讲 座

疾病预测预报方法

辽宁省卫生防疫站 章扬熙

疾病监测的信息反馈主要有两个方面，一是使对策、措施调优，一是预测预报。进行预测预报，可使防病工作有预见性，打主动仗，从而可防患于未然或防微杜渐，以控制疾病的发生和流行。疾病的预测就是根据疾病发生、发展规律及有关因素，收集有关资料，运用分析判断、数学模型等方法，对疾病未来发展的趋势和强度作出预测。本文就常用的疾病预测方法作一介绍。

疾病预测的依据

一、疾病资料：传染病疫情资料、慢性病资料、恶性肿瘤和心血管病等资料。通过漏报调查进行质量控制，计算估计发病率，用于预测。

二、致病因子资料：在传染病方面，包括病原体的群、型、毒力大小、变异情况、耐药性等；在非传染病方面，包括致病物质、危险因子等。

三、宿主资料：在传染病方面，要了解不同病型的患者和病原体携带者的分布及血清学调查资料。在自然疫源性疾病和寄生虫病方面，要了解疾病在动物中的情况、传染来源、储存宿主及中间宿主的种类及其分布等。在非传染病方面，包括疾病谱分布及遗传素质等。

四、环境资料：包括与疾病发生、发展有关的环境因子，如人口、气象、地理、媒介昆虫、环境污染、防制疾病的对策、措施实施情况等。

疾病的定性预测

对疾病的发展趋势和强度作定性的估计，是上升还是下降，是流行还是散发。

一、综合预测法：研究疾病的流行动力学，探求流行的动态规律，综合各有关因素，预测疾病的发展趋势。传染病的流行是有先兆的，可采用“审势法”进行预测。这种先兆常表现为上一流行病学年流行熄止较晚，上一年疫势已有上升，本年该病发生又起步早、病例多、病情重、发展速度快、大年龄组发病频率增加、涉及面扩大、有爆发点不断发生、密切接触传染指数高、感染谱左移等迹象；也可以结合有关因子进行预测。比如，三带喙库蚊出现早、高峰早，感染率高，人群乙脑HI抗体阳性率低，预兆流行性乙型脑炎流行。黑线姬鼠等鼠密度高，感染多，预兆流行性出血热的流行。脑膜炎双球菌A群带菌率明显上升，人群抗体阳性率低，预兆流行性脑脊髓膜炎的流行。又如，高血压、吸烟、高胆固醇、肥胖等是冠心病的危险因素，危险因素愈多，愈易患冠心病。

二、控制图法：本法适用于有明显季节性的传染病预测。以疾病最低发病月为起始月，划分流行病学年（比如，某地流脑最低发病月为9月，所以从每年9月起到次年8月止为流脑的流行病学年）。以过去若干年（至少包括一个流行周期）该病月发病率资料为基础，用最大、最小和中位数月发病率的数值绘制半对数疾病流行控制图。横坐标为月别，用普通尺度，纵坐标为该病的月发病率，用对数尺度。图中共有三条线，以各月发病率的最大值划出上限线，以各月发病率的最小值划出下限线，以各月发病率的中位数划出中位数线。为了预测某年该病的流行趋势，可以从该年9月起，把该病的每月发病率依次划在控制图上。如果这些点子在中位数以下，点的连线坡度比中位数线小，往往预兆散发；如果这些点子在中位数以上，线的坡度比中位数大，则往往预兆流行。应补充说明的是，控制图的纵坐标用半对数尺度的目的在于直观月发病率的变化速率。又由于传染病各月的发病率多呈偏态，故以中位数、最大值和最小值来作控制图。

三、尤度法：又叫最大似然法。它是把一系列条件概率连乘来求似然值。由于条件概率取值在0～1间，计算较繁，通过对数变换，把条件概率换成整数的指数，把连乘变成连加，得H值，哪个H值大，即判为哪一类，使计算简化。条件概率与指数的对应关系如表1。

