·论 著·

一种用于推算感染时间曲线的非参数极大似然估计方法

蔡全才[1,2],姜庆五[1]*,程 翔[3],郭 强[4],孙庆文[3],赵根明[1]
(1. 复旦大学公共卫生学院流行病学教研室,上海 200032;2. 第二军医大学卫生勤务学系流行病学教研室,上海 200433;3. 第二军医大学基础医学部数理学教研室;4. 第二军医大学训练部)

[摘要] **目的**:研究通过传染病发病时间曲线推算感染时间曲线的方法。**方法**:采用基于EMS算法的非参数极大似然估计方法,根据传染病的发病资料和潜伏期资料推算感染时间曲线,并用于中国内地SARS感染时间曲线推算。**结果**:通过Matlab编程,所采用的方法可以方便地估计出感染时间曲线,并可平滑个体数据的波动。用于推算中国内地SARS感染时间曲线,发现感染时间曲线平滑稳定,有明显的2个高峰,感染高峰与实际发病高峰间隔5 d左右;发病人数(4 634例)与估计的感染人数(4 633.6例)相差在0.5例以内。**结论**:基于EMS算法的非参数极大似然估计方法对于感染时间曲线的估计是可靠的,可以用于SARS感染时间曲线的推算。所推算的感染时间曲线有助于干预措施效果评价。

[关键词] 非参数极大似然估计;EMS算法;感染时间曲线;严重急性呼吸综合征
[中图分类号] R 563.19 [文献标识码] A [文章编号] 0258-879X(2004)12-1349-04

A non-parametric maximum likelihood estimation method for estimating infection curve of infectious diseases

CAI Quan-Cai[1,2], JIANG Qing-Wu[1]*, CHENG Xiang[3], GUO Qiang[4], SUN Qing-Wen[3], ZHAO Gen-Ming[1] (1. Department of Epidemiology, School of Public Health, Fudan University, Shanghai 200032, China; 2. Department of Epidemiology, Faculty of Health Services, Second Military Medical University, Shanghai 200433; 3. Department of Mathematics and Physics, College of Basic Medical Sciences, Second Military Medical University; 4. Division of Training, Second Military Medical University)

[**ABSTRACT**] **Objective**: To develop a method for constructing the infection curve with the incidence curve of infectious diseases. **Methods**: The incidence and incubation period data of infectious diseases were used to estimate the infection curve by a non-parametric maximum likelihood estimation (MLE) method based on EMS algorithm, which was used to construct the infection curve of severe acute respiratory syndrome (SARS) in mainland China. **Results**: The results showed that the programming method by software Matlab could easily construct the infection curve and smooth the individual fluctuations. When used for estimation of the infection curve of SARS in mainland China, it was found that the curve was smooth and had 2 obvious peaks. There were about 5 d interval from the peak of infection curve to that of incidence curve. The difference between incidence number (4 634) and estimated infection number (4 633.6) was under 0.5. **Conclusion**: The non-parametric MLE method based on EMS algorithm is reliable for the estimation of the infection curve and can be used to construct the infection curve of SARS, which will help us to evaluate the effectiveness of intervention measures.

[**KEY WORDS**] non-parametric maximum likelihood estimation; EMS algorithm; infection curve; severe acute respiratory syndrome

[Acad J Sec Mil Med Univ, 2004, 25(12): 1349-1352]

感染时间曲线在传染病的预防和控制中具有重要的意义。通过直接监测传染病感染人数的变化,结合干预措施采取落实情况,能够较为准确、敏感、直观地评价干预措施的效果,从而克服传统采用发病时间曲线或病例报告时间曲线进行措施效果评价的缺点。同时,对于感染时间曲线的分析还有助于我们预测未来疾病的发生情况。然而,在出现临床症状之前,我们无法直接观察到某个个体是否已经感染,也就无法直接监测到人群感染人数的变化情况。所以,直接构建疾病的感染时间曲线通常是不可能的。

Chan-Yeung 等[1]通过接触史调查来构建传染病的感染时间曲线。然而,由于真实可靠的接触史通常是难于获得的,而且即使获得了可靠的接触史,也往往会因为是多病例、多时间、多次接触而导致无法对感染时间做出准确的估计。同时,通过这种方法构建的感染曲线容易受到个体病例的影响而出现较大的波动,不利于措施效果的评价。因此,通过上述病例研究的方法来构建感染曲线还是相当困难和不可

