·应用研究·

COVID-19 潜伏期分布估计的统计学方法比较*

刘　裕[1,2]　卓冰婷[2]　陈俊宏[2]　杜志成[2]　郝元涛[3△]

【提　要】 **目的**　回顾和评估 COVID-19 潜伏期分布估计的统计学方法，为有效、快速、准确地收集和分析潜伏期数据提供参考和借鉴。**方法**　利用 COVID-19 疫情早期发表的数据，比较分析单区间删失、双区间删失和随机过程三类方法不同分布假设下获得的 COVID-19 潜伏期分布最大似然估计和贝叶斯估计。**结果**　同类方法不同分布假设间，非参数方法要比参数方法拟合效果更好，但非参数方法存在较多的跳跃点，且无法获得估计的 95%置信区间；同类方法相同分布假设条件下，最大似然估计与贝叶斯估计结果和拟合效果相近；同类方法的对数正态假设条件下获得的潜伏期分布的大分位数（>90%分位数）可能较大地偏离非参数估计结果；从数据利用的角度，双区间删失方法对数据的利用率最高；由于数据收集和利用的差异，不同方法得到的潜伏期分布估计可能存在较大差异。**结论**　采用双区间删失观测的参数模型获取传染病潜伏期分布的最大似然估计，可提高数据的收集、利用和分析效率；仔细比较不同分布假设下参数模型和非参数模型的结果，并谨慎解释潜伏期大分位数的估计结果，将有利于作出正确的防控决策。

【关键词】 潜伏期　分布估计　区间删失数据　COVID-19

【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2023.04.013

传染病的潜伏期是指宿主首次暴露于传染源到其首次出现疾病相关临床表现（体征或症状）的时间间隔[1]。掌握潜伏期分布对病例的定义、传染源的追溯、接触者追踪随访期和隔离期的设置、入境筛查隔离策略的制定、无症状感染人群医学观察期的确定等，以至疫情规模和传播潜力的测算，都具有重要意义[2-4]。

然而，潜伏期分布的准确估计并非易事，我们以新型冠状病毒感染（COVID-19）为例加以说明。首先，COVID-19 的感染暴露时间无法直接观测，往往只能知道感染暴露是在某个时间段发生的，也就是说，它是一种区间删失观测（interval censored data）[5]。这也是尽管截至 2020 年 1 月 22 日已报告 422 例 COVID-19 确诊患者，但 Linton 等[6] 只纳入 10 例具有明确暴露日期和发病日期的数据预估 COVID-19 潜伏期分布的可能原因。其次，感染以后患者出现症状的时间经常不能准确回忆，也就是说，患者出现症状（即发病）的时间也可能是区间删失观测。由于 COVID-19 疫情发生在冬季，疫情早期人们对该病知之甚少，其临床症状与呼吸道感染重叠，COVID-19 确诊患者对首次出现新型冠状病毒（SARS-Cov-2）感染症状的回忆往往摸棱两可。此时，调查得到的暴露感染和症状出现时间经常都是区间删失的情况，即我们获得的是双区间删失观测（doubly interval censored data）[7]。再者，对区间删失尤其是双区间删失观测数据的潜伏期分布估计远比精确观测复杂，结果稳定性也可能更差[8]。自 2019 年 12 月发生 COVID-19 疫情以来，研究人员采用不同的统计学方法分析各自收集的数据估计 COVID-19 的潜伏期分布，得到其潜伏期中位数在 4.0 天到 7.8 天之间[6,9-13]，相差较大。而他们获得的 COVID-19 潜伏期大分位数估计差异更大，这体现在对潜伏期超过 14 天的患者比例的估计。例如，Bi 等[9] 估计潜伏期超过 14 天的 COVID-19 患者在 5%左右，而 Qin 等[12] 的结果显示这个数值超过 10%。

分布假设和估计方法对潜伏期的分布估计具有深刻影响[14]。为了加深传染病潜伏期分布的理解，提升潜伏期分布监测中数据收集和利用的效率，我们有必要对现有的分析模型进行评估。本研究旨在综述潜伏期分布估计方法，采用 COVID-19 疫情早期 Lauer 等[11] 收集的数据对这些方法进行比较，以期为有效、快速、准确地预估潜伏期分布提供参考和借鉴。

资料与方法

1.数据收集

本文数据来源于 Lauer 等[11] 对 COVID-19 潜伏期分布估计的早期研究，该研究纳入 2020-01-04 至 2020-02-24 中国湖北以外确诊的 181 例 COVID-19 患者，这些患者的基本信息以及感染暴露和症状出现的时间区间均可从网络新闻或公共卫生报告中获取。

2.统计学方法

考虑包含 n 个独立样本的研究，假设样本 $i(i=1,2,\cdots,n)$ 感染暴露和出现症状的时间分别为 E_i 和 O_i（$O_i>E_i$），则该样本的潜伏期为 $T_i=O_i-E_i$。然而，在实践中我们往往只知道感染暴露或症状出现落在某个可能

* 基金项目：广东省基础与应用基础研究基金（2021A1515011591）；广东省医学科学技术研究基金（A2021104）

1.广州中医药大学公共卫生与管理学院（510006）
2.中山大学公共卫生学院医学统计系
3.北京大学公众健康与重大疫情防控战略研究中心
△通信作者：郝元涛，E-mail：haoyt@ bjmu.edu.cn

