\{H_E(t; \mathbf{x})\}, \tag{4}$$

其中， $\mathcal{L}_G\{s\} = \int_0^\infty e^{-sr} dG(r)$ 表示在 s 处取值的 G 的拉普拉斯变换。因此，边际风险函数为

$$\begin{aligned}\tilde{h}(t) &= -\frac{d}{dt}\log \tilde{S}(t;\mathbf{x}) \\ &= h_P(\text{age}+t;\text{year}+t,\mathbf{z}) - \frac{\mathcal{L}'_G\{H_E(t;\mathbf{x})\}}{\mathcal{L}_G\{H_E(t;\mathbf{x})\}} h_E(t;\mathbf{x}) \\ &= h_P(\text{age}+t;\text{year}+t,\mathbf{z}) + E[\lambda \mid T_o \geq t] h_E(t;\mathbf{x}), \end{aligned} \tag{5}$$

其中， $T_o = \min\{T_P, T_C\}$ 为观测生存时间， T_P 为其他原因死亡时间， T_C 为癌症死亡时间，且 $\mathcal{L}'_G\{u\} = \frac{\partial}{\partial z}\mathcal{L}_G\{z\}\big|_{z=u}$。最后一个等式的详细推导见补充材料。公式（5）表明，脆弱性对边际超额风险的影响随时间变化，也取决于脆弱性分布的选择 G。这意味着，对于 t的不同取值，边际超额风险的数值可能因脆弱性分布的选择不同而异（尤其是在脆弱性方差较大时）。观察不同分布对应的拉普拉斯变换表达式（例如，补充材料中展示的伽马分布和逆高斯分布）可以更清楚地看到这一点。此外，这使我们能够以与 Balan 等人类似的方式解读脆弱性超额风险模型[9],：在边际层面上，脆弱性被视为一种促使癌症患者中较健康个体被选择出来的因素。此外，这一因素也解释了未观测异质性以及与拟合模型的偏离。接下来，我们考虑分布 G的一种特定选择：伽马分布。除参数具有良好的灵活性和可解释性外，这一选择还可以得到边际生存函数的闭式表达式。补充材料中还给出了逆高斯分布和幂方差函数（PVF）脆弱性分布对应的拉普拉斯变换。PVF 是一种更一般的分布，具有闭式拉普拉斯变换，但多一个参数 [1]。伽马分布是 PVF 的极限情形。其他可导出闭式拉普拉斯变换的三参数脆弱性分布包括复合泊松分布[2] 和稳定分布 [22], ，后两者涵盖伽马分布、逆高斯分布等分布作为特例。Rondeau 等人回顾了其他一些脆弱性分布；使用这些分布时，需要通过数值积分计算边际风险和生存函数[31]。

伽马脆弱性

考虑条件风险模型 (2)，并假设 $\lambda \sim \text{Ga}(\mu, b)$，其中 $\text{Ga}(\mu, b)$ 表示均值参数为 $\mu > 0$、尺度参数为 $b > 0$、概率密度函数为 $g(r;\, \mu, b) = \frac{r^{\frac{\mu}{b}-1}}{\Gamma\left(\frac{\mu}{b}\right) b^{\frac{\mu}{b}}} \exp\left(-\frac{r}{b}\right)$ 的伽马分布。在将均值参数 μ 设为 1 的约束下，脆弱性的尺度参数 b 表示其方差。根据这一假设 ($\mu = 1$)，亚组特异性边际脆弱性生存函数为

$$\tilde{S}(t;\mathbf{x}) = \frac{\exp\{-[H_P(\text{age}+t;\text{year}+t,\mathbf{z}) - H_P(\text{age};\text{year},\mathbf{z})]\}}{\{1+bH_E(t;\mathbf{x})\}^{\frac{1}{b}}}. \tag{6}$$

亚组特异性边际脆弱性净生存函数为

$$\tilde{S}_N(t;\mathbf{x}) = \frac{1}{\{1+bH_E(t;\mathbf{x})\}^{\frac{1}{b}}}. \tag{7}$$

亚组特异性边际脆弱性净生存 (7) 是对脆弱性分布进行边际化、但以协变量 $\mathbf{x}$ 为条件的净生存。这是对经典净生存 $S_N(t;\mathbf{x}) = \exp[-H_E(t;\mathbf{x})]$, 的修正，并且可以看出 $\lim_{b\to 0} S_N(t;\mathbf{x}) = S_N(t;\mathbf{x})$。

亚组特异性边际脆弱性风险函数为

$$\tilde{h}(t;\mathbf{x}) = h_P(\text{age}+t;\text{year}+t,\mathbf{z}) + \frac{h_E(t;\mathbf{x})}{1+bH_E(t;\mathbf{x})}. \tag{8}$$

表达式 (8) 对我们的方法给出了很好的诠释：观测风险可以看作一个模型，其中超额风险上带有时变权重函数 $\omega(t;\mathbf{x}\ b) = \frac{1}{1+bH_E}(t;\mathbf{x})$。该函数形式涉及由未观测异质性引起的校正的尺度或离散程度。此外， $\omega(t;\mathbf{x}\ b)$ 是 $H_E(t;\mathbf{x})$ 和 b 的递减函数。因此，这些量增大时，脆弱性效应会更加显著（当 $0 < \omega(t;\mathbf{x}\ b) \leq 1$ 时如此；而 $\omega(t;\mathbf{x}\ b) = 1$ 表示不存在脆弱性效应）。因此，权重函数所引入的校正

函数 $\omega(t;\mathbf{x}\ b)$ 可以解释为：对于参数的特定取值，与未观测异质性相关的脆弱性会降低超额风险 $h_E(t;\mathbf{x})$。但是，需要注意的是，这并不意味着脆弱性校正总会使超额风险 $h_E(\cdot)$ 缩小，因为模型参数与权重函数中的尺度参数 b 是联合估计的。事实上，含脆弱性模型与不含脆弱性模型中的超额风险不能直接比较。模拟研究和真实数据应用将展示这一效应；在这两项分析中我们观察到，存在未观测异质性（UIH）时，含脆弱性模型与不含脆弱性模型的参数估计值不一定一致。

3 超额风险的灵活参数模型

To 为对超额风险建模，我们考虑灵活参数的一般风险（GH）结构 [14]:

$$h_E(t;\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}) = h_0\left(t\exp\{\mathbf{w}^\top\boldsymbol{\alpha}\};\boldsymbol{\theta}\right)\exp\left\{\mathbf{x}^\top\boldsymbol{\beta}\right\}. \tag{9}$$

