基于汇总现况数据的发病率估计

Ralph Brinks

医学生物统计与流行病学教席，维滕/黑尔德克大学健康学院/医学院，德国维滕 D-58448

生物统计与流行病学研究所，德国糖尿病中心，德国杜塞尔多夫 D-40225

通讯作者：电子邮箱：
ralph.brinks@uni-wh.de

摘要
我们使用英国煤矿工人呼吸困难的历史数据，以及德国糖尿病的近期数据，来说明一种从聚合当前状态数据（即年龄别患病率数据）推导年龄特异性发病率的方法。该方法以一个微分方程为核心，并特别关注置信区间的最大似然（ML）估计。

关键词：患病率、慢性病、疾病-死亡模型、最大似然估计

引言

关于疾病发病率与患病率之间关系的问题，至少可以追溯到1934年。当时，Muench研究了能否根据有关南美洲黄热病年龄特异性患病率的横断面样本，估计其年龄特异性发病率[Mue34]。Muench使用了“求和数据”和“催化曲线”这两个术语，后者至今仍可见于现代传染病流行病学教材中 [Vyn10]。在统计学文献中，“聚合当前状态数据”这一术语比“求和数据”更常用。对于当前状态数据，通常假设有一个样本，其中每位研究参与者都有一个（大多为）二元疾病状态Y和一个时间变量T。时间变量T描述的是疾病状态Y所对应的时间点。例如，考虑这样一种情况：在大流行的不同阶段（T），对人群样本进行SARS-CoV-2抗体诊断检测（其中Y表示是否存在抗体）。关于当前状态数据的综述见 [Jew03, Jew04]。关于当前状态数据的教材式介绍，请参阅 [Sun06]。在流行病学中，当前状态数据通常会按不同年龄组进行汇总。例如，Muench关于黄热病的文章将数据表示为五个年龄组中的阳性病例数（c）和检测次数（n）：5-9岁、10-14岁、15-19岁、20-39岁以及40+ （[Mue34]中的表1）。

本文简要回顾慢性疾病的疾病-死亡模型（IDM）及其与当前状态数据的关系。聚合当前状态数据可以用慢性疾病的患病率表示，本文重点探讨年龄特异性患病率与疾病-死亡模型中转移率之间的数学关系。因此，本文延续了 Keiding 的综述文章 [Kei91] 的传统。不过，我们关注的是基于类似 Kolmogorov前向方程 [Bri18] 的微分方程的数学关系。随后，我们将这些微分方程与最大似然估计相结合，以便从聚合当前状态数据中估计发病率。利用患病率数据中的数学关系估计发病率的方法已在 [Bri13] 中阐述，并在 [Lan16] 中与其他方法进行了比较。迄今为止，在推断统计中，我们一直使用重抽样技术，参见例如 [Bri15]。本文采用最大似然方法，并特别着重于估计置信区间。

作为应用案例，我们使用一份关于英国煤矿工人呼吸困难的历史数据集 [Ela14]，以及德国2型糖尿病的近期数据 [Tam16] ，来估计聚合当前状态数据中的年龄特异性发病率。

疾病死亡模型

近期研究表明，在某个时点 t，慢性疾病的年龄特异性患病率 p 与年龄特异性发病密度 i（同义词：发病率）以及死亡率 m_0 和 m_1 之间存在偏微分方程 (PDE) 关系 [Bri14, Bri15a]。图1展示了基础的多状态模型及其可能的转移和相应的率。在疾病-死亡模型（IDM）中，患病率 p 和率 i、m_0, 和 m_1 通常取决于日历时间 t 和年龄 a。例如，p(t, a) 和 i(t, a) 分别表示在时点 t 年龄为 a 的人群中，存活且患有该病的比例以及发病率。在流行病学背景下，日历时间 t 有时也称为时期 [Cla87]。