的区间，也就是说，我们一般获取如下形式的双区间删失观测（图 1）：

$X_i=\{(E_{iL},E_{iR}],(O_{iL},O_{iR}]\}$

其中，$E_i\in(E_{iL},E_{iR}]$，$O_i\in(O_{iL},O_{iR}]$，而且，$E_{iL}\leqslant E_{iR}$，$O_{iL}\leqslant O_{iR}$；特别地，当区间的左端点与右端点相等时（$E_{iL}=E_{iR}\triangleq E_i$ 或 $O_{iL}=O_{iR}\triangleq O_i$），表示观测到的是确切的感染暴露或症状出现时间。如果能够获取 E_i 或 O_i 确切的观测时间，则

$T_i\in(O_{iL}-E_i,O_{iR}-E_i]$，$T_i\in(O_i-E_{iR},O_i-E_{iL}]$ 或 $T_i=O_i-E_i$。

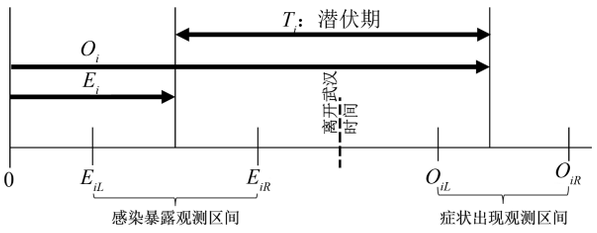

图 1 潜伏期观测数据示意图

我们关注的是潜伏期 T_i 的分布 $F(t)$，记 $S(t)=1-F(t)$ 为 T_i 的生存函数。以下简述基于区间删失观测的潜伏期分布估计方法（表 1）。

表 1 潜伏期估计方法汇总

分析方法	基本假设	统计模型	潜伏期分布假设	估计方法
单区间删失方法	症状出现时间都是确切已知的，而感染暴露时间属区间删失	–	–	NPMLE
		$\ln(T_i)=\mu+\sigma\cdot W_i$，其中，$\mu$ 是截距项，σ 是尺度参数，W_i 是随机误差	对数正态分布 威布尔分布 伽马分布	MLE 或 Bayes 估计
双区间删失方法	感染暴露时间和出现症状的时间都属区间删失	–	–	NPMLE
		同“单区间删失方法”的参数模型	对数正态分布 威布尔分布 伽马分布	MLE 或 Bayes 估计
随机过程方法*	病例是在武汉感染并在离开武汉后出现症状，更新过程达到稳态	$T_i=A_i+V_i$，A_i 感染到离开武汉的时间间隔，V_i 离开武汉到出现症状的时间间隔	对数正态分布 威布尔分布 伽马分布	MLE

*：将病毒感染到出现症状的整个过程视为更新过程，考察“离开武汉”这一删失事件截断的港伏期。PMLE：非参数最大似然估计；MLE：最大似然估计

（1）单区间删失方法

假定所有样本的症状出现时间都是已知的，即对任意 i，$O_{iL}=O_{iR}\triangleq O_i$，此时，$T_i\in(O_i-E_{iR},O_i-E_{iL}]\triangleq(T_{iL},T_{iR}]$。这样，潜伏期 T_i 的分布估计就简化为单个区间删失数据的分析。令 $\{s_j\}$ $m_j=0$ 为 $\{0,T_{iL},T_{iR}:i=1,2,\cdots,n\}$ 的唯一有序排列；记 $\alpha_{ij}=I(s_j\in(T_{iL},T_{iR}])$（$I$ 是示性函数），$p_j=F(s_j)-F(s_{j-1})$，则似然函数可以表示为：

$$L_F=\prod_{i=1}^{n}[F(T_{iR})-F(T_{iL})]=\prod_{i=1}^{n}\sum_{j=1}^{m}\alpha_{ij}p_j$$

满足 $\sum_{j=1}^{m}p_j=1$ 且 $p_j\geqslant 0(j=1,\cdots,m)$。最大化上述似然函数即可获得 $F(t)$ 的非参数最大似然估计（non-parametric maximum likelihood estimator，NPMLE）。这可以通过各种算法来实现，最经典的是 Turnbull 的自相合算法[15]。

如果假定潜伏期 T_i 服从某种特定的分布（如对数正态分布、威布尔分布或伽马分布），且具有对数线性模型的形式：$\ln(T_i)=\mu+\sigma W_i$，其中，μ 是截距项，σ 是尺度参数，W_i 是随机误差，则 T_i 的分布函数可以参数化为 $F(t)\triangleq F_\theta(t)$，而生存函数 $S(t)=P(T_i>t)=P(e^{\mu+\sigma W_i}>t)=P(e^{\mu}>t\,e^{-\sigma W_i})\triangleq S_0(t\,e^{-\sigma W_i})$，从而具有加速失效时间模型（accelerated failure time model，AFT model）[16] 的形式。此时，我们可以最大化似然函数：$L_{F\theta}=\prod_{i=1}^{n}\int_{T_{iL}}^{T_{iR}}f(t)\,dt=\prod_{i=1}^{n}[F_\theta(T_{iR})-F_\theta(T_{iL})]$，获得 $F_\theta(t)$ 参数的最大似然估计，从而得到潜伏期的分布估计；也可以假定参数 θ 服从某种形式的先验分布 π_θ（如均匀分布），则其后验分布正比于 $\pi_\theta\cdot L_{F_\theta}$，通过模拟算法（如马尔科夫链蒙特卡罗方法，Markov Chain Monte Carlo，MCMC）获得 θ 的后验估计，从而得到潜伏期分布的贝叶斯估计。