[基金项目] 上海市科委非典防治专项科研基金(NK2003-002);国家教育部防治非典科技攻关项目(No. 10).
[作者简介] 蔡全才(1969-),男(汉族),博士,副教授,硕士生导师.
*Corresponding author. E-mail: qwjiang@shmu.edu.cn

靠的。

基于 EM 算法的极大似然估计方法(MLE)可以用病例报告时间曲线来估计感染时间曲线，在 HIV/AIDS 研究中已有了成功的应用[2~5]。然而，由于病例报告时间曲线除了受到疾病潜伏期因素的影响以外，还受到发病至就诊、就诊至入院、入院至报告等时间间隔因素的影响，因此所推算的感染时间曲线可能出现较大误差。本研究对上述方法进行了改进，并用于中国内地 SARS 感染时间曲线的推算。

1　原理和方法

1.1　*构建似然函数*　假定感染事件是按照某种随机过程出现。N_t 为第 t 天的感染人数($t=1,2,\cdots,T$，T 是可利用发病数据的最后 1 d)；Y_t 为第 t 天的新发病例数；f_d 是潜伏期为 d 天的概率($d=0,1,2,\cdots$)；潜伏期服从 *Gamma* 分布(假设为其他分布也可以，只需要调整 f_d 的公式即可)。在这样的假设下，我们可以得到 $E[Y_t|N_1,N_2,\cdots,N_t]=\sum_{i=1}^{t}N_i f_{t-i}$，那么，第 t 天平均新发病例数为 $\mu_t=\sum_{i=1}^{t}\lambda_i f_{t-i}$。这里 $\mu_t=E[Y_t]$，$\lambda_i=E[N_i]$，$f_d=gamcdf(d+1,\alpha,\beta)-gamcdf(d,\alpha,\beta)$。其中，$gamcdf(x,\alpha,\beta)$ 是 Matlab 中计算 *Gamma* 分布累积分布的函数，α,β 为潜伏期 *Gamma* 分布的 2 个参数。

假设 $N_1,N_2,\cdots,N_T$ 是独立的泊松变量，那么我们可以得到 λ_t 的估计。同时，这也意味着 $Y_1,Y_2,\cdots,Y_T$ 也是独立的泊松变量。相对应于观察到的 $y_1,y_2,\cdots,y_T$，我们可以得到似然函数：

$$\prod_{t=1}^{T}(\sum_{i=1}^{t}\lambda_i f_{t-i})^{y_t}\exp(-\sum_{i=1}^{t}\lambda_i f_{t-i}$$

1.2　*EMS 算法*　假设 $N_{td}(d=0,1\cdots,T-t;t=1\cdots,T)$ 是第 t 天潜伏期为 d 天时受到感染的人数。那么，$Y_t=\Sigma_{d=0}^{t-1}N_{t-d,d}$，$Y_t$ 是第 t 天新发生的病例数，可以直接观察到。根据 Dempster 等[6]提出的 EM 算法，当使用下面迭代公式时似然函数将逐步递增：

$$\lambda_t^{new}=\frac{\lambda_t^{old}}{F_{T-t}}\sum_{d=0}^{T-t}\frac{Y_{t+d}f_d}{\Sigma_{i=1}^{t+d}\lambda_i^{old}f_{t+d-i}}$$

其中，$F_{T-t}=\Sigma_{d=0}^{T-t}f_d$，$\lambda_t$ 为第 t 天的感染人数。上述公式合并了 EMS 算法中的 E 步和 M 步。

最后，需要对感染曲线进行平滑处理，这个过程也就是 EMS 算法中的 S 步。首先，使得：

$$\Phi_t^{new}=\frac{\lambda_t^{old}}{F_{T-t}}\sum_{d=0}^{T-t}\frac{Y_{t+d}f_d}{\Sigma_{i=1}^{t+d}\lambda_i^{old}f_{t+d-i}}$$

接着，让 $\lambda_t^{new}=\sum_{i=0}^{k}w_i\Phi_{t+i-k/2}^{new}$。其中，$w_i$ ($i=0,1,\cdots,k$)为平滑计算中的权重。k 为偶数，它决定了权重平均的窗口宽度。k 和 w_i 在计算中需要事先确定。在本研究中，选择 Silverman 等[7]提出的双侧对称权重的方法来确定权重 w_i。计算公式为：

$$w_i=\left(\frac{k}{i}\right))/2^k,i=0,1,\cdots,k。$$

当 t 接近 1 或 T 时，$\Phi_{t+i-k/2}^{new}$ 的下标 $t+i-k/2$ 的取值有可能会超出范围。为了避免出现这种情况，我们作如下设定：当 $t+i-k/2<1$ 时，$\Phi_{t+i-k/2}^{new}=0$；当 $t+i-k/2>T$ 时，$\Phi_{t+i-k/2}^{new}=\Phi_T^{new}$。