〔例1〕某地对流脑进行预测，以10个地区1960～1980年的资料，按自上年11月份至当年10月份为一个

表1　　条件概率与指数的关系

P值	指数	P值	指数
0～	－10	0.113～	1
0.012～	－ 9	0.142～	2
0.015～	－ 8	0.178～	3
0.018～	－ 7	0.224～	4
0.023～	－ 6	0.284～	5
0.029～	－ 5	0.355～	6
0.036～	－ 4	0.447～	7
0.045～	－ 3	0.563～	8
0.057～	－ 2	0.708～	9
0.071～	－ 1	0.892～	10
0.090～	0		

注：指数＝〔lg P（Si/A_G）＋1〕×10取整

流行年度。流行分类标准为：流行期（11月～7月）的罹患率≥50/10万为流行，<50/10万为散发、因变量Y为流行规模，按流行、散发作两类定性预测，自变量（预报因子）6个，X_1～X_6。从10个地区1960～1980年中剔除罹患率值恰在50/10万左右的样本11个，实际应用了199个样本，其中流行年65个，散发年134个（详见表2）已知甲地区1983年X_1～X_6的观测值分别为②、③、②、①、②、②，试用尤度法对该年流行规模进行预测。

依据已知条件，对照表2中的相应值，得：

$H_1=8+6+6+8+4+4=36$

$H_2=6+3+9+3+8+9=38$

$H_2>H_1$，结果预测1983年为散发。

四、Bayes概率法：这种方法与尤度法相似。如果以X_1，X_2，…Xm表示指标（自变量），A_1，A_2，…Ag表示类别，当某样品各指标值分别为S_1，S_2，…Sm时，属于AG类的概率P（AG/S_1S_2…Sm）为下式（公式1）：

$$P(A_G/S_1S_2\cdots Sm)=\frac{P(A_G)P(S_1/A_G)P(S_2/A_G)\cdots P(Sm/A_G)}{\sum_{G=1}^{g}P(A\ )P(S_1/A_G)P(S_2/A_G)\cdots P(Sm/A_G)}$$

式中P（AG）为事前概率，不知时可用实际发生的频率来估计。比较计算所得的各P（AG/S_1S_2…Sm）值，若P（Af/S_1S_2…Sm）最大，则判断该年属Af类。其值即系判定为Af类的事后概率。

〔例2〕将例1用Bayes概率法进行预测

把原题的事前概率求出，得：

$P(A_1)=65/199\times100\%=32.38\%$

$P(A_2)=134/199\times100\%=67.62\%$

表2　　某10个地区流脑预测的资料

自变量		流行（Y_1）			散发（Y_2）		
		频数	百分比(%)	指数	频数	百分比(%)	指数
X_1	①≤13	8	12.31	1	53	39.55	6
（上年度流行期病例发	②＝14,15,16	42	64.61	8	54	40.30	6
生的众数旬次）	③≥17	15	23.08	4	27	20.15	3
X_2	①≤8	8	12.31	1	62	46.27	7
（上年度流行期病例累	②＝9,10,11	28	43.08	6	48	35.82	6
积发生的P_{10}旬次）	③≥12	29	44.62	6	24	17.91	3
X_3	①流行	41	63.08	8	25	18.66	3
（上年度流行期流行强度）	②散发	24	36.92	6	109	81.34	9
X_4	①升高	42	64.62	8	25	18.66	3
（上年度流行期流行趋势）	②非升高	23	35.38	5	109	81.34	9
X_5	①大	49	75.38	9	48	35.82	6
（上年度人群流动量）	②非大	16	24.62	4	86	64.18	8
X_6	①大	50	76.92	9	37	27.61	4
（本年度的人群流动量）	②非大	15	23.08	4	97	72.39	9

