

组合模型对恶性肿瘤死亡率拟合度评价及预测方法的研究

曲红梅 白亚娜 魁发瑞 胡晓斌 裴鸿波 任晓卫 申希平

730030 兰州,西北民族大学医学院预防医学教研室(曲红梅); 730000 兰州大学公共卫生学院流行病与统计学教研室(白亚娜、胡晓斌、裴鸿波、任晓卫、申希平); 730030 兰州,西北民族大学医院外科(魁发瑞)

通信作者:白亚娜, Email: baiyana@lzu.edu.cn

DOI: 10.3760/cma.j.issn.0254-6450.2017.01.022

【摘要】 目的 用6种单项预测方法对金昌队列13年恶性肿瘤死亡率进行拟合并预测。**方法** 采用动态数列、线性回归、指数平滑、自回归移动平均(ARIMA)模型、灰色模型、Joinpoint回归6种方法,利用金昌队列2001—2013年数据进行恶性肿瘤死亡率拟合及方法比较。采用组合模型进行模型优化,基于算术平均法、方差倒数法、均方误差倒数法、简单加权平均法计算组合模型权重系数。**结果** 对恶性肿瘤死亡率以Joinpoint线性回归拟合精度最高(87.64%),线性回归法、动态数列、GM(1,1)模型、指数平滑法、ARIMA(1,0,0)拟合精度分别为87.32%、86.99%、86.25%、85.72%、81.98%。基于灰色模型与线性回归的组合模型预测精度(>99%)高于基于ARIMA(1,0,0)与灰色模型的组合模型,其中算术平均法和均方误差倒数法权重系数组合模型(灰色模型与线性回归)拟合效果最好。**结论** 组合模型预测恶性肿瘤死亡率优于单项预测法,预测精度>95%。

【关键词】 恶性肿瘤; 死亡率; 组合模型

基金项目: 中央高校基本科研业务项目(31920150047); 西北民族大学引进人才项目(XBMUYJYC201626)

Effect of combination model on fitting cancer mortality and prediction Qu Hongmei, Bai Yana, Kui Farui, Hu Xiaobin, Pei Hongbo, Ren Xiaowei, Shen Xiping

Preventive Medicine Teaching and Research Section, Medical College, Northwest University for Nationalities, Lanzhou 730030, China (Qu HM); Center for Cancer Prevention and Control of Lanzhou University, Teaching and Research Section of Epidemiology and Biostatistics, School of Public Health, Lanzhou University, Lanzhou 730000, China (Bai YN, Hu XB, Pei HB, Ren XW, Shen XP); Department of Surgery, The Hospital of Northwest University for Nationalities, Lanzhou 730030, China (Kui FR)

Corresponding author: Bai Yana, Email: baiyana@lzu.edu.cn

【Abstract】 Objective To reduce the cancer burden in the Jinchang cohort and provide evidence for developing cancer prevention strategies and performing effectiveness evaluation in the Jinchang cohort. We are fitting thirteen years of cancer mortality data from the Jinchang cohort by using six kinds of predicting methods to compare relative fitness and to select good predicting methods for the prediction of cancer mortality trends. **Methods** The mortality data of cancer in Jinchang cohort from 2001–2013 were fitted using six kinds of predicting methods: dynamic series, linear regression, exponential smoothing, autoregressive integrated moving average (ARIMA) model, grey model (GM), and Joinpoint regression. Weight coefficients of combination models were calculated by four methods: the arithmetic average method, the variance inverse method, the mean square error inverse method, and the simple weighted average method. **Results** The cancer mortality was fitted and compared by using six kinds of forecasting methods; the fitting precision of the Joinpoint linear regression had the highest accuracy (87.64%), followed by linear regression (87.32%), the dynamic series (86.99%), GM (1, 1) (86.25%), exponential smoothing (85.72%) and ARIMA (1, 0, 0) (81.98%), respectively. Prediction accuracy of the combination model derived from GM (1, 1) and linear regression (>99%) was higher than that of the combination model derived from ARIMA (1, 0, 0) and GM (1, 1). The combination model derived from the GM (1, 1) and linear regression, with weight coefficients based on the arithmetic average method and the mean square error inverse method, had the

best prediction effect of the four weight calculation methods. **Conclusion** Prediction accuracy of the combination model, with accuracy >95%, was higher than that of the single prediction methods.

