·技术方法·

肿瘤发病人数估计的分层捕获-再捕获方法

高桂明 方积乾 柳青 刘颜

【摘要】 目的 建立一种估计广州市越秀区恶性肿瘤发病人数方法。方法 根据死亡统计和医院病案两种来源收集肿瘤病例,建立分层捕获-再捕获模型,应用 Bayesian 方法估计模型的参数。分析程序由 SAS 编程语言编写。结果 越秀区男性和女性恶性肿瘤人数的估计值分别为 610 例和 520 例,现有调查方式获得的男性和女性肿瘤发病资料的漏报率的估计值分别为8.5%和5.4%。结论 模拟研究的结果说明分层捕获-再捕获模型及对应的参数估计方法在估计肿瘤发病人数时比较可靠。

【关键词】 恶性肿瘤;分层;捕获-再捕获模型

Stratified ' capture-recaptured ' methodology in the estimation on the number of cancer incidence cases

GAO Guiming, FANG Jiqian, LIU Qing, et al. Sun Yat-sen University of Medical Sciences, Guangzhou 510089, *China*

【**Abstract**】 **Objective** To develop a new method in estimating the number of cancer incidence in Yuexiu district in Guangzhou. **Methods** Data of cancer incidence was collected thorugh hospital records and death certificates to develop a stratified capture-recaptured model. Related procedures for the parameter estimation were given and implemented on SAS language. **Results** Estimated numbers of both male and female cancer incidence in Yuexiu district were 610 and 520 respectively. The estimated miss-reporting rates on both male and female cancer incidence from hospital records and death certificates were 8.5% and 5.4% respectively. **Conclusion** The result of simulation studies showed that stratified ' capture-recaptured ' model proposed in this paper was reliable for the estimation of the number of cancer incidence cases.

【**Key words**】 Neoplasms; Stratification; Capture-recapture model

以人群为基础的恶性肿瘤发病登记是肿瘤流行病学研究的基础。在肿瘤发病登记制度不完善的地区,由于发病资料的漏报,肿瘤发病率偏低,不能反映实际发病水平。现在常用的一种估计肿瘤发病人数的方法是采用' 死补活 ",即从肿瘤死亡报告中查出某一年发病的资料(第一组资料),与当年医院的肿瘤发病登记资料(第二组资料)核对补漏。由于这两组资料都不完整,这样补漏后得到的肿瘤发病人数依然不完整。捕获-再捕获方法的对数线性模型(log-linear models for capture-recapture methods)[2]提供了一种根据某研究群体的来自不同调查途径的几组不完整资料估计该研究群体大小的方法,但肿瘤病人从发病到死亡的转移过程不能通过对数线性模型描述。我们建立一种估计肿瘤发病人数的分层捕获-再捕获模型,提供模型的参数估计方法,并应用于估计广州市越秀区的肿瘤发病人数。

基金项目:国家' 九五 '科技攻关项目(96-906-01-09)

作者单位:510080 广州,中山医科大学公共卫生学院卫生统计教研室

材料与方法

1.模型假设与框架:将某一年(第 1 年)的肿瘤发病病人按性别分为两组,收集第 1 年医院的发病记录和第 1 年至第 t 年(这里只考虑 $t=2$ 的情形)的死亡资料。假设肿瘤病人从发病到死亡的转移服从马尔可夫过程,则任一组发病病人的资料可整理成表 1 的形式。

表1 发病病人的观察数据(n/例)

第 1 年医院记录	死亡登记				
	有				没有
	第 1 年	第 2 年	…	第 t 年	
有	X_{11}^1	X_{11}^2	…	X_{11}^t	X_{10}
没有	X_{01}^1	X_{01}^2	…	X_{01}^t	X_{00}

设 X_{11}^j(X_{01}^j)分别表示在第 1 年医院的发病记录被观察到(未被观察到),在第 j 年的死亡登记中发现的发病病人个体数;X_{10}(X_{00})分别表示在第 1 年医院的发病记录被观察到(未被观察到),在死亡登记中未发现的发病病人个体数。

任一发病个体落入表 1 中各格子的概率如表 2。

表2　与表 1 对应格子的概率

第 1 年医院记录	死亡登记				
	有				没有
	第 1 年	第 2 年	…	第 t 年	
有	$p_1(1-\phi_1)\lambda$	$p_1\phi_1(1-\phi_2)\lambda$	…	$p_1\phi_1\phi_2\cdots\phi_{t-1}(1-\phi_t)\lambda$	$p_1(1-\lambda(1-\phi_1\phi_2\cdots\phi_t))$
没有	$(1-p_1)(1-\Psi_1)\lambda$	$(1-p_1)\Psi_1(1-\Psi_2)\lambda$	…	$(1-p_1)\Psi_1\Psi_1\cdots\Psi_{t-1}(1-\Psi_t)\lambda$	$(1-p_1)(1-\lambda(1-\Psi_1\cdots\Psi_t))$

表 2 中 ϕ_k（Ψ_k）分别表示在医院的发病记录中被观察到（未被观察到）的个体在第（$k-1$）年底活着的条件下活动到第 k 年底的概率，即第 k 年的年生存概率，$k=2,\cdots,t$；ϕ_1（Ψ_1）分别表示在医院的发病记录中被观察到（未被观察到）的个体活动第 1 年的概率；p_1 表示第 1 年发病病人当年到所调查医院住院治疗的概率；λ 表示死亡的肿瘤病人在死亡的当年被记录到死亡登记中的概率。当死亡登记比较完善时，可认为 $\lambda=1$；N 表示第 1 年肿瘤发病病人数。