1.3　*迭代收敛点*　当计算出来的 λ_t^{new} 值符合下述公式的要求时，则终止迭代，λ_t^{new} 值则为估计的第 t 天的感染人数。

$$\frac{|\Sigma_{t=1}^{T}\lambda_t^{new}-\Sigma_{t=1}^{T}\lambda_t^{old}|}{\Sigma_{t=1}^{T}\lambda_t^{old}}<\varepsilon$$

其中，$\varepsilon=10^{-4}$。

上述计算均采用 Matlab 6.1 软件编程完成。

2　实例分析

2.1　*资料来源*　SARS 病例资料来自于中国内地 2002～2003 年 SARS 个案调查数据库。由于 SARS 疑似病例在最后都被排除或重新分类为临床诊断病例，因此，在本研究中仅计算 SARS 临床诊断病例的病例数。SARS 病例的诊断标准符合中华人民共和国卫生部《传染性非典型肺炎诊断标准》。SARS 潜伏期分布数据从我们有关 SARS 潜伏期的研究结果中获得(另文发表)。

2.2　*感染时间曲线推算方法*　有关 SARS 潜伏期研究发现(另文发表)：SARS 潜伏期分布服从 *Gamma*(2.10，2.33)分布；95%的人在感染 SARS-CoV 后将在 11.42 d 内发病；99%的人在感染后将在 15.89 d内发病。因此，我们假定 SARS 潜伏期服从 *Gamma* 分布，最长潜伏期为 16 d。据此，选定先于中国内地首发病例发病时间 16 d 作为可能感染时间的起点，即 $t=1$。采用上述所建立的基于 EMS 算法的非参数 MLE 方法进行 SARS 感染时间曲线的推算。本研究发现 k 取不同偶数时仅对峰值的高低略有影响，因此，本研究选取 $k=10$ 计算 λ_t^{new} 值。

2.3　*结果*　估计的中国内地 SARS 感染时间曲线见图 1。从图中可以看出，以发病时间直接绘制的曲线波动大，且不规则，但似乎可见到 2 个主要的发病高峰。第 1 个发病高峰出现在 2003 年 2 月 8 日，第 2 个发病高峰出现在 2003 年 4 月 25 日。以估计的每日感染人数绘制的曲线较为平滑稳定，有明显的 2 个波峰。其中，第 1 个感染高峰出现在 2003 年 2 月 3 日至

4 日,与第 1 个发病高峰间隔 4～5 d。估计的第 2 个感染高峰更大,出现在 2003 年 4 月 19 日,与第 2 个发病高峰间隔 6 d。4 月 19 日以后疫情很快得到了控制,5 月 12 日开始估计的每日感染人数首次出现 10 例以下,估计的零感染从 5 月 22 日开始。