其中 $\mathbf{x}\in\mathbb{R}^p$，$\mathbf{w}\in\mathbb{R}^{p_t}$, $\boldsymbol{\alpha}\in\mathbb{R}^{p_t}$ 和 $\boldsymbol{\beta}\in\mathbb{R}^p$。通常 $\mathbf{w}\subseteq\mathbf{x}$。$h_0(\cdot;\boldsymbol{\theta})$ 表示参数为 $\boldsymbol{\theta}\in\Theta$ 的参数化基线风险函数。这是一种丰富的风险结构，包含以下特例：比例风险（PH）模型（$\boldsymbol{\alpha}=0$）、加速风险（AH）模型（$\boldsymbol{\beta}=0$）和加速失效时间（AFT）模型（$\boldsymbol{\alpha}=\boldsymbol{\beta}$，$\mathbf{w}=\mathbf{x}$）。关于 GH 结构的详尽讨论，见 [34] 。该结构允许纳入时间尺度效应（通过 $\mathbf{w}$）以及作用于风险水平的效应（$\mathbf{x}$）。采用灵活的超额风险模型有助于减少与模型设定错误相关的未观测异质性，从而有助于我们专注于捕捉潜在缺失协变量的影响。Gasparini 等人在总生存框架下讨论了这些问题[18]。相应的累积风险函数为

$$H_E(t;\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}) = H_0\left(t\exp\{\mathbf{w}^\top\boldsymbol{\alpha}\};\boldsymbol{\theta}\right)\exp\left\{\mathbf{x}^\top\boldsymbol{\beta}-\mathbf{w}^\top\boldsymbol{\alpha}\right\}. \tag{10}$$

超额累积风险函数可以写成闭式，这一事实有助于实现似然函数（将在下一节讨论），也便于从该模型模拟生存时间。具体而言，利用概率积分变换，从 GH 模型 (9) 模拟一个随机事件发生时间很简单（另见 Rossell 和 Rubio32）：

$$t = \frac{F_0^{-1}\left[1-\exp\left\{\log(1-u)\exp(\mathbf{w}^\top\boldsymbol{\alpha}-\mathbf{x}^\top\boldsymbol{\beta})\right\};\boldsymbol{\theta}\right]}{\exp(\mathbf{w}^\top\boldsymbol{\alpha})}, \tag{11}$$

其中 $u\sim$ (0,1)，F_0 是与基线风险 h_0相对应的累积分布函数。类似地，可以按以下方式从脆弱性模型 (2) 进行模拟

$$t = \frac{F_0^{-1}\left[1-\exp\left\{\log(1-u)\exp(\mathbf{w}^\top\boldsymbol{\alpha}-\mathbf{x}^\top\boldsymbol{\beta})/\lambda\right\};\boldsymbol{\theta}\right]}{\exp(\mathbf{w}^\top\boldsymbol{\alpha})},$$

其中 $u\sim$ (0,1)，且 $\lambda\sim G$。

关于基线风险的选择，Rubio 等人[34] 讨论了使用灵活参数分布时的不同选择。该风险函数理想的性质包括：数值上易于处理，并且能够捕捉风险的基本形态（递增、递减、单峰、浴缸形）。Rubio 等人[34] 提到指数化威布尔（EW）和广义伽马（GG）是基线风险的特定选择。本文考虑使用幂广义威布尔（PGW）分布 [7], , 但我们强调，也可以采用其他灵活的基线风险分布。PGW 分布是一种灵活的三参数分布（具有尺度参数 $\sigma>0$ 和两个形状参数 $\nu,\gamma>0$），既能捕捉风险的基本形态，又具有易于处理的风险函数和生存函数表达式 [6]。PGW 分布的概率密度函数、生存函数和风险函数见补充材料。我们也可以考虑更简单的（2 参数）基线风险，以捕捉特定的基本形态，例如对数正态分布（单峰）、对数逻辑斯蒂分布和伽马分布。

4 推断

全文中，令 $(t_1,\ldots,t_n)$ 表示总体中 n 名癌症患者的生存时间；$(\delta_1,\ldots,\delta_n)$ 表示相应的生命状态指标（0-存活，1-死亡）；$\mathbf{x}_i\in\mathbb{R}^p, i=,\ldots,n$ 表示第 i 位患者可用的协变量向量；age$_i$ 表示诊断年龄；year$_i$ 表示诊断年份。个体脆弱性模型 (4)–(5) 的似然函数为

$$\tilde{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma,b) \propto \prod_{i=1}^{n}\left[h_P(\text{age}_i+t_i;\text{year}_i+t_i,\mathbf{z}_i)+\frac{h_E(t_i;\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma))}{1+bH_E(t_i;\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma))}\right]^{\delta_i}\times\frac{1}{\{1+bH_E(t_i;\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma)\}^{\frac{1}{b}}}.$$

相比之下，不含脆弱性的模型 (1) 对应的似然函数为

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma) \propto \prod_{i=1}^{n}[h_P(\text{age}_i+t_i;\text{year}_i+t_i,\mathbf{z}_i)+h_E(t_i;\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma)]^{\delta_i}\times\exp\{-H_E(t_i;\mathbf{x}_i,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma,\nu,\gamma)\}.$$

这些模型的参数点估计将通过最大似然估计获得。置信区间将使用渐近正态近似计算；鉴于我们在癌症流行病学中的应用通常具有较大的样本量，这种做法是合理的。简言之，设 $\boldsymbol{\psi}$ 为所关注模型的完整参数向量，我们考虑使用 $\widehat{\boldsymbol{\psi}}\pm Z_{1-\frac{\tau}{2}}\text{diag}\left(J^{-\frac{1}{2}}(\widehat{\boldsymbol{\psi}})\right)$ 类型的置信区间，其中 $J(\boldsymbol{\psi})=-\frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}\log L(\boldsymbol{\psi})$ 是在适当参数化下对数似然函数的海森矩阵的负值，而 $1-\tau\in(0,1)$ 是置信水平。

我们使用赤池信息准则比较这两个模型，并通过模拟研究评估它们的性能。需要指出的是，使用似然比检验比较这两个模型更为复杂，因为似然比检验统计量的分布不一定渐近服从自由度为 1 的卡方分布（见第 8 章，共 20 章）。此外，已有一些针对混合模型制定的信息准则，也可以使用 [26]。特别是在我们的应用中，样本量足够大，使得赤池信息准则与边际赤池信息准则 [26] 非常接近。深入探讨不同信息准则的使用超出了本文的研究目标；关于这一点更详尽的讨论，请参阅 Müller 等人[26] 。

5 模拟研究

本节将介绍一项模拟研究，探讨脆弱性模型在若干情景下的表现。所提出的情景旨在实现以下具体目标 [12]:（目标 1）考察所提方法在不同设定下的有限样本性能；（目标 2）评估总体亚组中缺失协变量（具有异质分布）的影响。

5.1 目标 1：有限样本性能

数据生成与模拟设计

我们考察（i）不同样本量以及（ii）回归系数取不同值对所提方法性能的影响。这些回归系数的不同真实值对应 3 种不同情景（下文称为 Sc1、Sc2 或 Sc3）；对于每种情景，我们考虑四种不同的样本量（$n\in\{500,1000,2000,5000\}$）。每种情景下，我们生成 4 个协变量：一个连续协变量（如年龄）通过均匀分布混合进行模拟，其中取值范围为 (30,65) 的概率为 0.25，范围为 (65,75) 的概率为 0.35，范围为 (75,85) 岁的概率为 0.40；其余 3 个协变量（如性别、 X_1 和 X_2）均为二元变量，取值为 0 或 1 的概率相等（即 $P(\text{性别}=0)=P(\text{性别}=1)=0.5$）。

