应用Bayes概率法早期预测流脑的流行

曾 光[1] 胡 真[1] 来秀君[2] 吴贵坤[2]

为了研究适合基层防疫部门用于流行预测的一种切实可行的早期预测算法，本协作组以北京市10个边远地区的流脑疫情及人群流动资料为统计分析基础，进行了流脑流行早期预测的探索。在研究设计中作了大胆革新，应用Bayes概率法作出了定性预测，取得了满意的结果。同时，将该法的预测结果和用逐步判别分析法(Stepwise Discriminatant Analysis)的预测结果进行了比较。

资料及准备

一、资料来源：以北京市10个边远地区(A～J)自1960年以来的流脑旬报资料和人群额外流动量（简称人群流动量）为依据。这10地区人口相近，其中A区为半工矿半农业区，其余为农业区。10个地区自1960年以来流脑流行趋势相似，流脑病例的诊断标准相同且多年基本稳定。

二、资料的整理：

1.划分流行年度：规定自上年11月份至当年10月份为一个流行年度，其中自上年11月份至当年7月份为流行期，8～10月份为间歇期(据1959至1982年24年资料统计，间歇期病例仅占总病例数的1.1%)。逐年计算各地区各流行年度流行期的流脑罹患率及与上一流行期罹患率的环比值。

2.按旬整理流行期资料：将各地区历年流行资料，逐年分旬顺序编号，从上年11月上旬至7月下旬依次定为第1旬、第二旬……，第27旬，计算出或找出各地区、各年度流脑病例发生的中位数旬次，众数旬次和10%、20%、80%和90%分位数旬次。

三、规定变量：

1.因变量：本文研究的是两类定性预测。预测的因变量有两项：

y：流行规模。分类标准为：流行期罹患率≤50/10万为流行(y_1)，否则为散发(y_2)。

Z：流行趋势。分类标准为：年度流行期罹患率环比值≥150%为升高(Z_1)，否则为非升高(Z_2)。

2.自变量：所用的初选自变量有11个：

X_1：上年度流行期病例累积发生的中位数旬次；X_2：上年度流行期病例发生的众数旬次；X_3：上年度流行期病例累积发生的10%分位数旬次；X_4：上年度流行期80%病例发生的旬期。规定为病例发生的90%旬次减去10%旬次的差值；X_5：上年度流行期流行强度；X_6：上年度流行期流行趋势；X_7：上年度流行期病例累积发生的20%分位数旬次；X_8：上年度流行期病例累积发生的80%分位数旬次；X_9：上年度流行期60%病例发生的旬期，参照X_4的算法；X_{10}：上一流行年度的人群流动量。规定为≥5万人，或≥500万个工为流动量“大”，否则为“非大”；X_{11}本流行年度的人群流动量。规定同X_{10}。

计算方法

Bayes概率法和逐步判别分析法都要分别进行复测(回代检测)、回顾性预测和实际预测计算。并按下述方式进行两次。

1 中国预防医学中心流行病学微生物学研究所
2 北京市卫生防疫站

第一次计算：样本资料为A～J地区1960～1980年的资料，共210例样本。测算流行趋势时，应用了全部210例样本；测算流行规模时，因其中11例罹患率值，恰在50/10万(分类标准值)左右，用作判别分析样本的意义不大，予以舍去。实际用了199例。由此做统计分析建立数学模型后，将这些样本随机抽出的一部分反代入模型，做复测检验。对1981和1982年10个地区的流脑流行规模和流行趋势做回顾性预测检验(20例)。对1983年10个地区的同样指标做实际预测(10例)。

第二次计算，从第一次计算的样本中随机抽出了20例留作回顾性预测研究用。但将1981和1982年的资料加入样本中，使得本次计算的样本数仍为第一次计算。

一、Bayes概率法：本文的算法过程如下：

1.按变量水平归类：先将全部样本按因变量的不同水平归类，分别清点归入y_1和y_2类、Z_1和Z_2类的样本数。然后按照每个自变量的不同水平再分类，清点分入各类各水平的样本数，逐一计算出某水平的样本数占该自变量各水平总样本数的百分比。

2.选择用于预测的自变量：本文从11个初选自变量中选出了6个自变量(X_2、X_3、X_5、X_6、X_{10}、X_{11})用于预测计算。选择的标准为：如果因变量对应的两类(如Y_1和Y_2类)，所占某自变量的各个水平的百分数间呈明显的差距，则考虑该自变量入选；而且意义相近的自变量(如X_1与X_2、X_3与X_7)只选入一个。第一次计算时因变量和自变量对应的数值见表1。