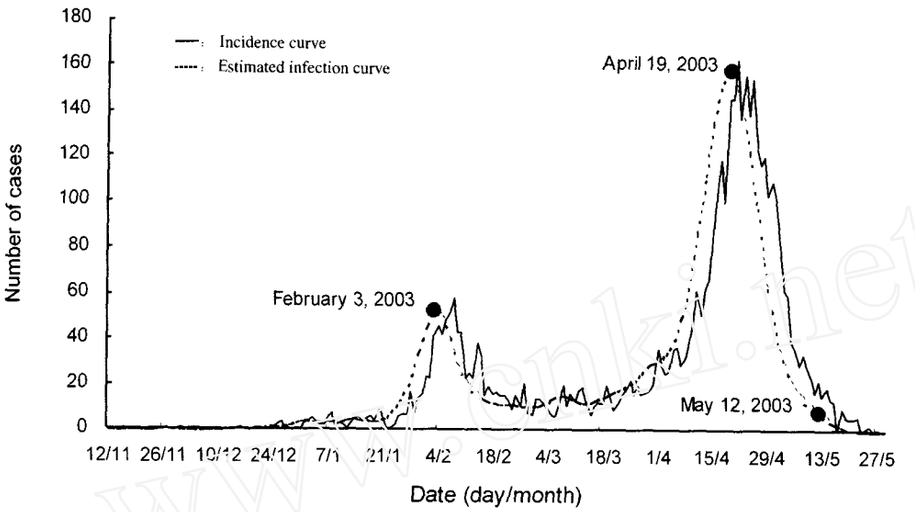

图 1 中国内地 SARS 发病时间曲线和估计的感染时间曲线

Fig 1 Incidence curve and estimated infection curve for SARS outbreak in China

3 讨 论

采用基于 EMS 算法的非参数 MLE 方法可以估计出每日感染人数,从而绘制出感染时间曲线。由于下列几方面原因,我们认为本研究所采用的方法是可靠和有效的:(1)EM 算法在不完全数据的研究中已有很多成功的应用[6]。本研究在 EM 算法的基础上引入了 S 步,以平滑个体数据波动。(2)本研究所采用的非参数方法可以确保所估计的感染人数不可能出现负值,克服了某些参数估计方法的缺陷[9,10]。(3)基于 EM 算法的 MLE 方法可以用病例报告时间曲线来估计感染时间曲线,在 HIV/AIDS 研究中已有了成功的应用[2～5]。病例报告时间曲线除了受到疾病潜伏期因素的影响以外,还受到发病至就诊、就诊至入院、入院至报告等时间间隔因素的影响,因此推算感染时间曲线的难度较大,也可能出现较大的误差。本研究采用个案病例的实际发病时间替代病例报告时间,简化了推算的过程,避免了上述可能出现的误差。(4)中国内地从 2002 年 11 月 16 日出现首例 SARS 病例以来,至 2003 年 5 月 29 日为止,SARS 临床诊断病例累计达到 5 316 例。其中,4 634 例(87.2%)有明确的发病时间记录。通过非参数 MLE 方法估计出的 SARS 感染人数累计约为 4 633.6 例,与实际的发病人数相差在 0.5 例以内。对中国内地 6 个重点疫区推算的结果也是类似的。同时,估计的感染高峰时间与实际的发病高峰时间间隔在 5 d 左右,这与我们有关 SARS 潜伏期均值研究的结果相一致(4.89 d,另文发表)。实例研究的结果说明了该方法对于感染时间曲线的估计是可靠的,可以用于 SARS 感染时间曲线的推算,所推算的感染时间曲线有助于干预措施效果评价。

[参 考 文 献]

[1] Chan-Yeung M, Yu WC. Outbreak of severe acute respiratory syndrome in Hong Kong Special Administrative Region: case report[J]. *BMJ*, 2003, 326(7394): 850-852.

[2] Brookmeyer R, Gail MH. Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States[J]. *Lancet*, 1986, 2(8519): 1320-1322.

[3] Becker NG, Watson LF, Carlin JB. A method of non-parametric back-projection and its application to AIDS data[J]. *Stat Med*, 1991, 10(10): 1527-1542.

[4] Chau PH, Yip PSF, Cui JS. Reconstructing the incidence of human immunodeficiency virus (HIV) in Hong Kong by using data from HIV positive tests and diagnoses of acquired immune deficiency syndrome[J]. *J Roy Stat Soc C (Appl Statist)*, 2003, 52(2): 237-248.

[5] Cui J, Becker NG. Estimating HIV incidence using dates of both HIV and AIDS diagnoses[J]. *Stat Med*, 2000, 19(9): 1165-1177.

[6] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from

incomplete data *via* the EM algorithm[J]. *J Roy Stat Soc B*, 1977, 39(1): 1-38.

[7] Silverman BW, Jones MC, Wilson JD. A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography[J]. *J Roy Stat Soc B*, 1990, 52(3): 271-324.

[8] Day NE, Gore SM, McGee MA, *et al*. Predictions of the AIDS epidemic in the UK: the use of the back projection method[J]. *Philos T Roy Soc B*, 1989, 325(1): 123-134.

[9] Rosenberg PS, Gail MH. Back calculation of flexible linear models of the human immunodeficiency virus infection curve [J]. *J Roy Stat Soc C(Appl Statist)*, 1991, 40(2): 269-282.

[10] Taylor JM. Models for the HIV infection and AIDS epidemic in the United States[J]. *Stat Med*, 1989, 8(1): 45-58.