生存时间的模拟基于总体死亡风险的加性分解，详见方程（1）。我们使用英国生命表模拟“其他死因”的事件发生时间，并使用逆变换法（11）模拟癌症事件时间（即由超额风险导致的事件发生时间），同时假设模型（9）。具体而言，我们采用GH结构，其中包括PGW基线超额风险、均值为 $\mu^*=1$ 且尺度为 b^* 的伽马脆弱性（其取值取决于模拟情景），以及协变量年龄、性别、 X_1 和 X_2 的效应（第3节详细介绍了使用 PGW 分布的逆变换法）。我们假设随机脱落时间服从具有给定速率 r的指数分布，并选择该速率的取值，使每个情景中约有5%的随机脱落。最后，我们假设在5年时进行行政删失。这两种删失来源（即随机脱落和行政删失）在所有情景下合计导致40%至45%的删失。我们在每个情景中生成并分析 $M=1000$个样本。

对于 Sc1，基线分布参数的真实值为 $\sigma^*=0.75$， $\nu^*=1.75$， $\gamma^*=8$，伽马脆弱性的尺度为 $b^*=0.5$，回归系数为 $\boldsymbol{\alpha}^*=\boldsymbol{\beta}^*=(1,\ 1,\ 1,1)$。我们强调，这些都是复杂的模拟情景，其中纳入了 4 个协变量，并同时包含时间依赖效应和风险层级效应，同时采用灵活的基线风险。在实践中，通常只会将部分协变量（例如 年龄）作为时间依赖效应，从而减少模型参数的实际数量。Sc1 旨在展现所提出的脆弱性模型在极具挑战性的情景下的表现。

在 Sc2 和 Sc3 中，我们假设了比 Sc1 更简单的真实模型，仅假设 2 个具有时间依赖效应的协变量。此外，为涵盖适当范围的可能情景，我们还假设基线风险分布的参数和随机效应的方差

与 Sc1 中的设定不同。Sc2 和 Sc3 使用的真实值分别见补充材料中的表 1 和表 2。

模拟数据分析

我们使用模型（9）分析超额死亡风险数据，并假设其参数化方式与模拟数据时采用的方式相同。我们报告估计回归系数的均值、偏倚（即估计回归系数均值与真实值之差）、估计回归系数的中位数、经验标准差、平均（估计）标准误，以及渐近置信区间的覆盖比例。

我们使用 R 软件完整实现了所提出的方法。优化步骤通过 R 命令 'nlminb' 完成。回归系数的标准误根据命令 'hessian'（R软件包 'numDeriv'）得到的海森矩阵计算，渐近 95% 置信区间则使用这些标准误近似得出。优化步骤的初始值采用基于 PGW 分布、但不包含脆弱性参数的比例风险模型所估计的参数（因此，伽马分布尺度的初始值设为 1；对于时间依赖效应对应的参数，则采用比例风险模型得到的估计值作为初始值）。对于优化器未能收敛（由 'nlminb' 指示），或命令 'hessian' 产生 'Inf' 或 'NaN' 值的情况，我们将其排除：Sc1 中，在观测值为 $n = 500$，1000，2000， 5000 时，分别有 1000 组估计中的 31、16、3 和 0 组被排除。Sc2 中，在观测值为 $n = 500$， 1000，2000， 5000 时，分别有 1000 组估计中的 10、1、0 和 0 组被排除；Sc3 中则没有排除任何组。

结果

Sc1 的结果详见下方表 1，而 Sc2 和 Sc3 的结果分别见补充材料中的表 1 和表 2。正如预期，随着样本量增加，偏倚会降低，覆盖率也会更接近名义值。采用不同的回归系数和脆弱性方差值，并不会改变模型性能及观察到的结果，如补充材料中的表 1 和表 2 所示。我们观察到，与基线风险参数和脆弱性尺度参数相比，协变量对应的回归系数估计表现有所不同。对于协变量对应的回归系数，即使样本量只有 500 例，模型表现也很好。对于基线分布参数和脆弱性的尺度参数，当样本量为 500 或 1000 时，偏倚较大；当样本量更大（2000 或 5000）时，模型表现良好，偏倚较小，覆盖率接近名义值。

对于 Sc1 的 500个观测值（因此基线分布参数和脆弱性尺度存在不可忽略的偏倚），我们检验该方法能否恢复亚组特异性边际净生存的真实值（即对脆弱性分布积分，但以协变量为条件，见方程（7））。我们可以看到，对于参考组（即 协变量取值设为 0），尽管存在上述偏倚，在所有模拟样本中取平均后，亚组特异性边际净生存的真实值仍得到了较好的恢复（图1）。

5.2 目标2：缺失协变量具有异质分布的影响

我们现在考察以下两种模拟情景。总体目标是比较两种方法：一种是使用基于总人群拟合的模型预测人群亚组的净生存，这在流行病学中是一种常见策略；另一种是进行分层分析。在第一种情景中，我们假设总体中存在两个亚组，由变量性别定义，且两组的 PGW基线风险略有不同（性别为 $= 1$ 时为 $\boldsymbol{\theta}_1 = (0.5, 1.5, 5)$，性别为 $= 0$ 时为 $\boldsymbol{\theta}_2 = (0.5, 1.5, 3)$）。变量性别为二元变量，取值为 1 的概率为 0.6，取值为 0 的概率为 0.4。对于每个亚组，我们生成 2 个协变量：一个连续协变量（例如年龄）通过均匀分布混合模拟，其中取值范围为 (30, 65)、(65, 75) 和 (75, 85) 年的概率分别为 0.25、0.35 和 0.40；变量 X_1 也是二元变量，在性别为 $= 1$ 时取值为 1 的条件概率为 0.8，在性别为 $= 0$ 时取值为 1 的条件概率为 0.4（也就是说，该变量在不同性别中的分布不同）。此外，我们假设协变量的效应因性别亚组而异：性别为 $= 1$ 时， $\boldsymbol{\alpha} = (0.7, 0.7, 0.5)$，且 $\boldsymbol{\beta} = (1, 0.5, 1)$；性别为 $= 0$ 时， $\boldsymbol{\alpha} = (0.7, 0.7, 0.25)$，且 $\boldsymbol{\beta} = (0.5, 0.5, 0.25)$。在第二种模拟情景中，我们再次假设总体中存在两个由变量性别定义的亚组，但两组的 PGW基线风险现在存在明显差异（性别为 $= 1$ 时为 $\boldsymbol{\theta}_1 = (0.5, 1.5, 5)$，性别为 $= 0$ 时为 $\boldsymbol{\theta}_2 = (0.5, 1.5, 0.75)$）。协变量的模拟方式与第一种情景相同，回归系数的真实值 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 也按第一种情景设定。对于两种情景，我们考虑的样本量为 $n = 500$、1000 和 5000，删失率约为 65%。