（2）双区间删失方法

对于感染暴露时间和出现症状的时间都是区间删失的情形，即观测值是双区间删失数据的情况，记 $\{u_1<u_2<\cdots<u_r\}$ 和 $\{v_1<v_2<\cdots<v_s\}$ 分别为 E_i 和 T_i 的所有可能取值的有序排列。定义 $e_j=P(E_i=u_j)$ 及 $f_k=P(T_i=v_k)$，则 $\sum_{j=1}^{r}e_j=\sum_{j=1}^{s}f_k=1$。令 $\alpha_{jk}^{i}=I(u_j\in(E_{iL},E_{iR}],u_j+v_k\in(O_{iL},O_{iR}])$，则似然函数为 $L_F=\prod_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{s}\alpha_{jk}^{i}e_jf_k$。最大化似然函数 L_F，就可得到 $\{f_k\}$ $s_j=1$（即 $F(t)$）的 NPMLE 估计。De Gruttola 等[17] 首先给出了该估计的自相合算法，但由于计算效率比较低，Sun[18] 改进了该算法。

类似地，如果假定潜伏期 T_i 服从某种特定的分布且可以表示成上述线性模型的形式，则我们同样可以通过 AFT 模型来刻画潜伏期的分布。令 $\delta_i=I(E_{iL}<E_{iR},O_{iL}<O_{iR})$，$\omega_i=I(\{E_{iL}=E_{iR},O_{iL}<O_{iR}\}\ or\ \{E_{iL}<E_{iR},O_{iL}=O_{iR}\})$，即 $\delta_i=1$、$\omega_i=1$ 和 $(1-\delta_i)(1-\omega_i)=1$ 分别表示观测数据是双区间删失、单区间删失和确切观测的

情形，此时，似然函数为

$$L_{F_\theta} = \prod_{i=1}^{n} \left[\int_{E_{iL}}^{E_{iR}} \int_{O_{iL}}^{O_{iR}} g_\varphi(e) f_\theta(o-e)\, dode\right]^{\delta_i} \times \left[F_\theta(T_{iR}) - F_\theta(T_{iL})\right]^{\omega_i} \times \left[f_\theta(T_i)\right]^{(1-\delta_i)(1-\omega_i)}$$

其中，g_φ和f_θ分别为感染暴露时间和潜伏期的概率密度函数，φ 和 θ 分别为各自的分布参数。通常，假定感染暴露时间在观测区间（E_{iL}，E_{iR}］均匀分布，这样，我们最大化似然函数就可以获得潜伏期分布参数 θ 的 MLE 估计，从而得到潜伏期分布的估计。与单区间删失方法一样，我们也可以通过贝叶斯方法获得潜伏期的分布估计。

（3）随机过程方法

在 COVID-19 流行早期，有大量无症状者离开武汉并在武汉以外确诊的病例。假设这些病例在离开武汉之前已被感染，则离开武汉与出现症状之间这段时间就是对他们潜伏期的删失观测（censored observations，图 2）。Qin 等[12]将感染暴露开始的 COVID-19 疾病发展视为随机过程。具体地，对纳入的任一病例 i，从感染到出现症状的整个过程可视为更新过程（renewal process）；将该病例从感染暴露到其离开武汉之间的时间间隔A_i看作一个更新过程的后向复发时间（backward recurrence time），而将该病例离开武汉到其出现症状之间的时间间隔V_i看作是这个更新过程的前向复发时间（forward recurrence time）。显然，V_i容易观测，而A_i回忆偏移往往较大。对单个病例，A_i和V_i未必相同；而当更新过程达到稳态（equilibrium status）时，它变得可逆，即在稳态条件下，可认为A_i与V_i的分布相同[19]。研究人员认为，武汉地区人口超过 1100 万，在 2020 年春节前几天（2020-01-19 至 2020-01-23）每日进出武汉的人口数量超过 100 万，从而这个更新过程达到了稳态。经验证，A_i和V_i具有相同的边际分布，满足更新过程的条件。按照以上假设，$T_i = A_i + V_i$。我们可以求得V_i的条件分布为 $h(v) = \dfrac{1-F(v)}{\mu}$，其中 $F(v)$是潜伏期的分布函数，μ 是平均潜伏期。如果假定潜伏期T_i服从某种特定的分布（如威布尔分布），则我们可以通过最大化似然函数 $L_F = \prod_{i=1}^{n} \dfrac{1-F(v_i)}{\mu}$，获取分布参数 θ 的 MLE，从而得到潜伏期的分布估计。

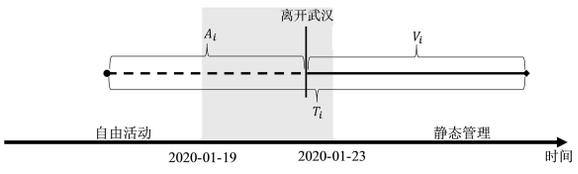

图 2　COVID-19 潜伏期估计的更新过程方法模型

3. 模型评价

对于最大似然估计，我们计算负对数似然函数值进行同类方法内的比较；同样计算贝叶斯估计的负对数似然函数值，并与最大似然估计进行比较。此外，我们对各种方法的数据利用情况及影响传染病防控政策制定的潜伏期分位数估计（2.5%、25%、50%、75%、90%、95%、97.5%和 99%分位数）进行仔细比较。