为了得到有效的估计，需要减少参数的个数，我们假设：①死亡的肿瘤病人在死亡的当年被记录到死亡登记中的概率为 1；②在发病后的每一年，没有在所调查的医院住院治疗的肿瘤病人与在这些医院住院治疗的肿瘤病人的年生存率之比 Ψ_i/ϕ_i（$i=1,2,\cdots,t$）为常数 μ（若 $\mu<1$，则说明在上述医院的治疗有效果。μ 越小，说明治疗效果越好）。

2. 参数估计方法：在上述模型假设下，依照多项分布建立似然函数，根据似然函数可求出各参数的满条件分布[3]，通过 Metroplis 子链法（subchains）[5] 估计参数 ϕ_1,ϕ_2,p_1,μ,N。估计参数的计算过程可由 SAS 语言编制程序完成。

3. 资料来源：本研究从两个方面收集广州市越秀区 1996 年肿瘤发病资料[1]：①医院病历：从广州市越秀区附近的主要大医院（市人民医院、广州医学院附一院、中山医科大学附二院和肿瘤医院、市肿瘤医院）收集 1996 年所有 ICD-9 编码 140～208 和 225 的病例，选出户籍在越秀区的病历；②死亡登记卡：收集越秀区公安部门 1996～1997 年死因为恶性肿瘤的死亡证书。所得数据按性别分别整理列表如表 3。

表3　越秀区恶性肿瘤发病病人观察数据（n/例）

性别	1996 年医院记录	死亡登记		
		有		没有
		1996 年	1997 年	
男性	有	18	37	232
	没有	86	185	
女性	有	43	19	194
	没有	12	124	

结　果

将分层的捕获-再捕获模型及其参数的估计方法应用于表 3 中的数据，得参数的估计值如表 4。

表4　1996 年越秀区恶性肿瘤发病数据的参数估计值

性别	ϕ_1		ϕ_2		μ		p_1		N	
	M	$s_{\bar{x}}$	M	$s_{\bar{x}}$	M	$s_{\bar{x}}$	M	$s_{\bar{x}}$	M	$s_{\bar{x}}$
男	0.88	0.00	0.80	0.00	0.70	0.00	0.47	0.00	610	0.35
女	0.68	0.00	0.78	0.00	0.68	0.00	0.49	0.00	520	0.18

从表 3 和表 4 可知 1996 年男性和女性发病人数的估计值分别为 610 例和 520 例，而从医院和死亡统计得到的 1996 年男性和女性发病人数分别为 558 例和 492 例，由此可得到从医院和死亡统计得到的资料关于男性和女性发病人数的漏报率的估计值分别为8.5%和5.4%。由于 μ 的两个估计值（0.70，0.68）都小于 1，说明在上述大医院住院治疗的肿瘤发病病人的年生存率明显高于不在上述大医院住院治疗的肿瘤发病病人。

讨　论

为了说明上述模型的 Metropolis 子链算法的功效，将这种方法应用于一系列大小不等的模拟数据集，结果显示，参数 ϕ_1,ϕ_2,μ 的估计值与真值有一定的偏差，但 p_1 的估计值与真值几乎没有差别，N 的估计值与真值偏差也很小。例如我们取参数的真值为：$N=1\,000$，$p_1=0.6$，$\phi_1=0.7$，$\phi_2=0.6$，$\mu=0.65$，随机产生 100 个模拟数据集，应用上述模型的 Metropolis 子链算法得到 N 的估计值的均值对真值的相对偏差为 28/1 000。这说明了上述模型和算法在估计肿瘤发病人数时有一定的可靠性。产生有偏估计的原因可能与 Metropolis 子链算法中各子链的建议分布选取不当有关，这需要在以后的研究中探索功效更强的算法。

由于只收集到两年的死亡资料（$t=2$），限制了模型的参数的个数和模型的假设。如果能收集到 3 年以上的死亡资料（$t\geqslant 3$），可以提出更为合理的模型假设，得到更有效的估计方法。

分层的捕获-再捕获模型最早产生于野生动物

的研究,它通过对在一个地区观察到的动物进行标记,观察它们随后几年向其他几个地区的转移情况,从而估计动物在不同地段之间的转移率[6]。本文建立的模型不仅考虑可在医院(可当作一个地区)观察到的病人从发病到死亡的转移过程,而且考虑在医院未观察到的病人从发病到死亡的转移过程,进而估计发病病人的总人数。本文提出的模型假设每个肿瘤发病病人到所调查医院住院治疗的概率相等,故不适合个体间差异较大的肿瘤发病群体。

参 考 文 献

1 柳青,刘颜,曹卡加,等. 广州市越秀区恶性肿瘤发病率分析. 癌症, 1999,18:538-540.

2 高桂明,方积乾,夏海晖,等. 捕获-再捕获资料的对数线性模型及其在估计婴儿死亡人数中的应用. 数理医药学杂志,2000,13:193-194.

3 茆诗松,王静龙. 高等数理统计. 北京:高等教育出版社,1998. 444-459

4 Vounatsou P, Smith AF. Bayesian analysis of ring-recovery data via markov chain monte carlo simulation. Biometrics, 1995, 51:687-708

5 Tanner MA. Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. 3rd edition. New York: Springer-Verlag, 1996.177-182.

6 Schwarz CJ, Schweigert JF, Arnason AN, et al. Estimating migration rates using tag-recoery data. Biometrics, 1993, 49:177-193.

(收稿日期:2000-08-26)