对于每种情景，我们首先对每个模拟数据集拟合不含脆弱性的经典模型 (1) 和采用 PGW基线风险的脆弱性模型 (8)，并省略变量 X_1。我们通过

参数	True	MLE均值	Bias	MLE中位数	覆盖率	标准误均值	经验标准差
情景 N=500							
σ	0.750	0.974	0.224	0.761	0.817	0.446	0.622
ν	1.750	1.983	0.233	1.873	0.932	0.397	0.498
γ	8.000	8.690	0.690	8.609	0.787	4.097	4.986
α_1	1.000	1.097	0.097	1.006	0.928	0.607	1.333
α_2	1.000	0.984	-0.016	0.995	0.914	0.273	0.639
α_3	1.000	1.027	0.027	0.991	0.924	0.565	1.197
α_4	1.000	1.010	0.010	0.986	0.917	0.608	1.420
β_1	1.000	0.940	-0.060	1.016	0.941	0.384	0.846
β_2	1.000	1.007	0.007	0.996	0.964	0.165	0.369
β_3	1.000	0.989	-0.011	1.009	0.957	0.359	0.675
β_4	1.000	1.003	0.003	1.014	0.946	0.378	0.821
b	0.500	1.231	0.731	0.549	0.828	0.810	1.675
情景 N=1000							
σ	0.750	0.870	0.120	0.740	0.885	0.333	0.453
ν	1.750	1.847	0.097	1.820	0.937	0.246	0.269
γ	8.000	8.364	0.364	8.246	0.869	3.065	3.404
α_1	1.000	1.029	0.029	1.000	0.957	0.319	0.582
α_2	1.000	1.023	0.023	1.001	0.955	0.162	0.373
α_3	1.000	1.004	0.004	0.972	0.953	0.319	0.570
α_4	1.000	0.969	-0.031	0.989	0.945	0.326	0.584
β_1	1.000	0.994	-0.006	1.008	0.954	0.189	0.263
β_2	1.000	0.984	-0.016	0.995	0.948	0.090	0.207
β_3	1.000	0.997	-0.003	1.005	0.952	0.191	0.327
β_4	1.000	1.007	0.007	0.999	0.956	0.194	0.317
b	0.500	0.836	0.336	0.541	0.923	0.616	1.063
情景 N=2000							
σ	0.750	0.821	0.071	0.759	0.942	0.227	0.309
ν	1.750	1.778	0.028	1.766	0.962	0.165	0.162
γ	8.000	8.061	0.061	8.010	0.929	2.170	2.292
α_1	1.000	1.015	0.015	0.997	0.959	0.192	0.272
α_2	1.000	1.010	0.010	1.005	0.954	0.093	0.154
α_3	1.000	1.009	0.009	0.991	0.945	0.197	0.338
α_4	1.000	1.009	0.009	0.995	0.945	0.193	0.310
β_1	1.000	1.003	0.003	1.008	0.947	0.113	0.147
β_2	1.000	0.994	-0.006	0.998	0.956	0.048	0.087
β_3	1.000	0.996	-0.004	1.004	0.946	0.116	0.185
β_4	1.000	0.997	-0.003	1.003	0.939	0.114	0.175
b	0.500	0.660	0.160	0.550	0.964	0.409	0.607
情景 N=5000							
σ	0.750	0.760	0.010	0.747	0.950	0.123	0.125
ν	1.750	1.764	0.014	1.760	0.954	0.102	0.102
γ	8.000	8.108	0.108	8.026	0.951	1.352	1.369
α_1	1.000	0.991	-0.009	0.987	0.952	0.109	0.110
α_2	1.000	1.001	0.001	1.002	0.949	0.051	0.052
α_3	1.000	0.988	-0.012	0.986	0.948	0.110	0.109
α_4	1.000	0.987	-0.013	0.989	0.936	0.110	0.113
β_1	1.000	0.999	-0.001	0.997	0.945	0.066	0.066
β_2	1.000	1.000	-0.000	0.998	0.949	0.027	0.028
β_3	1.000	1.001	0.001	1.000	0.943	0.066	0.067
β_4	1.000	1.002	0.002	1.001	0.938	0.066	0.068
b	0.500	0.565	0.065	0.543	0.941	0.229	0.233

表 1：目标 1、情景 1 的模拟结果。MLE均值：最大似然估计的均值；MLE中位数：最大似然估计的中位数；标准误均值：标准误的均值；经验标准差：经验标准差

对两种模型中与各患者对应的个体净生存曲线取平均，计算相应的总体净生存曲线。我们将真实的总体净生存曲线与使用这些模型得到的拟合净生存曲线进行比较。然后，我们通过对相应观测值子集中的个体净生存曲线取平均，计算按性别 $= 0,\ 1$ 定义的亚组净生存率。我们的目的是比较整个总体的净生存曲线，以及在存在未观测异质性（UIH）的情况下，使用基于整个样本拟合的模型估算亚组净生存率的效果。

第二阶段，我们进行分层分析，按照变量性别 $= 0,1$ 定义的每个数据子集，分别拟合不含脆弱性的经典模型（1）和脆弱性模型（8）。然后，我们利用这些模型得到的个体拟合净生存曲线的平均值，计算与亚组性别 $= 0,1$ 对应的净生存曲线。我们再次将相应的净生存曲线与这些亚组的真实净生存率进行比较。

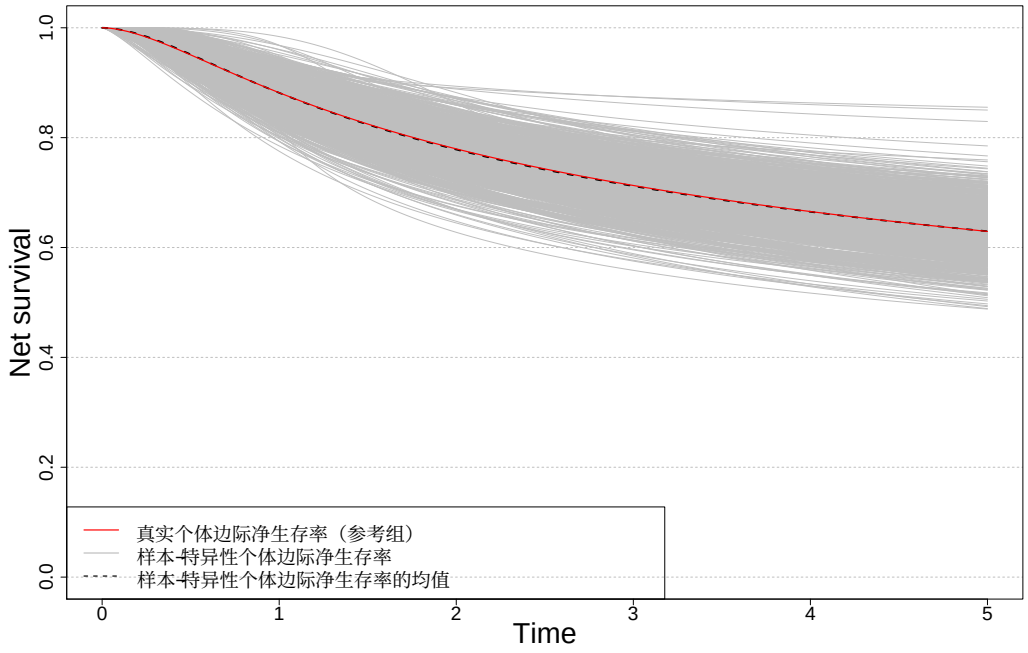


图1：Scena rio Sc1（500个观测值）：参考组的亚组特异性边际净生存

结果见补充材料（图1–2）。在第一种情景下，对于所有样本量（补充材料中的图1），我们发现，拟合的总体净生存曲线平均值与真实总体净生存率非常接近。亚组分析得到的净生存曲线也较为接近，但经典模型的拟合略差（且这一情况在所有样本量下均存在）。分层分析的结果最佳，拟合净生存曲线的平均值几乎与真实净生存曲线完全相同。在第二种情景下（补充材料中的图2），我们发现，拟合的总体净生存曲线平均值接近真实总体净生存率，但两种模型在所有样本量下都存在轻微偏倚。亚组分析得到的两条平均曲线均存在很大的偏倚，而且经典模型与脆弱性模型之间也存在差异。对于所有样本量，采用分层分析都能大幅降低这种偏倚。