4. 统计软件

本研究所有数据处理和建模过程均通过 R 软件实现。其中，单区间删失方法的 NPMLE 估计采用 survival 程序包，而 MLE 估计和 Bayes 估计采用 icenReg 程序包；双区间删失方法的 NPMLE 估计采用 doubcens 程序包，而 MLE 估计和 Bayes 估计采用 coarseDataTools 程序包；随机过程方法基于 Qin 等[12]提供的 R 代码实现。

结　果

1. 基线特征

研究数据来源于 2020-01-04 至 2020-02-24 中国湖北以外确诊的 COVID-19 患者，总共 181 例。这些患者来自以亚洲为主的五大洲；年龄跨度较大，从 2 岁到 80 岁，平均年龄为 46.0（±15.4）岁；108 例（61.0%）为男性；159 例（90.9%）有武汉旅居史，137 例（75.7%）有明确的症状出现日期。具体信息见表 2。

表 2　研究对象的基本特征［n(%)］

汇总分类	统计量
病例数	181
年龄（岁）	
病例数（缺失数）	170（11）
均数±标准差	46.0±15.4
中位数（最小值~最大值）	44.5（2.0~80.0）
性别	
男	108（61.0%）
女	69（39.0%）
病例数（缺失数）	177（4）
武汉暴露史	
武汉旅居史	159（90.9%）
无武汉旅居史	16（9.1%）
例数（缺失数）	175（6）
是否观测到确切的症状出现日期	
是	137（75.7%）
否	44（24.3%）
病例数（缺失数）	181（0）
COVID-19 确诊地区	
亚洲	157（86.7%）
欧洲	9（5.0%）
北美洲	7（3.9%）
澳洲	6（3.3%）
南美洲	2（1.1%）
病例数（缺失数）	181（0）

2. 潜伏期分布估计

将 137 例具有明确症状出现日期的 COVID-19 患者纳入单区间删失方法分析。图 3 的结果显示，Turnbull 的 NPMLE 存在较多的跳跃“阶梯”；对于参

数模型，相同的分布假设下，MLE 估计与 Bayes 估计结果相近；尽管各模型对 COVID-19 的中位潜伏期的估计接近（5.4~6.0 天），但对于大分位数（如>95%分位数）的估计与 Turnbull 的 NPMLE 估计相比差别有变大趋势，置信区间变长，尤其是潜伏期的对数正态假设下，其 MLE 估计和 Bayes 估计与 Turnbull 的 NPMLE 估计差距最大，且 99%分位数估计超过 14 天。

纳入所有 181 例数据的双区间删失方法分析结果见图 4。可见，NPMLE 估计存在较多的“跳跃”点；相同分布假设下的参数模型，其 MLE 估计与 Bayes 估计接近；相比之下，不同分布假设的参数模型估计的结果差别要大，估计的中位潜伏期在 5.0~5.5 天之间；对于潜伏期大分位数（如>95%分位数）的估计与 NPMLE 估计差距变大，95%置信区间变宽，在对数正态假设下尤为明显。这些结果都与单区间删失方法得到的结果类似。

从更新过程的角度，研究数据中包含 59 例 2020-01-19 至 2020-01-21 期间离开武汉并在武汉以外确诊且获得确切症状出现日期（即前向复发时间明确）的患者，得到 COVID-19 潜伏期分布的 MLE 估计（图 5）。除威布尔分布假设下潜伏期分布的小于 50%分位数估计明显偏离其他两种分布假设（对数正态分布和伽马分布）外，其他各分位数估计接近，而且潜伏期中位数估计在 4.0 天左右。

3.模型评价

为了进行模型间的比较，我们计算各模型拟合结果的负对数似然函数值。尽管 Bayes 估计目标函数的优化采用的是后验分布函数，但本研究的结果显示，相同分析方法和分布假设条件下，按潜伏期分布的 Bayes 估计计算得到的负对数似然函数值，略大于 MLE 估计的结果（表 3），数值非常接近，提示 Bayes 估计与 MLE 估计吻合度很高。因此，这里仅比较不同模型的 MLE 估计。