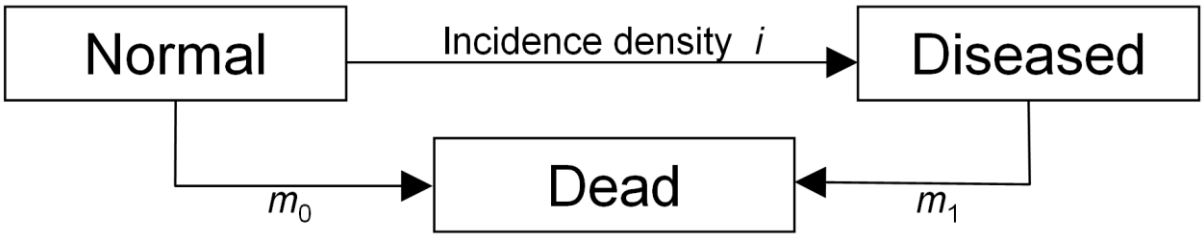

图1：慢性疾病（“患病”）的疾病-死亡模型及相关转移密度：发病密度 i、无疾病时的死亡率（m_0）和患病时的死亡率（m_1）。

将图1中疾病-死亡模型的转移率 i、m_0, 和 m_1 与年龄特异性患病率 p 联系起来的偏微分方程为

$$(\partial p/\partial t + \partial p/\partial a) = (1 - p)\ \{i - p\ (m_1 - m_0)\}, \tag{1}$$

其中，∂/∂x 表示对 x 求偏导数。

最大似然估计

假设我们有按年龄组汇总的当前状态数据，年龄组以 k 编号，k = 1，.....，K；年龄组 k 中有 c_k 人报告患有慢性疾病，该年龄组总人数为 n_k。那么，聚合当前状态数据的二项似然函数 L 为

$$L = \prod_{k=1}^{K} \binom{n_k}{c_k} p_k^{c_k} (1 - p_k)^{n_k - c_k} \tag{2}$$

其中，k 是年龄组的编号，p_k 是（未知的）患病率，即年龄组 k 中患有相关病症的受试者所占比例。

如果方程（2）中的患病率 p_k，k = 1，.....，K，可以表示为率 i、m_0，和 m_1，的函数，即 p = p(i, m_0, m_1)，那么参数估计就会很直接；我们将在英国煤矿工人呼吸困难的数据中对此加以说明。表1显示了观察到的呼吸困难聚合当前状态数据。为纳入患病者与未患病者之间死亡率的差异，我们还将使用死亡率数据（表2）。

通常，在流行病学背景下无法获得死亡率 m_0 和 m_1 。更常见的情况是，能够获得总体死亡率 m 和死亡率比 R = m_1/m_0 。在这种情况下，患病率 p 没有解析表示 p = p(i, m, R)，其具体解释将在后文给出。随后，可以通过代入估计量估计发病率，这将在表3所示的2型糖尿病数据中演示。

数据集

表1展示了关于英国煤矿工人呼吸困难的观察到的聚合当前状态数据 [Ela14]。

年龄组 k（以年为单位）	人数 观察到的（n_k）	患有.....的人数 病症（c_k）
20至24岁	1952	16
25至29岁	1791	32
30至34岁	2113	73
35至39岁	2783	169
40至44岁	2274	223
45至49岁	2393	357
50至54岁	2090	521
55至59岁	1750	558
60至64岁	1136	478

表1：英国煤矿工人呼吸困难数据（**[Ela14]**中的表14.2a）。

之后，我们还会考虑纳入死亡率。因此，我们也使用了[Ela14]中列出的生命表。

年龄组 k（以年为单位）	生命表 常规人口	患有.....的人口 呼吸困难
20至24岁	481185	343937
25至29岁	478683	333343
30至34岁	476150	320446
35至39岁	472641	304305
40至44岁	467066	284325
45至49岁	457729	260806
50至54岁	441895	233060
55至59岁	415262	200561
60至64岁	372908	163241

表2：英格兰和威尔士的生命表（**[Ela14]**中的表14.2b）。

为展示一种基于代入估计量估计发病率的替代方法，我们使用表3所示的数据。该表展示了德国法定医疗保险中女性在2009年和2010年确诊2型糖尿病的聚合当前状态数据。有关这些数据的详细信息见[Tam16]。由于严格限制使用原始数据，我们在原始数据中加入了随机噪声（2%）；随后将数据下采样100倍，并四舍五入为最接近的整数。