6 真实数据应用：肺癌流行病学

我们分析了一组数据，该数据来自基于人群的全国性非小细胞肺癌（NSCLC）登记系统，涵盖了2012年在英格兰确诊的女性患者 [10]。我们纳入以下协变量：标准化诊断年龄（agec，连续变量）；诊断时肿瘤分期（stage，分类变量，I–IV期）；是否存在心血管合并症（CVD，二元变量）；是否存在慢性阻塞性肺疾病（COPD，二元变量）；以及根据2010年英格兰多重剥夺指数（IMD）计算的标准化收入领域指标（连续变量），其定义至下层超级输出区层级（作为剥夺程度的衡量指标）。诊断时分期和合并症信息来自关联数据（医院就诊统计 -HES- 和肺癌审计数据）。该队列的患者随访至2015年12月31日，当时仍存活的患者被右删失。我们将分析范围限定为协变量无缺失的女性。最终样本包含 $n = 14557$ 名完整病例，其中 $n_o = 12138$ 人在2015年12月31日前死亡。患者诊断年龄的第25%、50%和75%分位数分别为 64.93、72.64 和 80.23 岁，平均年龄为 72.00 岁。在这些患者中，2434 人诊断为 I期，1131 人为 II期，3241 人为 III期，7751 人为 IV期。最后，1954 名患者被归类为患有心血管合并症，3260 名患者患有慢性阻塞性肺疾病。

为了评估未观测异质性的影响，我们比较了该数据集的八个模型。模型*C1*：（经典）不含脆弱性的模型，其中年龄具有时间层级效应（$\mathbf{w}_i$），而年龄、多重剥夺指数、分期、心血管疾病和慢性阻塞性肺疾病具有风险层级效应（$\mathbf{x}_i$）。模型 *F1*：脆弱性模型，其中年龄具有时间层级效应（$\mathbf{w}_i$），而年龄、多重剥夺指数、分期、心血管疾病和慢性阻塞性肺疾病具有风险层级效应（$\mathbf{x}_i$）。模型*C2*：不含脆弱性的模型，其中年龄具有时间层级效应（$\mathbf{w}_i$），而年龄、多重剥夺指数、心血管疾病和慢性阻塞性肺疾病具有风险层级效应（$\mathbf{x}_i$）。模型*F2*：脆弱性模型，其中年龄具有时间层级效应（$\mathbf{w}_i$），而年龄、多重剥夺指数、心血管疾病和慢性阻塞性肺疾病具有风险层级效应（$\mathbf{x}_i$）。模型*C3*：不含脆弱性的模型，其中年龄具有时间层级效应（$\mathbf{w}_i$），而

年龄和多重剥夺指数具有风险层级效应（$\mathbf{x}_i$）。模型 *F3*：脆弱性模型，其中年龄具有时间层级效应（$\mathbf{w}_i$），而年龄和多重剥夺指数具有风险层级效应（$\mathbf{x}_i$）。模型 *C4*：采用 PGW基线风险且不含协变量。模型 *F4*：采用 PGW基线风险且不含协变量的脆弱性模型。

图2a展示了使用模型 C1–C4 和 F1–F4 得到的整个队列的净生存率。这些结果通过以下公式计算：

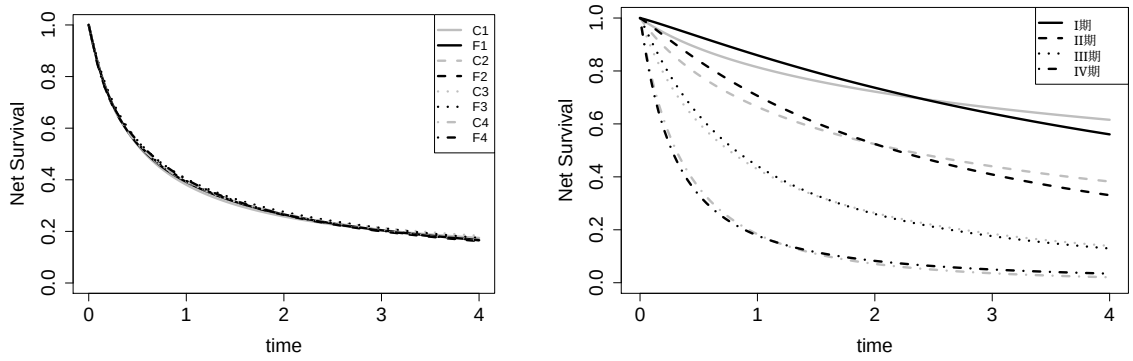
$$
\begin{aligned}
S_N(t) &= \frac{1}{n}\sum_{i=1}^{n}\exp\{-H_E(t;\mathbf{x}_i,\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\sigma},\widehat{\nu},\widehat{\gamma})\},\ \text{Models C1–C4,}\\
\tilde{S}_N(t) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left\{1+\widehat{b}H_E(t;\mathbf{x}_i,\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\sigma},\widehat{\nu},\widehat{\gamma})\right\}^{\frac{1}{\widehat{b}}}},\ \text{Models F1–F4.}
\end{aligned}
$$

模型 C1–C4 和 F1–F4 的赤池信息准则值见表2。总体上，最佳模型为模型 F1，显然更支持纳入所有协变量和脆弱性的模型。在经典模型（C1–C4）中，最佳模型为模型 C1，同样支持纳入所有协变量。模型比较结果和图2得出了一些有趣的结论。尽管赤池信息准则明确支持模型 F1，但图2a显示，使用模型 C1–C4 和 F1–F4 得到的总体净生存曲线估计值非常相似。这表明，即使数据更支持脆弱性模型，基于包含和不包含脆弱性的模型得到的某些估计量仍可能相似。不过，这绝不是普遍结论。图2b显示，按分期分层的模型 C1 和 F1 的净生存率在早期肿瘤分期之间存在明显差异。由于肿瘤分期协变量的分布以晚期患者为主（见描述性分析），这些差异在图2a中并不明显。也就是说，由于我们关注的是总体亚组平均值的估计，相应协变量（患者特征）的分布起着关键作用。因此，尽管估计得到的模型差异很大（见表2），协变量的分布会使人群层面得到相似的边际净生存曲线。不过，这一结论并不适用于按分期分层的净生存函数。表3给出了 $t = 1$, 2, 3, 4 年时的净生存率置信区间。这些置信区间通过基于最大似然估计量渐近正态性的蒙特卡罗近似得到（更多详情见 Rubio 等人[34]）。这些置信区间证实，净生存率估计的差异也体现在区间估计中。