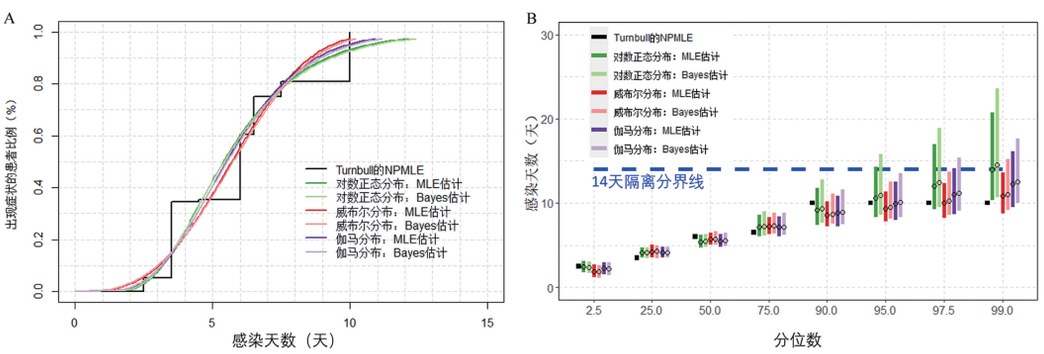

图 3 COVID-19 潜伏期分布的单区间删失方法分析

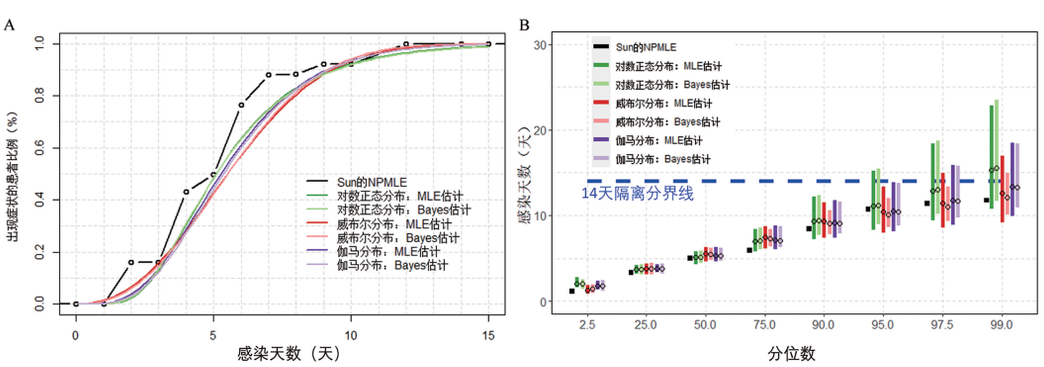

图 4 基于双区间删失数据的 COVID-19 潜伏期分布估计

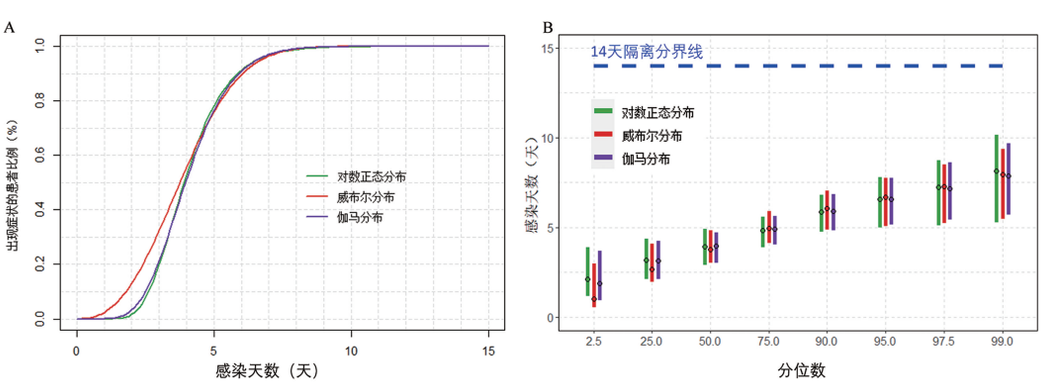

图 5 基于更新过程的 COVID-19 潜伏期分布估计

表 3　不同分布假设及分析方法获得的 COVID-19 潜伏期估计结果

统计学方法	分位数								-Log likelihood
	2.5%	25.0%	50.0%	75.0%	90.0%	95.0%	97.5%	99.0%	
单区间删失方法（n=137）									
NPMLE	2.5	3.5	6.0	6.5	10.0	10.0	10.0	10.0	40.754
对数正态分布假设									
MLE 估计	2.4 (1.8,3.1)	4.1 (3.5,4.8)	5.4 (4.7,6.2)	7.1 (6.0,8.6)	9.1 (7.3,11.8)	10.6 (8.3,14.4)	120 (9.1,17.1)	14.0 (10.2,21.2)	42.663
Bayes 估计	2.4 (1.8,3.1)	4.1 (3.5,4.7)	5.4 (4.7,6.3)	7.2 (6.2,9.0)	9.3 (7.8,11.5)	10.9 (8.6,15.6)	12.5 (9.5,18.7)	14.6 (10.8,23.1)	42.682
威布尔分布假设									
MLE 估计	1.9 (1.2,2.7)	4.2 (3.5,5.0)	5.7 (5.0,6.5)	7.2 (6.3,8.3)	8.5 (7.2,10.2)	9.3 (7.8,11.5)	10.0 (8.2,12.5)	10.8 (8.7,13.8)	43.936
Bayes 估计	1.8 (1.1,2.6)	4.2 (3.5,4.9)	5.8 (5.0,6.6)	7.3 (6.4,8.7)	8.7 (7.5,10.8)	9.5 (8.1,12.1)	10.2 (8.6,13.4)	11.0 (9.2,14.9)	43.944
伽马分布假设									
MLE 估计	2.3 (1.5,3.0)	4.1 (3.5,4.8)	5.5 (4.8,6.3)	7.1 (6.0,8.4)	8.8 (7.2,10.9)	9.9 (8.0.12.6)	11.0 (8.7.14.3)	12.3 (9.5,16.5)	42.985
Bayes 估计	2.2 (1.4,2.9)	4.1 (2.5,4.8)	5.5 (4.8,6.4)	7.2 (6.2,8.7)	8.9 (7.5,11.4)	10.1 (8.3,13.3)	11.2 (9.0,15.1)	12.6 (9.9,17.4)	
双区间删失方法（n=181）									
NPMLE	1.2	3.3	5.0	5.9	.5	10.7	11.4	11.7	52.177
对数正态假设									
MLE 估计	2.0 (1.6,2.6)	3.7 (3.2,4.3)	5.1 (4.4, 5.9)	7.0 (5.7,8.7)	9.3 (7.2,12.3)	11.1 (8.3,15.1)	12.9 (9.2,18.0)	15.3 (10.4,22.1)	56.872
Bayes 估计	2.0 (1.4.2.6)	37 (3.1,4.3)	5.1 (4.55.9)	71 (6.1,8.6)	95 (7.7,12.3)	11.2 (8.9.15.4)	13.1 (10.0.18.7)	15.6 (11.4,23.5)	56.963
威布尔分布假设									
MLE 估计	1.3 (1.8,3.1)	3.8 (3.3,4.5)	5.5 (4.8,6.4)	7.5 (6.3,8.7)	9.3 (7.4,11.1)	10.4 (8.1,12.7)	11.4 (8.6,14.2)	12.6 (9.2,16.1)	53.416