表 1　北京10个边远地区流脑早期预测Bayes概率法计算表（第一次计算）

自变量		流行规模y				流行趋势z			
		流行y_1		散发y_2		升高z_1		非升高z_2	
X_2	①≤13	8	(12.31)	53	(39.55)	17	(24.28)	52	(37.14)
	②=14、15、16	42	(64.61)	54	(40.30)	33	(47.14)	65	(46.43)
	③≥17	15	(23.08)	27	(20.15)	20	(28.57)	23	(16.42)
X_3	①≤8	8	(12.31)	62	(46.27)	21	(30.00)	54	(38.57)
	②=9、10、11	28	(43.08)	48	(35.82)	21	(30.00)	58	(41.43)
	③≥12	29	(44.62)	24	(17.91)	28	(40.00)	28	(20.00)
X_5	① 流行	41	(63.08)	25	(18.66)	14	(20.00)	53	(37.86)
	② 散发	24	(36.92)	109	(81.34)	56	(80.00)	87	(62.14)
X_6	① 升高	42	(64.62)	25	(18.66)	13	(18.57)	51	(36.43)
	②非升高	23	(35.38)	109	(81.34)	57	(81.43)	89	(63.57)
X_{10}	① 大	49	(75.38)	48	(35.82)	33	(47.14)	63	(45.00)
	② 非大	16	(24.62)	86	(64.18)	37	(52.86)	77	(55.00)
H_{11}	① 大	50	(76.92)	37	(27.61)	39	(55.71)	50	(35.71)
	② 非大	15	(23.08)	97	(72.39)	31	(44.29)	90	(64.29)

注：（）内数字为百分数

3.分别计算Y_1、Y_2和Z_1、Z_2占总样本数的百分比——Sy_1、Sy_2和Sz_1、Sz_2。即Bayes概率法的“事前概率”。例如第一次计算时Sy_1和Sy_2分别为32.38%和67.62%。

4.预测计算：

①计算概率乘积$P'y_1$、$P'y_2$和$P'z_1$、P^1z_2：

例如：预测1983年A地区流行规模时，要分别计算出流行和散发的概率乘积$P'y_1$和$P'y_2$。

已知，A地区该年X_2、X_3、X_5、X_6、X_{10}、X_{11}观测值的水平分别为2、3、2、1、2、2，则第一次计算时：

$p'y_1 = 64.61\% \times 44.62\% \times 36.92\% \times 64.62\% \times 24.62\% \times 23.08\% = 0.003908$

$p'y_2 = 40.30\% \times 17.91\% \times 81.34\% \times 18.66\% \times 64.18\% \times 72.39\% = 0.005090$

②分别计算流行和散发、升高和非升高的概率py_1、py_2、pz_1、pz_2：

$$py_1 = \frac{Sy_1 \cdot p'y_1}{Sy_1 \cdot p'y_1 + Sy_2 \cdot p'y_2} \times 100\%$$

$$= \frac{32.38\% \times 0.003908}{32.38\% \times 0.003908 + 67.62\% \times 0.005090} \times 100\%$$

$$= 26.82\%$$

$$py_2 = \frac{Sy_2 \cdot p'y_2}{Sy_2 \cdot p'y + Sy_2 \cdot p'y_2} \times 100\%$$

$$= \frac{67.62\% \times 0.005090}{32.38\% \times 0.003908 + 67.62\% \times 0.005090} \times 100\%$$

$$= 73.12\%$$

根据计算结果可知，1983年A地区的流行规模为流行和散发的概率分别为26.82%和73.12%。用同样的算法，也可对A地区流行趋势作出预测，以此类推。

二、逐步判别分析法：应用的因变量、初选自变量和样本资料与Bayes概率法完全相同。对于自变量指标的定性数值首先要按0，1分布予以数量化。

如：X_{10}：(大：1，非大：0)，然后，按一定的F值对初选自变量进行筛选，用选得自变量作判别分析计算。当F=24时，选得X5、X_6、X_{10}和X_{11} 4个自变量用于判别流行强度(Y_1、Y_2)，选得X_5、X_6和X_{11} 3个自变量用于判别流行趋势(z_1、z_2)。判别方程如下：

$$Y_1 = 5.8794X_5 + 2.1299X_6 + 4.3798X_{10} + 4.4662X_{11} - 10.9278$$

$$Y_2 = 7.1619X_5 + 3.4377X_6 + 3.3427X_{10} + 2.4851X_{11} - 14.6945$$

$$Z_1 = 6.7108X_5 + 1.6956X_6 + 7.9722X_{11} - 12.9348$$

$$Z_2 = 5.4632X_5 + 2.1909X_6 + 6.3907X_{11} - 10.0437$$

结果与分析

应用两种不同算法的复测、回顾性预测和实际预测结果见表2。

结果表明，应用Bayes概率法对10个地区流脑流行规模和流行趋势的复测、回顾性预测和实际预测的符合率均在70～100%之间，其中对流行规模的预测结果更优。预测与复测的符合率相一致，说明用过去资料建立的模型，在当前仍具有适用性。应用该法两次不同计算的结果亦相近。其中第二次计算中，对计算前随机选取的20例所做的回顾性预测，可说明模型的适用性。因为这20例包括了这10个地区从1960至1980年流脑疫情动态变化的各种情况，其中有流行年9例(包括解放后发病率最高的年份)和疫情升高年8例。