注：指数系根据百分比查表1而求得

再把有关数据代入公式1，得：

$$P(A_1/S_1S_2\cdots S_6)=\frac{3238\times6461\times4462\times3692\times6462\times2462\times2308}{3238\times6461\times4462\times3692\times\times6462\times2462\times2308+6162\times403\times1791\times8134\times1866\times6418\times7239}$$

$$=26.87\%$$

$$P(A_2/S_1S_2\cdots S_6)=\frac{6762\times403\times1791\times8134\times1866\times6418\times7239}{3238\times6461\times4462\times3692\times6462\times2462\times2308+6162\times403\times1791\times8134\times1866\times641.5\times7239}$$

$$=73.13\%$$

由于$P(A_2/S_1S_1\cdots S_6)$大，故预测1983年流脑为散发，这个预测的事后概率为73.13%。

五、逐步判别分析：这是一种多类判别的方法，它根据各预报因子的重要性大小，逐个经过F检验，把有显著意义的变量选入判别函数，当引入新的预报因子后，将原判别函数中的判别能力已减弱到使F检验不显著的变量剔出函数，直到再无变量能引入函数，亦无已引入函数中的变量需要剔除时为止。计算判别系数，把判别函数最终确定下来。将各个预报因子Xi（i=1,2,…m）代入判别函数，计算各Yj值，若Yk在Yj中为最大，即将该年预测为第K类。由于逐步判别分析计算量大，需借助于电子计算机来完成计算。

〔例3〕应用某省1970～1979年的疫情资料及气象资料，选取了七个预报因子（当年4～7月平均月气温（x_1～x_4），前一年10～11月乙脑月发率（x_5～x_6），当年7月乙脑月发率（x_7），来预报乙脑的流行程度（年发病率<4/10万为第一类散发，>4/10万、<15/10万为第二类小流行，>15/10万为第三类流行）。试应用逐步判别分析建立预报判别函数。

经电子计算机的逐步判别程序运算，得到以下判别函数：

$$Y_1=-2107.13-358.07X_1+341.84X_3+700.14X_5-38894.78X_6+253.03X_7+1n0.4$$

$$Y_2=-1716.57-312.04X_1+303.70X_3+625.17X_5-34213.04X_6+2170.93X_7+1n0.4$$

$$Y_3=-1312.05-271.39X_1+264.09X_3+540.83X_5-29471.17X_6+1916.21X_7+1n0.2$$

表3　　数　据　与　结　果

年别	原始数据							原分类	计算后分类
	X_1	X_2	X_3	X_4	X_5	X_6	X_7		
1970	10.1	17.1	22.0	23.6	0.28	0.020	0.216	2	2
1971	9.7	17.1	22.3	24.3	1.05	0.068	0.427	3	3
1972	9.8	15.6	21.1	25.7	0.90	0.020	0.115	2	2
1973	10.6	15.2	21.3	25.4	0.47	0.014	0.178	2	2
1974	9.0	16.9	20.2	25.0	0.43	0.034	0.254	3	3
1975	12.3	17.4	22.3	24.4	0.31	0.006	0.221	2	2
1976	8.4	16.3	20.1	23.2	0.56	0.008	0.099	1	1
1977	10.4	17.3	21.8	25.5	0.08	0.003	0.210	1	1
1978	10.2	16.6	22.8	25.3	0.12	0.005	0.113	1	1
1979	7.5	17.2	21.8	23.8	0.25	0.024	0.058	1	1

将10年的资料回代判别函数，结果判别全部正确，又对1980～1982年进行预报，结果1980年判别为一类散发，实际也是一类散发；1981年判别为三类流行，实际为二类小流行；1982年判别为二类小流行，实际也是二类小流行；预报效果基本满意。

六、其他：还有模糊聚类法等。

疾病的定量预测

一、直线预测模型：对于有趋向性的时间序列，

如果时间不长，可以用直线来拟合，而把其波动性作为剩余误差来处理。

〔例4〕某市市区1973～1985年肺癌的标化死亡率如下表，试求其直线预测方程，并预测1986年的肺癌标化死亡率。

表4 某市1973～1985年肺癌标化死亡率

年别	肺癌标化死亡率(/10万)	年别	肺癌标化死亡率(/10万)
1973	83.8	1981	99.4
1975	86.2	1983	111.9
1977	92.0	1985	117.2
1979	95.1		