Bayes 估计	1.4 (0.8,2.0)	38 (3.1,4.5)	5.5 (4.7,6.3)	73 (6.4,8.6)	9.1 (7.9,11.1)	10.1 (8.7.12.8)	11.0 (9.4,14.2)	12.1 (10.1.16.0)	53.508
伽马分布假设									
MLE 估计	18 (1.3,2.3)	3.8 (3.4,4.3)	5.3 (4.6,6.1)	71 (6.0,8.3)	9.1 (7.3,10.7)	10.5 (8.3,12.5)	11.7 (9.1,14.2)	13.3 (10.1,16.2)	55.644
Bayes 估计	18 (1.2,2.4)	38 (3.2,4.4)	53 (4.7,6.2)	71 (6.3.8.7)	90 (7.9,11.6)	10.4 (8.9.13.8)	117 (9.8,15.8)	13 3 (10.9.18.4)	55.699
随机过程方法（n=59）									
对数正态分布假设	2.1 (1.2,3.7)	3.2 (2.2,4.3)	3.9 (3.0,4.9)	4.8 (3.9,5.6)	5.9 (4.8,6.8)	6.6 (5.0,7.8)	7.2 (5.1,8.8)	8.1 (5.4,10.2)	101.981
威布尔分布假设	10 (0.5,3.1)	2.7 (1.9,4.1)	3.8 (3.0,4.8)	4.9 (4.1,5.8)	60 (4.9.6.9)	6.7 (5.0,7.7)	7.3 (5.2,8.4)	7.9 (5.4,9.3)	102.950
伽马分布假设	1.9 (0.9,3.7)	3.1 (2.1,4.3)	4.0 (3.0,4.8)	4.9 (4.0,5.7)	5.9 (4.9,6.9)	6.6 (5.3,7.8)	7.2 (5.5,8.7)	7.9 (5.8,9.8)	102.103

本研究的结果显示，无论是单区间删失方法、双区间删失方法，还是随机过程的角度，各种方法不同分布假设条件下，其 MLE 估计的负对数似然函数值都非常接近，且都大于非参数方法。这提示，从拟合优度的角度，非参数方法的结果优于参数方法。如果我们以非参数模型结果为基准，无论是单区间删失方法还是双区间删失方法，对数正态分布假设条件下的潜伏期大分位数（≥95%）估计更倾向于偏离非参数模型；而随机过程方法在三个分布假设条件下的潜伏期大分位数估计基本一致。从数据利用的角度，由于受诸多假设条件的限制，随机过程方法能够利用的样本数目（n=59）明显少于单区间删失方法（n=137）和双区间删失方法（n=181）。

讨　论

本研究首先回顾了 COVID-19 潜伏期分布的统计估计方法，即单区间删失方法，双区间删失方法和随机过程方法，从收集数据的结构、数学符号化过程到模型的构建和实现，以及模型的评价，逐一进行了详细介绍；其次，利用 Lauer 等[11]收集的 181 例确诊患者感染暴露和出现症状的信息，对三种方法的 MLE 估计和 Bayes 估计结果进行了比较。我们的比较结果显示，同类方法不同分布假设间，非参数方法要比参数方法拟合效果更好，但非参数方法存在较多的跳跃点，且

无法获得估计的95%置信区间;同类方法相同分布假设条件下,MLE估计与Bayes估计结果和拟合效果相近;同类方法的对数正态假设条件下获得的潜伏期分布的大分位数(>90%分位数)可能较大地偏离非参数估计结果;从数据利用的角度,双区间删失方法对数据的利用率最高;由于数据收集和利用的差异,不同方法得到的潜伏期分布估计可能存在较大差异。

区间删失数据的NPMLE估计被认为是分析该类数据的金标准[7]。但非参数方法依赖于对潜伏期可能取值点的"猜测",一般只能从样本数据获得,对于样本数据以外的取值点,在估计结果则体现为无信息的"水平线"或"线性插值",这就是我们看到NPMLE存在较多"跳跃"点的原因(图1A和图3A)。另外,因为NPMLE估计不需要任何的分布假设条件,从而无法进行统计推断,也就没法计算估计的置信区间。基于此,研究人员普遍选择的是潜伏期分布的参数模型估计[6,9-13]。然而,由于我们难于像非删失数据估计方法那样方便地检查统计分布假设的准确性(如残差),我们完全有必要先获得区间删失数据的NPMLE,并将参数模型结果与之比较,只有在参数模型并未严重偏离NPMLE结果情况,才能有理由相信我们的参数模型结果的有效性和可靠性[16]。