年龄组	女性 2009年		女性 2010年	
k（以年计）	观察到的（n_k）	患有糖尿病（c_k）	观察到的（n_k）	患有糖尿病（c_k）
20至24岁	19029	33	18939	36
25至29岁	19549	65	18917	69
30至34岁	19391	109	19388	122
35至39岁	20885	200	19722	208
40至44岁	28844	402	26543	403
45至49岁	28856	706	29509	742

50-54	25641	1145	25870	1178
55-59	23223	1826	23238	1850
60-64	18845	2134	20112	2423
65-69	21964	3160	19714	2887
70-74	22965	4281	23452	4446
75-79	15944	3628	16509	3909
80-84	13310	3114	13083	3295
85-89	8796	2159	8637	2220
90-94	2380	569	2760	681
95-99	892	188	833	177

表3：2009年德国女性的糖尿病法定健康保险数据和**2010 [**Tam**16]**。

年龄特异性发病率的估计我们区分患有和未患有该疾病的人群所处的情况

是否具有不同的死亡率。后者称为非差异性死亡率[Kei91]。

非差异性死亡率

首先，我们考虑没有迁移且没有差异死亡率的情况（即 $m_0 = m_1$），此时偏微分方程 (1) 变为

$$(\partial p/\partial t + \partial p/\partial a) = (1 - p)\, i. \tag{3}$$

此外，若假设方程 (1) 中的患病率 p 不随时间 t 变化，则偏微分方程 (1) 变为常微分方程（ODE）

$$\partial_a p := \mathrm{d}p \,/\, \mathrm{d}a = (1 - p)\, i, \tag{4}$$

其通解为

$$p(a) = 1 - (1 - p_0) \times \exp\left(- \int_{a_0}^{a} i(\tau)\, d\tau \right) \tag{5}$$

其中，$p_0 = p(a_0)$ 是初始条件。

方程 (4) 是年龄特异性发病密度 i 的一个直接估计量的基础。若能根据年龄特异性患病率 p 估计导数 $\partial_a p$，则可得到 $i = \partial_a p \,/(1 - p)$。

例1：我们以经 logit 转换的患病率作为结果变量，拟合线性回归模型 $\mathrm{logit}(p(a)) = \log(p(a)/\{1 - p(a)\}) = \beta_0 + \beta_1 \times a$，其中 a 取年龄组中点 = 22.5、27.5、...、62 5 1 β_0 = 7 02 β_1 = 0 110 = β_0 .，并使用表 中的数据。 + β_1 = ∂_a 1 = β_1 我们得到 – . 和 . 。将 p(a) expit(a) 代入 i p/(– p)，得到 i(a) p(a)。expit 函数是 logit 函数的反函数，即 $(\mathrm{logit})^{-1} = \mathrm{expit} = \exp/(1+\exp)$，其导数为 expit′ = expit . (1 – expit)。对应的年龄特异性发病率 $i(a) = \beta_1$ expit($\beta_0 + \beta_1$ a)，其中 $\beta_0 = -7.02$、$\beta_1 = 0.110$，在图 2 中以黑线表示。作为比较，蓝线显示了 [Ela14]给出的估计值。注意，[Ela14]在 57.5 岁时给出了负估计值。