另一个值得关注的问题是参数的解释。从表2可以看出，模型 C1 和 F1 的基线风险参数估计值存在显著差异。这是因为模型 C1 对应的基线风险估计值呈下降趋势，而模型 F1 对应的基线风险估计值呈上升趋势。这解释了参数估计值为何不同；这些差异只是在反映两者截然不同的形态。不过，我们想强调，模型 C1 和 F1 的基线风险（更一般地说，经过脆弱性校正与未经脆弱性校正的模型）不能直接比较，因为（i）模型 F1 的估计值对应于一个边际模型（对脆弱性项取边际化），且（ii）脆弱性模型（8）中的基线风险同时控制分子中的超额风险和时变权重。尽管两者的性质不同，比较它们仍然有用，因为两个模型的估计值存在差异，说明存在未观测异质性；只有当脆弱性项的方差不可忽略时，才会出现这种差异。

模型	C1	F1	C2	F2	C3	F3	C4	F4
$\widehat{\sigma}$	0.763	8.760 (7.072,10.851)	0.109	0.769	0.107	0.737	0.099	0.562
$\widehat{\nu}$	0.987	1.077 (1.046,1.108)	1.196	1.273	1.199	1.288	1.197	1.407
$\widehat{\gamma}$	4.980	0.813（0.675，0.980）	4.185	0.323	4.173	0.314	4.325	0.272
年龄$_t$	0.308	2.633 (1.367,3.900)	0.271	0.442	0.272	0.436	–	–
agec	0.328	0.124（0.000 0.248）	0.331	0.289	0.340	0.303	–	–
IMD	0.530	0.827（0.578，1.075）	0.305	1.079	0.319	1.126	–	–
第2期	0.706	0.964（0.825，1.102）	–	–	–	–	–	–
阶段3	1.460	2.085 (1.963,2.207)	–	–	–	–	–	–
第4阶段	2.202	3.312（3.169，3.456）	–	–	–	–	–	–
CVD	0.214	0.352 (0.267,0.437)	0.147	0.295	–	–	–	–
COPD	0.142	0.242（0.172，0.311）	-0.018	-0.054	–	–	–	–
$\widehat{b}$	–	0.764（0.687，0.850）	–	4.542	–	4.785	–	6.836
AIC	20548.43	**20140.06**	26328.84	26311.62	26306.3	26302.23	26864.78	26808.51

表2：最大 li 模型 C1–C4 和 F1–F4 的似然估计值和赤池信息准则。另列出选定模型的参数 F1 的置信区间。



(a) (b)

图2：肺癌数据：（a）总人群的净生存曲线，以及（b）按诊断时肿瘤分期分层的净生存曲线（模型 C1——灰色线，模型 F1——黑色线）。

按分期划分的总人群

		经典（C1）			脆弱性（F1）		
分期	year	NS	下限	上限	NS	下限	上限
I	1	0.814	0.801	0.827	0.859	0.850	0.873
	2	0.722	0.704	0.740	0.737	0.721	0.760
	3	0.661	0.640	0.681	0.639	0.620	0.668
	4	0.615	0.590	0.637	0.560	0.542	0.591
II	1	0.664	0.642	0.688	0.706	0.685	0.732
	2	0.524	0.497	0.551	0.524	0.501	0.557
	3	0.440	0.413	0.466	0.409	0.385	0.442
	4	0.382	0.355	0.411	0.331	0.309	0.359

表3： $t = 1$，2，3，4 年净生存率（NS）及 95% 置信区间：经典模型（模型 C1）和脆弱性模型（模型 F1）。

分层分析

图2中 I–IV期的净生存曲线存在差异，这表明存在未观测异质性，而且未观测协变量在各个分层中可能具有不同的条件分布。为探究这一观点，我们现在进行分层分析，研究三个不同分层中的患者：I–II期、III期和IV期肿瘤。进行这种分层有临床和生物学依据，因为不同的肿瘤分期反映了肿瘤大小及其向其他器官扩散（转移）的情况。我们对每个组分别比较脆弱性模型和非脆弱性模型，以评估未观测异质性的影响。对于 I–II期，我们拟合以下模型：模型 *C1-I.II*：不含脆弱性的模型，年龄的时间层级效应为（$\mathbf{w}_i$），年龄、多重剥夺指数、分期、心血管疾病和慢性阻塞性肺疾病的风险层级效应为（$\mathbf{x}_i$）。模型 *F1-I.II*：脆弱性模型，年龄的时间层级效应为（$\mathbf{w}_i$），年龄、多重剥夺指数、分期、心血管疾病和慢性阻塞性肺疾病的风险层级效应为（$\mathbf{x}_i$）。模型 *C1-III*、*F1-III*、*C1-IV*和 *F1-IV* 分别代表III期和IV期对应的模型。结果见表4。对于 I–II期，不含脆弱性的模型略占优势，估计的脆弱性方差较小。对于III期，脆弱性模型更受支持，且我们观察到模型参数估计值存在差异。同样，对于IV期，脆弱性模型更受支持，估计的脆弱性方差较大，模型参数的估计值也存在明显差异。因此，我们注意到，不同肿瘤分期分层中的未观测异质性程度各不相同。从图3可以看出，含脆弱性和不含脆弱性的模型所得到的净生存曲线重合。这表明，按分期对数据进行分层后，未观测异质性对净生存估计的影响较小，尽管III期和IV期存在未观测异质性的证据。值得注意的是，在模型中遗漏重要协变量会影响基线风险参数和回归系数的估计（例如，模型 F1、F2 和 F3 在表2中的估计值）。这是一个众所周知的现象

与风险回归模型中遗漏重要协变量有关（关于在生存回归模型中遗漏协变量的影响，详见 Rossell 和 Rubio[32] 的讨论）。

	C1-I.II	F1-I.II	C1-III	F1-III	C1-IV	F1-IV
$\widehat{\sigma}$	2.172	2.839	0.430	1.315	0.090	0.689
$\widehat{\nu}$	1.099	1.095	1.150	1.074	1.241	1.277
$\widehat{\gamma}$	3.203	2.560	2.286	0.856	3.110	0.235
年龄$_t$	0.399	0.354	0.516	-2.119	0.287	0.314
agec	0.564	0.568	0.363	0.691	0.291	0.277
IMD	0.731	0.758	0.172	0.292	0.631	1.507
第2期	0.770	0.836	–	–	–	–
CVD	0.446	0.490	0.262	0.359	0.131	0.333
COPD	0.482	0.511	0.092	0.147	0.085	0.260
$\widehat{b}$	–	0.211	–	0.666	–	3.075
AIC	**7817.53**	7818.49	6986.52	**6962.63**	5218.65	**5214.64**