在我们的研究里,同类方法相同分布假设条件下的MLE估计与Bayes估计结果和拟合效果相近。但是,一般模型的Bayes估计,通常以MLE估计为初始估计,采用模拟算法(如MCMC方法)通过最大化后验函数获得。前期关于COVID-19潜伏期分布估计,Backer等[10]和Linton等[6]利用stan语言[20]实现,而且一般需要额外计算留一法交叉验证(leave-one-out cross validation,LOO-CV)或泛化信息量准则(widely applicable information criterion,WAIC)参数[21]进行模型比较,模型的收敛性有时难以保证。因此,尽管Bayes估计有其优势[22],但无论是从理论还是计算的复杂度而言,基于区间删失数据的潜伏期分布Bayes估计不如其MLE估计直接和便捷。

不同分析方法之间,数据利用的效率差异较大,结果的变异也较大。显然,基于双区间删失方法利用了所有收集的181例数据,显示了最高的数据利用效率。理论上,双区间删失方法对单区间删失数据同样适用,为此,我们采用双区间删失方法对137例单区间删失数据重新进行了分析,结果与单区间删失方法完全一致。而Qin等[12]提出的随机过程方法,虽然最终分析计算过程比较简单,而且在一定程度上可纠正数据收集过程中的回忆偏倚,但其假设条件较多,导致满足条件的数据较少,从而产生样本选择偏倚,使其计算结果与区间删失方法得到的估计差别较大。另外,在新发传染病流行早期,数据采集和分析利用效率直接影响防控决策及其效果。因此,基于双区间删失数据分析方法是潜伏期的分布估计较好的选择。

综上所述,采用双区间删失数据的最大似然法估计传染病潜伏期分布,可以提高数据的收集、利用和分析效率,减少样本的选择偏倚;潜伏期分布估计过程中,除了比较不同分布假设下的估计结果,还要与非参数模型估计进行比较,并在不同数据集之间验证结果的可靠性;对潜伏期分布大分位数的估计和解释要谨慎,仅依赖于模型拟合优度统计量获得的"最佳"估计,有可能高估最长潜伏期。

参 考 文 献

[1] Boslaugh S. Encyclopedia of epidemiology. Thousand Oaks, CA: SAGE Publications, Inc.2008:583.

[2] Zhang M, Xiao J, Deng A, et al. Transmission dynamics of an outbreak of the COVID-19 Delta variant B.1.617.2-Guangdong province, China, May-June 2021. China CDC Weekly, 2021, 3(27): 584-586.

[3] 邱明悦,胡涛,崔恒建. 双区间删失下新冠病毒肺炎潜伏期分布的参数估计. 应用数学学报,2020,43(2):200-210.

[4] 万时雨,刘珏,刘民. 新型冠状病毒肺炎潜伏期的研究进展. 科学通报,2021,66(15):1802-1811.

[5] 方积乾. 生物医学研究的统计方法.第2版. 北京:高等教育出版社,2019:299-300.

[6] Linton NM, Kobayashi T, Yang Y, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. Journal of Clinical Medicine, 2020, 9(2): 538.

[7] Sun J. The statistical analysis of interval-censored failure time data. New York: Springer, 2006.

[8] 梁洁,崔燕,刘晓萌,等. 含有Ⅱ型区间删失数据的回归模型参数估计. 中国卫生统计,2017,34(4):546-549.

[9] Bi Q, Wu Y, Mei S, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. The Lancet Infectious Diseases, 2020, 20(8): 911-919.

[10] Backer JA, Klinkenberg D, Wallinga J. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. Eurosurveillance, 2020, 25(5): 10-15.

[11] Lauer SA, Grantz KH, Bi Q, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Annals of Internal Medicine, 2020, 172(9): 577-582.

[12] Qin J, You C, Lin Q, et al. Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. Science Advances, 2020, 6(33): eabc1202.

[13] 杜志成,顾菁,李菁华,等. 基于区间删失数据估计方法的COVID-19潜伏期分布估计. 中华流行病学杂志,2020,41(7):1000-1003.

[14] Nishiura H. Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. Emerging Themes in Epidemiology, 2007, 4: 2.

(下转第548页)

值高得的多”的机制；通俗理解就是医保部门“定工分”（每个病种的分值），医院“挣工分”（实际获得的分值），医保部门年度按“工分”给医院付费；DIP 付费有利于激励医院、医生提高医疗技术，处理疑难病例，降低医疗成本，最终使患者受益[13]。

但 DIP 付费受区域经济、医疗水平等因素影响，施行过程中难免出现问题，需要不断总结改进[14]。本研究选择相邻时间同一团队治疗的单侧初次手术的甲状腺乳头状癌患者为研究对象，从年龄、性别、手术操作、合并症、住院时间、住院费用、术后并发症方面相比两组差异无统计学意义；但研究组临床路径完成率明显低于对照组，原因为住院费用超过标准。进一步对住院费用分析发现，住院费用增多的原因是麻醉药费增加。这与 DIP 付费是先医保垫付，年底统一结算，而医院、科室、医生暂无相应监管而放松对月或季度患者住院费用控制有关。