应用最小二乘法求直线方程 $\hat{Y}=a+bx$。首先以年别为x，以肺癌标化死亡率为Y，求出 $\Sigma x=13853$，$\Sigma x^2=27415199$，$\Sigma Y=685.6$，$\Sigma Y^2=68098.7$，$\Sigma XY=1357120.4$，$n=7$，回归系数b的计算用下式（公式2）：

$$b=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{n}}{\Sigma X^2-\frac{(\Sigma X)^2}{n}}$$

$$=\frac{1357120.4-\frac{13853\times685.6}{7}}{27415199-\frac{(13853)^2}{7}}=2.8393$$

年别平均数为

$$X=\frac{\Sigma X}{n}=\frac{13853}{7}=1979$$

肺癌死亡率的平均数为

$$\bar{Y}=\frac{\Sigma Y}{n}=\frac{685.6}{7}=97.9429$$

截距a的计算用下式（公式3）：

$$a=\bar{Y}-b\bar{X}=97.9429-2.8393\times1979$$

$$=-5521.0036$$

直线预报方程为

$$\hat{Y}=-5521.0036+2.8393X$$

令X=1986，得1986年该市市区预期肺癌标化死亡率的估计值为 $\hat{Y}=-5521.0036+2.8393\times1986=117.82/10万$

$S_{Y.X}$为各Y值距回归线的标准差，它是衡量预测误差大小的统计指标，用下式（公式4）计算：

$$S_{Y.X}=\sqrt{\left[\Sigma Y^2-\frac{(\Sigma Y)^2}{n}-\frac{(\Sigma XY-\frac{\Sigma X\Sigma Y}{n})^2}{\Sigma X^2-\frac{(\Sigma X)^2}{n}}\right]\div(n-2)}$$

$$=\sqrt{\left[68098.7-\frac{(685.6)^2}{7}\frac{(1357120.4-\frac{13853\times685.6)^2}{7}}{2741.5199-\frac{(13853)^2}{7}}\right]\div(7-2)}$$

$$=3.04$$

$S_{Y.X}$越小，预测越精确。

二、指数曲线预测模型：有些传染病已有行之有效的对策，所以其年发病率的时间序列呈单调下降，而下降的速度又往往与其当时的发病率呈正比，对于这一类型疾病的预测方程，可采用指数曲线的数学模型。

〔例5〕某地百日咳由于开展了计划免疫，1980～1984年发病率逐年下降，年发病率（/10万）依次为62.82、51.25、42.07、32.62、21.06。试求其指数曲线预测模型，并求1985年该病年发病率的预测值。

为了简化计算，把年度取其缩减值=年度—1980。求各年发病率Y的自然对数lnY，结果列于表的第4行。分别求得n=5，$\Sigma x=10$　$\bar{X}=2$，$\Sigma x_2=30$，

表5 某地百日咳1980～1984年的年发病率(/10万)

年度	年度缩减值X	年发病率Y	LnY=Y′	年发病率的估计值$\hat{Y}$	剩余误差
1980	0	62.82	4.1403	66.50	−3.68
1981	1	51.25	3.9367	51.08	0.17
1982	2	42.07	3.7393	39.24	2.83
1983	3	32.63	3.4849	30.14	2.48
1984	4	21.06	3.0474	23.15	−2.09
1985	5	(14.36)		17.79	−3.43

$\Sigma Y'=18.348625$，$\bar{Y}'=3.669725$，$\Sigma XY'=34.059664$，$\Sigma Y^2=68.053424$，将所得有关值代入公式2得：

$$b=\frac{34059664-\frac{10\times 18.348625}{5}}{30-\frac{(10)^2}{5}}=-0\ 2637586$$