[收稿日期] 2004-09-15　　[修回日期] 2004-09-30

[本文编辑] 尹　茶

· 临床研究 ·

乳腺癌术前快速活检定性诊断方法及其应用

Application of a fast biopsy method for pre-operative diagnosis of breast cancer

蔡清萍，王　强，项洪刚，高文超，周　辉

（第二军医大学长征医院普通外科，上海 200003）

[摘要] **目的**：借助乳腺旋切技术，探索建立一种新的乳腺癌术前定性诊断方法。**方法**：利用 Mammotome 乳腺微创旋切活检系统，对 26 例临床诊断为乳腺癌的患者进行乳腺肿块旋切，标本送病理学冰冻检查；另一组 26 例乳腺癌患者采用常规手术切除后送病理学冰冻检查，对两种方法的操作时间、送检标本病理学冰冻检查时间、病理学检查的阳性率、假阴性率及根治手术全身麻醉时间进行比较。**结果**：新的快速活检方法操作仅需(8.2±2.3) min，而传统的手术活检方法需(26.8±4.1) min，二者间有显著性差异($P<0.05$)。病理学诊断的阳性率均为 100%。送检标本病理学冰冻检查时间及根治手术全身麻醉时间两组间无明显差异。**结论**：乳腺癌术前施行旋切快速定性诊断方法操作时间短，阳性率高，无痛苦，明显减少手术时的全身麻醉时间，有良好的应用前景，值得推广。

[关键词] 乳腺肿瘤；活组织检查；Mammotome 乳腺微创旋切活检系统

[中图分类号] R 737.9　　[文献标识码] B　　[文章编号] 0258-879X(2004)12-1352-01

对于怀疑为乳腺癌的患者，确切的病理学诊断是指导外科手术治疗的金标准，因此寻找一种快速、简便、实用的定性诊断方法显得尤为重要，为此我们借助近年来发展起来的乳腺旋切技术建立了乳腺癌术前快速活检定性诊断方法，并与传统的手术切除活检进行比较，现介绍如下。

1　资料和方法

1.1　一般临床资料和分组　选择 2002 年 4 月至 2003 年 12 月来我院普通外科就诊、行手术治疗的乳腺癌患者 52 例，均为女性患者，平均年龄(61.3±8.4)岁，均为单发性乳腺肿块，B 超显示乳腺内浸润性肿块，病灶最大 3.2 cm×3.8 cm，最小 1.2 cm×1.8 cm，肿块位于乳晕区 8 例(15.4%)，外上象限 27 例(51.9%)，内上象限 6 例(11.5%)，外下象限 8 例(15.4%)，内下象限 3 例(5.8%)，其中 8 例伴有腋窝淋巴结肿大。所有患者随机分为 2 组(每组 26 例)：(1)快速活检定性诊断组；(2)手术切除活检定性诊断组。

1.2　方法　所有患者均进行术前告知，知情同意，两组患者在拟行外科根治性手术前分别采用快速活检定性诊断方法和手术活检定性诊断方法，比较两种定性诊断方法所需要的操作时间、切除标本病理学冰冻检查时间、病理学检查阳性率及后续的根治手术麻醉的时间等指标。具体操作方法如下：(1)快速活检定性诊断组：采用美国爱惜康内镜外科公司生产的 Mammotome 乳腺微创旋切活检系统(由槽式旋切刀、真空吸引泵、控制手柄和相应软件等组成)进行乳腺肿块微创旋切活检。患者取仰卧位，常规消毒铺单，以左手拇指和食指触及并固定乳腺肿块，皮肤表面以 1%利多卡因局部麻醉，尖刀切开皮肤约 0.3 cm，将 11 G 旋切刀经切口插入至病灶深处，使旋切刀凹槽对准病灶后进行多次扇形旋切(5～7 次)，旋切标本送病理学冰冻检查，局部压迫止血。(2)手术活检定性诊断组：按照传统常规方法进行全麻，肿块切除和病理学冰冻检查。所有患者在明确为乳腺癌的病理学诊断后立即进行根治性手术。

1.3　统计学处理　采用 t 检验。

2　结　果

采用两组方法进行活检均可获得充分的组织标本，足以进行病理学冰冻检查，切除的标本病理学冰冻检查时间无显著差异；采用微创旋切活检定性诊断方法操作所用时间较经典的手术切除活检方法明显缩短，患者无疼痛不适。所取得的旋切标本均一次性明确诊断，阳性率为 100%(26/26)，与手术活检定性方法阳性率相同。由于快速微创旋切活检仅在

(下转第 1356 页)

[作者简介] 蔡清萍(1969-)，男(汉族)，博士，主治医师.
E-mail: caiqingping@hotmail.com