表4：按分期分层数据拟合的模型的最大似然估计。

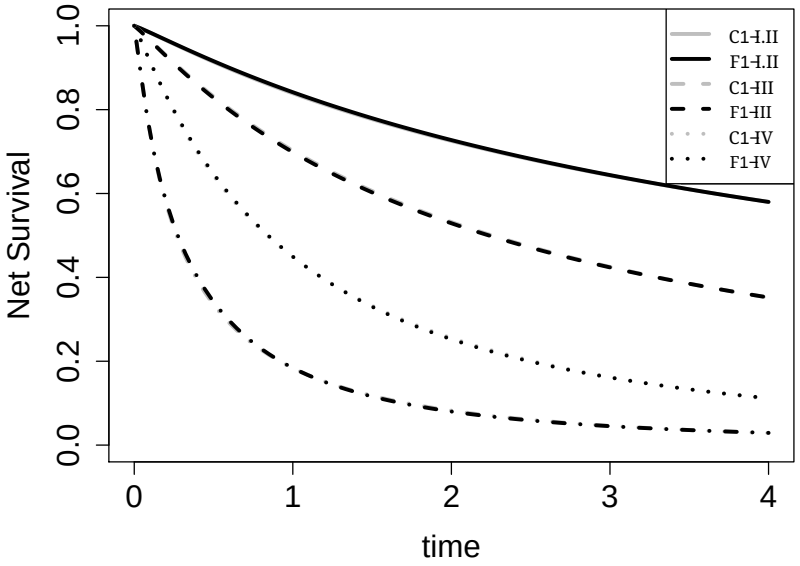


图3：肺癌数据：按分期分层数据拟合的模型的净生存曲线。灰线表示模型 C1 在相应分层（I–II期、III期或IV期）拟合得到的净生存曲线。黑线表示模型 F1 在相应分层（I–II期、III期或IV期）拟合得到的净生存曲线。

7 讨论

我们将脆弱性模型扩展到相对生存框架，并表明这些模型会使癌症幸存者中的个体呈现健康状况较好的选择效应。我们还表明，所提出的超额风险脆弱性模型能够捕捉未观测个体异质性，这是人群癌症流行病学中常见的挑战。我们提出了一类通用的参数化超额风险回归模型，该模型能够保证参数的可识别性，并包含具有实际意义的风险结构。所提出的模型便于实现似然函数，因为我们可以纳入时间层级效应，同时避免数值积分。尽管如此，我们指出，也可以采用其他风险结构。虽然我们重点研究了伽马脆弱性分布，因为它便于处理且易于解释，但我们还介绍了几种其他的

脆弱性分布选择，这些分布可导出闭式表达式（见补充材料）。全面比较不同脆弱性分布选择所产生的影响，不在本研究的目标范围内，但我们将在未来的研究中对此进行探讨。

我们的模拟表明，所提出的模型具有良好的推断性质。我们在极具挑战性的情景下评估了所提出方法的性能：样本量较小（$n = 500$）、删失比例较高（40%），且生成模型非常复杂（所考虑的4个协变量既对时间尺度有影响，也对风险水平有影响）。在实际应用中，我们不会预期所有协变量都同时影响时间尺度和风险水平。如我们的模拟目标2所示，如果使用拟合于总人群的模型来获取异质性亚组的净生存曲线，未观测异质性可能会产生不可忽视的影响。这表明，使用经典模型和脆弱性模型得到的、与总人群亚组相关的净生存曲线之间存在差异，可能意味着未观测异质性具有不可忽视的影响；分层分析或有助于减少这种异质性。

我们已经表明，不出所料，在总生存框架中报告过的未观测异质性（UIH）的影响及其解释 [9] 在相对生存框架中同样存在。我们的研究提供了一个补充视角，即探讨 UIH 对边际量（如净生存）估计的影响。我们表明，所提出的脆弱性模型能够检测 UIH。不过，务必牢记这样做在实践中的局限性，因为 UIH 的存在也可能表明模型不准确或模型设定错误 [27]。我们提出了一种脆弱性 GH 模型，可纳入时变效应。因此，采用通用且灵活的模型有助于减少与模型设定错误相关的 UIH。当拟合于总人群的模型也用于估计与总人群亚组相关的边际量时 [39], ，区分 UIH 与模型设定错误尤为重要，因为目标亚组可能具有不同的生成模型，或所纳入协变量的分布不同。因此，存在 UIH 时，将结果与分层分析所得结果进行比较会很有帮助。对于有意估计患者亚组边际量的分析人员而言，这是一个值得关注的发现。当然，由于 UIH 会导致模型参数估计产生偏倚（对于不含个体脆弱性的模型），条件量（如特定协变量取值下的超额风险）的估计也会产生偏倚。我们还表明，UIH 会在随访时间内引发选择效应。Vaupel 和 Yashin [39], 也曾在总生存框架中讨论过这类效应，他们发现，在某些情景下，UIH 会严重影响条件量（或个体量）的估计。

我们展示了需要采用灵活的三参数分布（例如，PGW）来建模基线风险形状的情景。也存在一些情景，数据可能更支持较简单的基线风险模型（例如，对数正态分布）或较简单的风险结构（如加速失效时间模型或比例风险模型）。在这类情况下，模型选择工具（如赤池信息准则或贝叶斯信息准则）有助于确定最适合数据的模型（关于讨论，参见 Rubio 等人[34] 以及 Rubio 和 Drikvandi[33]）。

真实数据应用清楚地说明，使用包含和不包含脆弱性项的模型所得到的净生存估计存在差异，而这种差异取决于患者特征的分布。这与模拟研究目标2所得结论一致。净生存是一种边际度量，因为它是对观测到的协变量取平均后计算得出的。因此，其估计自然取决于协变量的分布。在实践中，首先比较不同模型以检测未观测异质性（UIH）十分重要，随后还应考察存在 UIH 对感兴趣量（如净生存）估计的影响。此外，真实数据应用表明，UIH 的程度可能取决于观测到的协变量（例如，我们示例中的肿瘤分期）。因此，将我们的方法扩展为采用依赖于观测协变量的脆弱性分布，可能是一个非常有价值的研究方向。

我们工作的可能拓展包括采用其他参数化基线风险分布，并结合其他风险结构（例如加性风险结构）。在癌症流行病学中，还有其他一些值得关注的量可由估计的超额风险推导得出，例如损失寿命年数等 [11]。因此，量化未观测异质性对这些量的估计所造成的影响，将具有研究价值。由于我们的真实数据应用涉及使用机密数据，因此无法公开这些数据。不过，我们创建了一个代码库（https://github.com/FJRubio67/IFNS），其中包含一个使用“The Simulacrum”数据集的应用。该数据集包含人工生成的类患者癌症数据。此代码库展示了所提出的方法和提供的软件。

据我们所知，这是首项在相对生存框架内评估 UIH 对估计量影响的研究。鉴于净生存用于国际比较和为政策制定提供依据，因此，仔细分析可能导致偏倚的因素（如 UIH）对该指标估计的影响是恰当的，而本文提出的方法为解决这一问题提供了一种工具。