鉴于本研究发现的 DIP 付费的不足，今后可以从下面三点进行改进：（1）提高临床路径完成率：降低病种成本最直接有效的办法就是推行临床路径，DIP 付费联合临床路径在医保控费、节省医疗资源等方面有优势[15]。（2）各相关科室均需有 DIP 概念和费用额度：出台与病种分值和费用相适应的激励约束机制，对超费情况进行处罚，结余给予奖励，摆脱以前“吃大锅饭”、“超支是笔糊涂账”的现状，实现奖惩责任到人，这样医生才能真心配合医疗费用的管理[14]。（3）动态调整 DIP 付费政策：根据地区一定时间段疾病分值和费用情况，动态调整相应分值和费用，并倾向支持医疗新技术和危重病例抢救。

综上所述，无论是单病种付费（含 DRGs），还是 DIP 付费，都是为了给参保对象提供合理、优质、高效的医疗服务，促使医疗机构技术发展、优化资源配置、节约医疗成本。在总额费用控制的前提下，由于各地方医疗技术、经济水平差异，DIP 付费实施过程中难免出现不足，需要及时改进。本研究发现，DIP 付费施行后降低了临床路径变异率，需要建立监管机制，控制医疗费用过快增长、提高临床路径完成率。

参 考 文 献

[1] 冯其柱，卢曼曼，王琦.甲状腺乳头状癌患者施行临床路径对医疗费用及单病种付费完成率的影响.中国医疗管理科学，2021，11（3）：19-23.

[2] 单红燕.DIP 分值付费下公立医院建立多维度成本管控模式的思考.中国总会计师，2021（2）：110-111.

[3] 胡晓梅，陈迎春，周福祥，等. 基于单病种付费的分级诊疗实施效果及影响因素研究.中国卫生政策研究，2019，12（10）：51-57.

[4] 付子英，井琨，王军. 以膀胱癌为例探讨医保单病种付费方式对医疗服务质量的影响.中华医院管理杂志，2017，33（12）：893-896.

[5] 秦成，泮露萍，陈秋月，等. 精益管理在卒中中心脑梗死单病种费用控制中的应用研究.中华危重病急救医学，2019，25（5）：637-640.

[6] 林振威，吴风琴，程斌.杭州市甲状腺恶性肿瘤单病种定额付费实施效果评价.江苏卫生事业管理，2019，30（8）：999-1001+1005.

[7] Qing F，Liu C.Forecasting Single Disease Cost of Cataract Based on Multivariable Regression Analysis and Backpropagation Neural Network. Inquiry，2019，56：1-15.

[8] 成柠，王鹏，庞宇，等.基于 DRG 数据的 DEA-Tobit 模型在临床科室运行效率评价上的应用.中国卫生统计，2021，38（3）：462-463.

[9] 金凤，赵阳，夏文斌，等. 疾病诊断相关分组（DRGs）在医院绩效考核的应用.中国病案，2020，21（10）：55-57.

[10] 冯玉莹，蔡鑫宇.基于 DRG 支付的云南省定点医疗机构冠心病住院患者医疗费用分析.医学与社会，2021，34（3）：102-106+122.

[11] 杨蓉，涂晓贤，陈锦华.肺癌手术患者的 DRGs 分组研究.中国卫生统计，2021，38（5）：769-772.

[12] 张丹.DRG/DIP 付费下公立医院绩效管理的实践与探讨.财会学习，2021，16（8）：160-161.

[13] 姜丽萍，蔡嘉敏，刘晓童.按病种分值付费（DIP）相关影响因素分析及中医院管理对策.按摩与康复医学，2021，12（16）：92-94.

[14] 伍敏琦，钟碧霞，蔡进，等.DIP 付费下子宫壁内平滑肌瘤超费原因分析及对策.现代医院，2021，21（3）：399-401.

[15] 陈维雄，林雯琦，欧凡，等.DIP 与临床路径对医疗资源消耗影响的实证研究.中国医疗保险，2021，14（3）：56-61.

（责任编辑：张　悦）

（上接第 545 页）

[15] Turnbull BW. The empirical distribution function with arbitrarily grouped，censored and truncated Data. Journal of the Royal Statistical Society：Series B（Methodological），1976，38（3）：290-295.

[16] Cowling BJ，Muller MP，Wong IO，et al. Alternative methods of estimating an incubation distribution：examples from severe acute respiratory syndrome. Epidemiology，2007，18（2）：253-259.

[17] De Gruttola V，Lagakos SW. Analysis of doubly-censored survival data，with application to AIDS. Biometrics，1989，45（1）：1-11.

[18] Sun J. Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. Biometrics，1995，51（3）：1096-1104.

[19] Qin J. Biased sampling，over-identified parameter problems and beyond. Singapore：Springer Nature Singapore Pte Ltd，2017.

[20] Carpenter B，Gelman A，Hoffman M，et al. Stan：A probabilistic programming language. Journal of Statistical Software，2017，76（1）：1-32.

[21] Gelman A，Hwang J，Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing，2014，24（6）：997-1016.

[22] Samaniego FJ，Reneau DM. Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. Journal of the American Statistical Association，1994，89（427）：947-957.

（责任编辑：张　悦）