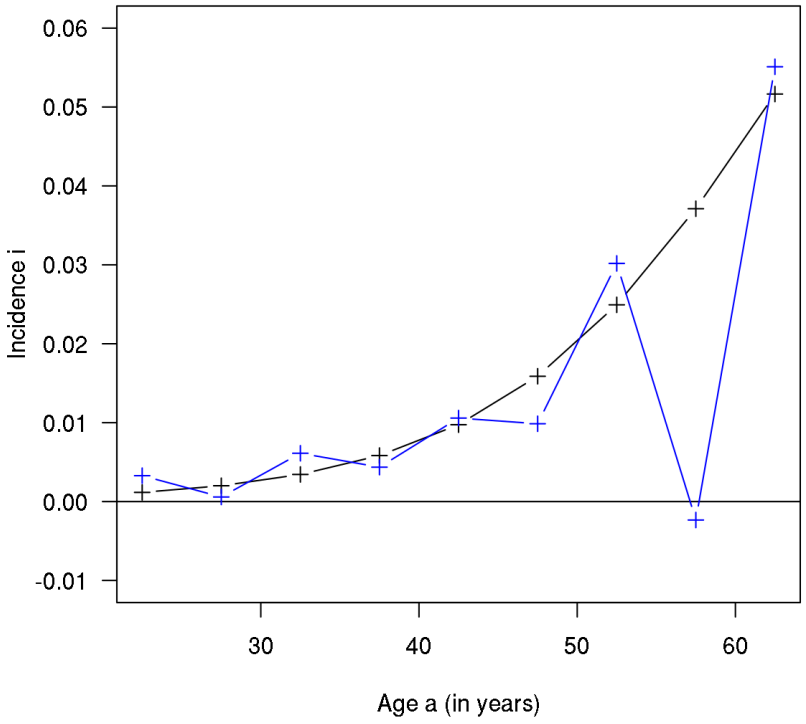

图 2：例1中根据表1数据计算得出的年龄特异性发病率估计值。

例1展示了如何利用微分方程理论，根据慢性疾病的聚合当前状态数据估计年龄特异性发病率。由于我们还希望估计年龄特异性发病率的置信区间，因此选择了最大似然方法。

结合图 2，我们发现年龄特异性发病密度 i(a) 似乎随年龄呈指数增长。因此，我们采用 $i(a) = \exp(\beta_0 + \beta_1 \cdot a)$ 的形式，系数为 β_0 、 β_1。将这一发病率代入方程（ 5 ），并设初始条件为 $a_0 = 20$ 且 $p(20) = p_0 = 0$，可得 $p(a) = 1 - \exp(h(20) - (a))$，其中 $h(z) = \exp(\beta_0 + \beta_1 \cdot z)/ \beta_1$。接着，在年龄组中点 $a_k = 22.5$、.....、62.5 处计算 p，并最终将其代入方程（2）。由此得到似然函数 L = L(β_0, β_1)，通过优化该函数，可计算出两个系数 β_0 和 β_1的最大似然估计量。此外，若根据费希尔信息矩阵的逆矩阵推导出的标准误，计算方差—协方差矩阵 [Woo15]，的大样本近似下的 95%置信区间，则可得到表 3所示的结果。

	点估计	95%置信区间
β_0	–7.8237	–8.0590 至 –7.5883
β_1	0.07559	0.07006 至 0.08111

表4：用于无差异死亡率情况下年龄特异性发病率参数化的系数 β_0 和β_1 的极大似然估计量。

差异死亡率

在差异死亡率的情况下，我们必须区分给定的是哪一种死亡率。在下一个示例中，我们已知患病者的死亡率（m_1）以及总体人群中的总体死亡率（m）。随后，我们考虑已知总体死亡率 m 和死亡率比 $R = m_1/m_0$ 的情况。

患病者死亡率与总体死亡率

从方程（1）出发，并利用总体死亡率 m 是 m_0 和 m_1, $m = (1 - p)\, m_0 + p\, m_1$, 的凸组合这一事实，我们得到

$$(\partial/\partial t + \partial/\partial a)p = i - p(m_1 - m + i). \tag{6}$$

同样，我们假设方程（6）不依赖于时间 t，并得到常微分方程

$$\partial_a p = i - p\,(m_1 - m + i) \tag{7}$$

其通解为

$$p(a) = \exp(-G(a)) \times \left\{ p_0 + \int_{a_0}^{a} i(\tau) \times \exp(G(\tau))d\tau \right\} \tag{8}$$

with $_{a}=\int\{ - + \}_{a}$ 聚合当前 且 $p_0 = p(a_0)$ 为初始条件。
状态数据中的发病率估计。)( )
($_1$ τ τ

总体死亡率 m 和呼吸困难的矿工的死亡率 m_1 根据表2所示的生命表进行估计。具体做法是，利用单减过程理论（例如，见 [Pre01] 中的第3章），将表2第二列和第三列中的五年死亡概率转换为年度死亡率 m 和 m_1。

与上一节类似，我们令发病率 $i(a) = \exp(\beta_0 + \beta_1 \cdot a)$，其系数为$\beta_0$、 β_1。若将该发病率代入方程 (8)，初始条件为 $a_0 = 20$ 且 $p(20) = p_0 = 0$。针对式 (2) 中的似然函数，两个系数 β_0 和 β_1 的最大似然估计量得出了表4所示的结果。