财务披露

AB 获得英国癌症研究中心项目拨款资助（C7923/A29018）。

参考文献

[1] O. Aalen、O. Borgan 和 H. Gjessing。《生存与事件史分析：过程视角》。施普林格科学与商业媒体，纽约，2008年。[2] O.O. Aalen。使用复合泊松分布对生存分析中的异质性进行建模。《应用概率年鉴》，2(4):951–972，1992年。[3] O.O. Aalen。脆弱性在生存分析中的影响。《医学研究统计方法》，3(3):227–243，1994年。[4] O.O. Aalen、M. Valberg、T. Grotmol 和 S. Tretli。理解疾病风险的差异：难以捉摸的脆弱性概念。《国际流行病学杂志》，44(4):1408–1421，2014年。[5] C. Allemani、T. Matsuda、V. Di-Carlo、R. Harewood、M. Matz、M. Nikši´c、A. Bonaventure、M. Valkov、C.J. Johnson 和 J. Estève。全球监测癌症生存趋势（2000–14年，CONCORD-3）：分析来自71个国家322个基于人群的登记机构的个体记录，涵盖37 513 025名确诊患有18种癌症之一的患者。《柳叶刀》，391(10125):1023–1075，2018年。[6] D. Alvares 和 F.J. Rubio。一个易于处理的纵向数据与生存数据贝叶斯联合模型。《医学统计学》，40(19):4213–4229，2021年。[7] V. Bagdonavicius 和 M. Nikulin。《加速寿命模型：建模与统计分析》。Chapman and Hall/CRC，2001年。[8] T.A. Balan 和 H. Putter。聚类生存数据中的非比例风险与未观测异质性：何时能够区分二者？《医学统计学》，38(18):3405–3420，2019年。[9] T.A. Balan 和 H. Putter。脆弱性模型教程。《医学研究统计方法》，29:3424–3454，2020年。[10] A. Belot、H. Fowler、E.N. Njagi、M Luque-Fernandez、C. Maringe、W. Magadi、A. Exarchakou、M. Quaresma、A. Turculet、M.D. Peake、N. Navani 和 B. Rachet。英格兰非小细胞肺癌患者的年龄、剥夺程度及特定合并症与接受重大手术之间的关联：一项基于人群的研究。《胸腔》，74(1):51–59，2019年。[11] A. Belot、A. Ndiaye、M.A. Luque-Fernandez、D.K. Kipourou、C. Maringe、F.J. Rubio 和 B. Rachet。根据受众总结并传达生存数据：通过基于人群的癌症登记数据示例说明不同测量指标的教程。《临床流行病学》，11:53，2019年。[12] A. Burton、D.G. Altman、P. Royston 和 R.L. Holder。医学统计学中模拟研究的设计。《医学统计学》，25(24):4279–4292，2006年。[13] H. Charvat、L. Remontet、N. Bossard、L. Roche、O. Dejardin、B. Rachet、G. Launoy、A. Belot 和 CENSUR Working Survival Group。一种多层超额风险模型，用于在分层数据上估计净生存，并允许协变量产生非线性和非比例效应。《医学统计学》，35(18):3066–3084，2016年。[14] Y.Q. Chen 和 N.P. Jewell。关于一类一般半参数风险回归模型。《生物统计学》，88(3): 687–702，2001年。[15] L. Duchateau 和 P. Janssen。《脆弱性模型》。施普林格科学与商业媒体，2007年。[16] A. Eletti、G. Marra、M. Quaresma、R. Radice 和 F.J. Rubio。用于灵活超额风险建模的统一框架及其在癌症流行病学中的应用。《英国皇家统计学会会刊C辑：应用统计》，71:1044–1062，2022年。[17] M. Fauvernier、L. Roche、Z. Uhry、L. Tron、N. Bossard、L. Remontet 和“净生存估计挑战工作组”。含连续协变量的多维惩罚风险模型：用于研究癌症生存趋势与社会不平等。《英国皇家统计学会杂志：C辑（应用统计）》，2019年。[18] A. Gasparini、M.S. Clements、K.R. Abrams 和 M.J. Crowther。模型设定错误对共享脆弱性生存模型的影响。《医学统计学》，38:24477–4502，2019年。[19] J.J. Goeman、S. Le Cessie、R.J. Baatenburg De Jong 和 S.A. Van De Geer。利用疾病史预测生存：结合相对生存与脆弱性的模型。《荷兰统计学》，58(1):21–34，2004年。[20] D.D. Hanagal。《使用脆弱性模型建模生存数据》。施普林格·自然，2019年。

[21] R. Henderson 和 P. Oman。脆弱性对生存分析中边际回归估计的影响。英国皇家统计学会杂志：B辑（统计方法学），61(2):367–379，1999年。[22] P. Hougaard。由稳定分布导出的异质总体生存模型。《生物统计学》，73(2): 387–396，1986年。[23] P. Hougaard。生存数据的脆弱性模型。《寿命数据分析》，1(3):255–273，1995年。[24] N. Keiding、P.K. Andersen 和 J.P. Klein。脆弱性模型与加速失效时间模型在描述遗漏协变量所致异质性方面的作用。《医学统计学》，16(2):215–224，1997年。[25] P.C. Lambert 和 P. Royston。生存分析灵活参数模型的进一步发展。《Stata期刊》，9(2):265–290，2009年。[26] S. Müller、J.L. Scealy 和 A.H. Welsh。线性混合模型中的模型选择。《统计科学》，28(2): 135–167，2013年。[27] J. O'Quigley 和 J. Stare。包含脆弱性和随机效应的比例风险模型。《医学统计学》，21(21):3219–3233，2002年。[28] M.P. Perme、J. Stare 和 J. Estève。相对生存的估计。《生物统计学》，68(1):113–120，2012年。[29] M. Quaresma、J. Carpenter 和 B. Rachet。使用低秩薄板样条的灵活贝叶斯超额风险模型。《医学研究统计方法》，29(6):1700–1714，2020年。[30] P.M.V. Rancoita、M. Valberg、R. Demicheli、E. Biganzoli 和 C. Di Serio。肿瘤休眠与脆弱性模型：一种新方法。《生物统计学》，73(1):260–270，2017年。[31] V. Rondeau、Y. Mazroui 和 J.R. Gonzalez。frailtypack：一个R软件包，使用惩罚似然估计或参数估计，借助脆弱性模型分析相关生存数据。《统计软件杂志》，47(4): 1–28，2012年。[32] D. Rossell 和 F.J. Rubio。删失与模型误设下的加性贝叶斯变量选择。《统计科学》，38:13–29，2023年。[33] F. J. Rubio 和 R. Drikvandi。MEGH：一类用于聚类生存数据的参数化一般风险模型。《医学研究统计方法》，31:1603–1616，2022年。[34] F.J. Rubio、L. Remontet、N.P. Jewell 和 A. Belot。基于风险的回归模型的一般结构：在基于人群的癌症研究中的应用。《医学研究统计方法》，28:2404–2417，2019年。[35] F.J. Rubio、B. Rachet、B. Giorgi、C. Maringe 和 A. Belot。关于生命表分层不足时超额死亡风险估计模型的研究。《生物统计学》，22(1):51–67，2021年。[36] P.D. Sasieni。比例超额风险。《生物统计学》，83(1):127–141，1996年。[37] M.J. Stensrud、M. Valberg、K. Røysland 和 O.O. Aalen。利用因果脆弱性模型探究选择偏倚。《流行病学》，28(3):379–386，2017 年。[38] A. Tsiatis。竞争风险问题中的一个不可识别性方面。《美国国家科学院院刊》，72(1):20–22，1975年。[39] J.W. Vaupel 和 A.I. Yashin。异质性的诡计：选择对人口动态的一些意外影响。《美国统计学家》，39(3):176–185，1985年。[40] J.W. Vaupel、K.G. Manton 和 E. Stallard。个体脆弱性异质性对死亡率动态的影响。《人口学》，16(3):439–454，1979年。[41] A. Wienke。《生存分析中的脆弱性模型》。Chapman and Hall/CRC，2010年。[42] P.H. Zahl。超额风险的脆弱性建模。《医学统计学》，16:1573–1585，1997年。