表2 北京10个地区流脑流行的复测、回顾性预测和实际预测的符合率(%)

计算方法	计算类别	流行规模y			流行趋势z		
		复测	回顾预测	预测	复测	回顾预测	预测
Bayes概率法	第一次计算	85	100	100	80	75	70
	第二次计算	85	85	100	85	75	70
逐步判别分析法(F=2.4)	第一次计算	79	85	100	71	85	90
	第二次计算	85	85	100	68	85	85

对1983年的疫情做了实际预测。对流行规模的预测符合率为100%，颇为满意。对流行趋势的预测符合率为70%，比前者低，但可根据实际情况做进一步分析：在10个地区中，流脑疫情回升最为显著的为G地区。该地区的报告病例数，由上一流行期的18例骤升至57例，对该地区的预测结果完全正确。对流行趋势的3例误判现象，都发生在病例基数在10例以下的情况。说明应用该法判别病例基数低的流行趋势时仍有缺陷，有待进一步提高。

讨　论

一、所用自变量的意义：多年来，A～J地区流脑主要致病菌株为A群。目前，在这些地区尚未较大规模地开展预防接种的情况下，

流脑疫情主要按其自然演变规律而波动，同时要受到社会因素的影响。A～J地区的流脑波动趋势，与国内大多数地区所报道一致，即每8～10年出现一次流行高峰，在一个自然流行期内，季节高峰的前偏或后移，反映了人群免疫水平的升高或下降，必然会对下一流行期的流行规模和流行趋势带来影响。根据这些特点，本文确定了多数自变量。此外，人群的大规模流动，是酿成流脑流行的重要社会因素。因而，亦考虑到以人群额外流动量为自变量。

二、定量预测与定性预测：一般说来，采用因子的时间区间，越远离流行高峰则越不容易得到精确的信息，反之亦然。我们在研究流脑流行的早期预测时，现实地选择了定性预测。我们实际做的虽然是两类判别预测，但是成功率较高。将得到的流行规模和流行趋势的预测结果进行综合分析，实际可得到四种不同的组合：

流行-升高年，流行-非升高年，散发-升高年，散发-非升高年。

四种结果，各具有不同的流行病学意义，据此来指导流脑的防治工作，具有重要的实际意义。

三、对模型可靠性的估计：过去的数理预测研究，往往要等待一年一度的预测考验，在短期内很难对模型的可靠性做到心中有数。针对这一问题，本文研究设计了两次不同的算法。第一次算法中，取1960～1980年资料为样本，以1981、1982年作预测考验年(即回顾性预测年)，以1983年的资料做实际预测。第二次算法中，以1960～1982年资料作样本，但从1960～1980年的样本中，随机选取了20例作预测考验，仍以1983年资料作实际预测。这样做的结果，使我们大大地增加了考验模型的机会。特别是第二次计算时随机抽取的20例，包括了各种不同水平的流行规模和流行趋势，从多方面考验了模型。因此，我们有理由认为，预测模型是有实用意义的。

摘　要

为了早期预测流脑在北京市10个远郊居民区的流行，应用贝叶斯（Bayes）概率法建立了数学模型，并以逐步判别分析法作为对比算法。所用基础资料为1960～1980年这10个居民区的流脑旬报数据和人群额外流动量。以此算得1960～1980年的理论发病率及1981～1983年的预测发病率与实际报告值颇相符，且与用逐步判别分析法算得理论值一致。由于该算法简单易学，故适应于县级防疫站应用。

ABSTRACT

A mathematical model,based upon Bayes Method of probability (BMP), was established for the early forecast of epidemic cerebrospinal meningitis (ECM) in 10 communities of Beijing outskirts,with the comparision of stepwise Discriminatant Analysis Method (SDAM). Every 10 days'case report of ECM and the records of extra population flowing during 1960～1980 in these communities were analysed in establishing this model.The results of 1981～1983 calculated from this model were 70～100% precise to the observed morbidity levels and trends of (ECM) from 1960 to 1983, and agreed with that calculated from SDAM.Since BMP is far easier to learn than SDAM, it can be used by antiepidemic stations in county level.

（谨对河南唐河县陈长志大夫，医科院基础所高润泉教授，赵珩和韩少梅同志协助计算表示谢意）