	点估计	95%置信区间
β_0	–8.4706	–8.7296 至 –8.2116
β_1	0.10107	0.094791 至 0.10734

表5：用于对差异死亡率下的年龄特异性发病率进行参数化的系数 β_0 和 β_1 的极大似然估计量。

总体死亡率和死亡率比

根据表3中的数据，我们设定方法 $p(t, a) = \text{expit}(\beta_0 + \beta_1 \cdot t + \beta_2 \cdot a + \beta_3 \cdot a2)$，并计算系数 $\beta_0, \ldots, \beta_3$ 的极大似然估计量。我们假设已知德国女性的总体死亡率 m。此外，假设死亡率比 $R = m_1/m_0$ 已知。给定 $m = (1 - p)\, m_0 + p\, m_1$ 和 $R = m_1/m_0$, , 偏微分方程 (1) 可写为

$$\left(\partial/\partial t+\partial/\partial a\right)p=(1-p)\left\{i-m\frac{p\left(R-1\right)}{1+p\left(R-1\right)}\right\}, \tag{9}$$

由此可求解发病率

$$i=\frac{\left(\partial/\partial t+\partial/\partial a\right)p}{1-p}+m\frac{p\left(R-1\right)}{1+p\left(R-1\right)} \tag{10}$$

对于 m 和 R，我们分别取自德国联邦统计局和丹麦糖尿病登记册 [Car19], 。

找到 p 的最大似然估计后，公式（10）通过插件原理 [Was04] 得出发病率 i 的最大似然估计。通过使用德尔塔法，将 p 的标准误转换为 i 的标准误，从而计算 i 的置信区间 [Sha03]。年龄特异性发病率及其 95% 置信区间的估计值如图 3 所示。在较高年龄段，置信区间的长度会增加。

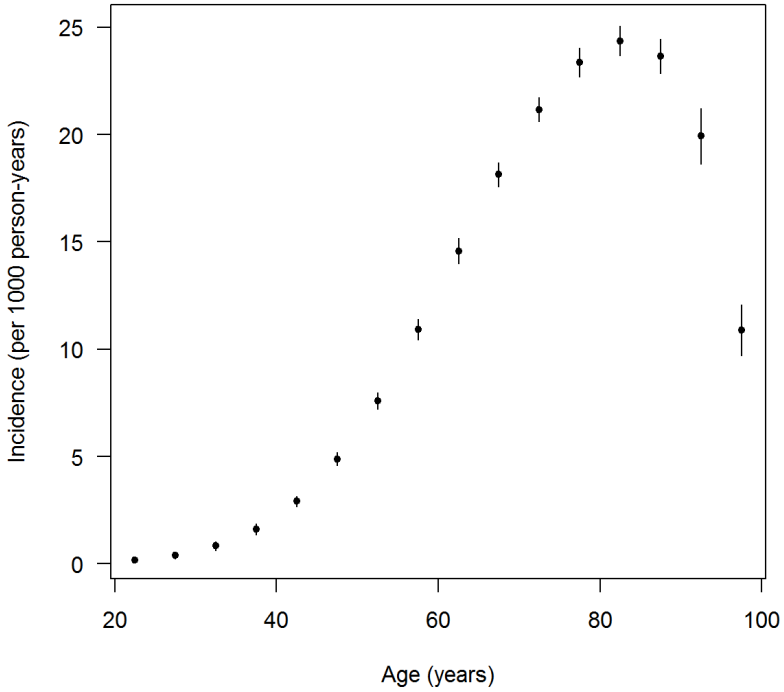

图3：使用公式（9）的代入估计所估计的女性糖尿病年龄特异性发病率。竖线表示95%置信界限。

讨论

在本研究中，我们描述了一种关于微分方程统计推断的新方法，该微分方程将慢性疾病的年龄特异性患病率与其潜在的年龄特异性发病率联系起来。该方法可以利用聚合当前状态数据（年龄别患病率数据）估计年龄特异性发病率。该微分方程的渊源至少可以追溯到[Bru99]。除估计发病率外，该微分方程还被用于其他应用，例如根据患病率和发病率估计死亡率[Toe18], 、预测慢性疾病患者人数[Voe22] ，或估计卫生政策的影响 [Toe21]。一个具有历史意义的流行病学应用，是根据患病率估计发病率。在存在差异死亡率的情况下，如果已知一般人群的死亡率以及患病者相对于未患病者的死亡率比，便可以实现这一目标。