应用公式2,得:

$$a=\bar{Y}-b\bar{X}=3.669725-(-0.2637586)\times 2$$
$$=4.1972422$$

所求的指数曲线预测模型为下式(公式5)得:

$$\hat{Y}=e^{bX+a}=e^{-0.2637586X+4.1972422}$$

预测1985年，该年缩减值x＝1985－1980＝5，代入上式得：$\hat{Y}=17.79$

即1985年预测百日咳的年发病率为17.79/10万，结果实际其年发病率为14.36/10万，二者基本吻合。

预测误差大小可用相关指数来考察，用下式（公式6）：

$$R^2=1-\frac{\Sigma(Y-\hat{Y})}{\Sigma(Y-\bar{Y})}$$

本例，$R^2=1-\frac{32.0987}{1045.5013}=0.9693$，$R^2$接近于1，说明误差较小。若再预测1986年，宜把1985年的发病率加进去，用1980～1985年疫情重新建立预测模型，这样误差小。

三、应用多元逐步回归分析建立预报方程：用于根据影响疾病流行的因素（预报因子）对疾病的流行强度进行定量预测。逐步回归分析的原理是按预报因子对疾病流行的作用大小，由大到小依次逐个引入回归方程，每引入一个因子，都要对方程中每个因子作F检验，当发现作用无显著意义的因子即予剔除，每剔除一个因子后，还要对留在方程中的因子逐个检验，这样引剔反复，直到无因子可再引入，亦无因子需要剔除为止。由于计算冗繁，本法通常应用电子计算机程序上机计算。

〔例6〕应用某省1970～1978年的流脑疫情及气象资料（如下表），采用逐步回归分析方法建立流脑预报方程（规定选剔预报因子的F界值为2.25）。

表6　　原始数据及用预报方程测得值与实际值的比较

测查类别	年别	X_1	X_2	X_3	X_4	X_5	X_6	X_7	流脑发病率(/10万) 实际值	流脑发病率(/10万) 预测值	差数
	1970	1.737	0.273	0.371	0.637	22.4	5.7	6.5	22.15	22.65	−0.50
复	1971	1.914	0.172	0.330	0.947	17.5	22.0	1.7	21.85	20.77	1.08
	1972	1.705	0.160	0.300	0.645	12.7	9.3	12.2	23.70	26.02	−2.32
	1973	1.524	0.189	0.428	0.775	21.4	0.8	3.2	18.90	20.08	−1.18
	1974	1.637	0.204	0.266	0.630	17.7	0.6	0.4	27.94	24.92	3.02
	1975	2.774	0.153	0.385	0.929	0.7	0	9.8	47.93	45.84	2.09
	1976	2.642	0.174	0.257	0.914	17.5	18.6	7.2	36.41	37.98	−1.57
测	1977	1.717	0.151	0.255	0.638	73.6	4.3	3.6	26.25	28.01	−1.76
	1978	0.882	0.131	0.171	0.376	20.7	31.7	4.2	10.94	9.78	1.16
预	1979	0.549	0.073	0.151	0.274	1.1	13.7	0.9	10.21	11.52	−1.31
	1980	0.912	0.052	0.110	0.259	36.1	32.4	2.3	16.17	16.37	−0.20
	1981	1.490	0.095	0.185	0.444	3.7	19.5	4.6	28.15	26.05	2.10
测	1982	1.742	0.177	0.316	0.652	48.0	18.6	9.3	21.48	23.74	−2.26

注：x_1～x_4分别为当年1月，前1年10～12月流脑月发病率，x_5～x_7分别为前一年11～12月，当年1月的降水量

经电子计算机进行逐步回归分析，在七个预报因子中最后选中四个，得预报方程为：

$$\hat{Y}=21.7235x_1-44.9128x_2-20.7680x_4-$$
$$0.2345X6+11.7477$$

该方程的剩余标准差为2.6726，复相关系数为0.9841，应用此方程对1970～78年进行复测，并对1979～82年进行预测，结果满意（表6）。