如果已知患病者相对于未患病者的死亡率比。总体死亡率可能来自国家统计局，死亡率比则可以从其他环境中转用，例如从其他国家获取。这是可行的，因为率比在人群[Bre80]中作为关联性指标具有很强的稳定性。迄今为止，尚无关于该微分方程统计推断的理论，因此我们必须采用重抽样技术来获得置信界限。在重抽样过程中，根据所报告的输入参数分布抽取一定数量的随机样本，以估计输入参数中的不确定性如何通过微分方程传播并影响结果。例如，[Bri15]中介绍了这种重抽样方法。

由此引出一个问题：本文所述方法能否以及如何加以推广。在涉及差异死亡率的示例中，可以考虑这样一种情况：不仅患病率，关于死亡率的信息也仅在统计不确定性下已知。在这种情况下，发病率的标准误需要更复杂的误差传播概念，例如影响函数 [Was06] 或随机变量的泰勒级数近似 [Wol07]。推广本文发现的另一种方式涉及患病率 p 的函数形式。这样做可以拟合更复杂的患病率数据，例如时间 — 年龄交互作用。在此类情况下，将 p(t, a) 写成 expit 函数 p(t, a) =expit(f(t, a)) 的形式，其中 f 是任意可微函数（例如关于 t 和 a 的多项式），即可立即得到发病率 i 的代入估计 = p (∂f + $m_1 - m_0$)，其中 ∂f = ∂f/∂t + ∂f/∂a。这是因为 expit 函数的导数为 expit’ = (1 – expit) · expit，并将 p = expit(f) 代入式（1）即可得出。

在非慢性疾病中可以看到一个重要的推广。迄今为止，我们只考虑了不会缓解的疾病。如果图1中存在一条从患病状态回到正常状态的路径，缓解率为 r，那么偏微分方程 (1) 可改写为 [Bri15b]:

$$(\partial/\partial t + \partial/\partial a)p = (1-p)\{i - p(m_1 - m_0)\} - r\,p\,. \qquad (11)$$

在非差异性死亡率的情况下，偏微分方程 (11) 是线性的，可以像煤矿工人的例子那样求得解析解。如果 $m_1 \neq m_0$, ，则可以进行类似于糖尿病示例的分析。然而，无论是差异性死亡率还是非差异性死亡率，要推断发病率 i，都需要有关缓解率 r 的额外信息（或假设）。尽管黄热病可能缓解，但在引言中的示例里，Muench 隐含地假设感染后阳性的血清学状态不会改变。对其他非慢性疾病也可以作类似的考量。

附录

资金声明

作者未因本研究的任何方面获得任何资金支持。

利益冲突

作者声明，其在本研究的任何方面均不存在利益冲突 .

参考文献

[Bre80] Breslow NE，Day NE：《癌症研究中的统计方法》，第1卷《病例对照研究分析》，国际癌症研究机构（IARC），里昂，1980年

[Bri13] Brinks R, Landwehr S, Icks A, Koch M, Giani G. 利用常微分方程根据患病率推导年龄特异性发病率。《统计医学》. 30;32(12):2070-8, 2013 doi: 10.1002/sim.5651.

[Bri14] Brinks R, Landwehr S：非传染性疾病患病率的年龄和时间依赖模型及其在德国痴呆症研究中的应用。《理论人口生物学》，92:62-68, 2014

[Bri15a] Brinks R, Landwehr S. 慢性病患病率与发病率之间的新关系。《数学医学与生物学》：IMA 期刊，32(4), 425-435, 2015, DOI 10.1093/imammb/dqu024

[Bri15b] Brinks R, Landwehr S (2015) 二分变量的变化率与患病率：模拟与应用。《PLOS ONE》10(3): e0118955. DOI 10.1371/journal.pone.0118955

[Bri15] Brinks R、Hoyer A、Kuss O、Rathmann W：德国城市地区主动出行增加对2型糖尿病风险的预计影响。《PLOS ONE》10(4)：e0122145，2015。https://doi.org/10.1371/journal.pone.0122145

[Bri18] Brinks R、Hoyer A。疾病-死亡模型：统计学视角与微分方程。《寿命数据分析》24(4), 743-754，2018，DOI 10.1007/s10985-018-9419-6

[Bri21a] Brinks R：Skinfaxi项目的统计分析，Zenodo存储库 DOI 10.5281/zenodo.4816749 https://doi.org/10.5281/zenodo.4816749

[Bri21b] Brinks R：基于患病率估计年龄特异性发病率：一种新颖的马尔可夫链蒙特卡罗方法, Zenodo存储库 DOI 10.5281/zenodo.4770055 https://doi.org/10.5281/zenodo.4770055

[Bru99] Brunet RC, Struchiner CJ：一种非参数方法，用于根据不可逆疾病的患病率数据重建年龄和时间依赖性发病率，并考虑差异死亡率。《理论人口生物学》56(1)：76-90，1999年。

[Car19] Carstensen 10.1136/bmjdrc-2019-001071

[Chu10] Chubb MC, Jacobsen KH：数学建模与流行病学研究过程。《欧洲流行病学杂志》，25(1), 13-19，2010年

[Cla84] Clayton D, Schifflers E. 癌症发病率时间变化模型。I：年龄–时期模型和年龄–队列模型。《医学统计学》，6(4)，449-467，1987年

[Ela14] Elandt-Johnson RC, Johnson NL：生存模型与数据分析。约翰·威立父子出版公司，2014年

[Kei91] Keiding N，年龄特异性发病率与患病率：统计学视角，《英国皇家统计学会会刊 A辑》154：371-412，1991

[Jew03] Jewell NP，van der Laan M。当前状态数据：综述、近期进展与未解决的问题。《统计学手册》23：625-642，2003，DOI 10.1016/S0169- 7161(03)23035-2

[Lan16] Landwehr S, Brinks R. 基于患病率的发病率估算技术的比较研究及其在德国痴呆数据中的应用。《统计医学》28；35(5)：768-81，2016年，DOI：10.1002/sim.6736

[Mue34] Muench H：通过催化曲线从求和数据推导率，《美国统计学会会刊》29 (185)：25-38，1934年

[Pre01] Preston SH, Heuveline P, Guillot M.《人口学》，Blackwell Publishing，马尔登（马萨诸塞州），2001年

[Tam16] Tamayo T, Brinks R, Hoyer A, Kuss O, Rathmann W. 德国糖尿病的患病率和发病率：对2009年和2010年6500万人法定医疗保险数据的分析。《德国国际医学杂志》，113(11)，177，2016年

[Toe18] Tönnies T, Hoyer A, Brinks R. 2012年德国确诊2型糖尿病患者的超额死亡率——基于7000万德国人的理赔数据进行的估计。《营养、代谢与心血管疾病》28(9):887-891，2018年

[Toe21] Tönnies T, Heidemann C, Paprott R, Seidel-Jacobs E, Scheidt-Nave C, Brinks R, Hoyer A. 估算税收政策干预对德国2020年到2040年2型糖尿病成年人的预测人数和患病率的影响。《BMJ开放糖尿病研究与护理》9(1):e001813，2021年

[Sha03] Shao J：《数理统计》，Springer New York，2003年

[Sun06] Sun J.《区间删失失效时间数据的统计分析》。纽约：Springer，2006年

[Vyn10] Vynnycky E, White RG：《传染病建模导论》，牛津大学出版社，2010年

[Was04] Wasserman L：《统计推断简明课程》，Springer New York，2004

[Voe22] Voeltz D、Tönnies T、Brinks R、Hoyer A。《2型糖尿病未来患病率——慢性病预测方法的比较分析》。《PLOS ONE》7;17(3):e0264739，2022

[Was06] Wasserman L：《非参数统计学》，Springer New York，200 6

[Wol07] Wolter KL：《方差估计导论》，Springer New York，2007

[Woo15] Wood SN，《核心统计学》，剑桥大学出版社，2015

[Xun13] Xun X, Cao J, Mallick B, Maity A, Carroll RJ：偏微分方程模型的参数估计。美国统计学会会刊，108(503)，1009-1020，2013