【Key words】 Cancer; Mortality rate; Combination model

Fund programs: Research Fund Central University-Northwest University for Nationalities (31920150047); Project of "Bringing in Talent" in Northwest University for Nationalities (XBMUYJYC201626)

为了解金昌队列恶性肿瘤未来发展趋势,制定切实可行的恶性肿瘤防治策略,应用6种预测方法对该队列2001—2013年恶性肿瘤死亡率拟合,并探讨恶性肿瘤预测方法,最终遴选单项模型建立组合模型并对恶性肿瘤死亡率进行趋势预测。

资料与方法

应用金昌队列恶性肿瘤死亡率2001—2013年恶性肿瘤死亡率数据来建立死亡率预测模型,并预测2014—2016年恶性肿瘤死亡率。

1. 模型的遴选:通过动态数列^[1]、线性回归^[2]、指数平滑^[3]、自回归移动平均(ARIMA)模型^[4]、灰色模型[GM(1,1)]^[5]、Joinpoint回归(Joinpoint regression)^[6]6种模型进行单项模型的拟合并评价模型的方法学,遴选拟合精度较高的预测模型。恶性肿瘤死亡率与年份序列值之间呈线性关系,因此资料满足以年份序列值为自变量 x ,恶性肿瘤死亡率为因变量 y 做曲线拟合的条件。利用年份与恶性肿瘤死亡率建立Joinpoint回归,并对2001—2013年队列人群恶性肿瘤死亡率进行拟合。

2. 模型评价:为了评价预测方法的准确性,即拟合的精确度,本研究采用4个误差度量指标评价拟合效果^[7]:均方误差(MSE)、平均绝对误差(MAE)、平均绝对百分误差(MAPE)和拟合精度。

3. 模型的优化组合:组合模型(combination model)是将 ≥ 2 种模型赋予适当的权重系数进行预测,以提高预测的精度。①组合模型权重系数的计算基于算术平均法、方差倒数法、均方误差倒数法、简单加权平均法4种方法^[8];②组合模型的拟合精度评价采用以上模型评价方法;③模型的预测是基于2001—2013年恶性肿瘤死亡率数据预测2014—2018年恶性肿瘤死亡率。

4. 统计学分析:采用Excel 2010、SPSS 18.0软件进行动态数列、GM(1,1)、线性回归、指数平滑、ARIMA模型的预测分析。利用美国国家癌症研究所研发的Joinpoint Regression Program 4.1.1.3软件进行Joinpoint回归分析。检验水准 $\alpha=0.05$ 。

结果

1. 单项拟合预测方法的效果评价:采用6种单

项预测方法拟合的2001—2013年恶性肿瘤死亡率见表1。6种预测单项方法用于恶性肿瘤死亡率的预测方法的比较,Joinpoint线性回归拟合精度最高(87.64%),线性回归、动态数列、GM(1,1)、指数平滑、ARIMA模型拟合精度分别为87.32%、86.99%、86.25%、85.72%、81.98%(表2)。

表1 采用6种单项预测方法拟合2001—2013年金昌队列恶性肿瘤死亡率(/10万)

年份	恶性肿瘤死亡率	动态数列	线性回归	指数平滑	ARIMA模型(1,0,0)	GM(1,1)	Joinpoint回归
2001	161.86	-	153.40	160.71	224.95	-	149.10
2002	134.85	169.99	165.26	161.38	187.21	171.97	160.85
2003	193.27	178.11	177.13	145.97	171.06	180.87	172.60
2004	182.95	186.24	188.99	173.44	206.00	190.22	184.35
2005	201.72	194.36	200.86	178.96	199.83	200.05	196.10
2006	194.73	202.49	212.72	192.18	211.06	210.39	207.86
2007	176.61	210.61	224.59	193.66	206.88	221.27	219.61
2008	301.86	218.74	236.45	183.76	196.04	232.71	231.36
2009	289.18	226.86	248.32	252.34	270.95	244.74	243.11
2010	315.32	234.99	260.18	273.73	263.37	257.40	254.86
2011	228.94	243.11	272.05	297.88	279.01	270.71	266.61
2012	246.87	251.24	283.91	257.85	227.34	284.70	278.36
2013	291.43	259.36	295.78	251.47	238.06	299.42	290.11

表2 6种单项预测方法拟合恶性肿瘤死亡率评价(/10万)

预测方法	MSE	MAE	MAPE	拟合精度
动态数列	1 769.99	31.59	13.01	86.99
线性回归	1 242.86	28.75	12.68	87.32
指数平滑	2 104.08	34.10	14.28	85.72
ARIMA(1,0,0)	2 223.02	39.09	18.02	81.98
GM(1,1)	1 431.51	31.49	13.75	86.25
Joinpoint回归	1 267.81	28.47	12.36	87.64

2. 组合模型拟合及方法学评价:将ARIMA模型(1,0,0)与GM(1,1)、GM(1,1)与线性回归建立组合模型进行模型优化,2001—2013年恶性肿瘤死亡率组合预测模型拟合值见表3。结果提示GM(1,1)与线性回归组合模型4种赋予权重的方法,MAE、MAPE及MSE均小于ARIMA(1,0,0)与GM(1,1)的组合方式。GM(1,1)与线性回归组合方式优于ARIMA(1,0,0)与GM(1,1)组合方式。在两种模型组合方式中,算术平均法与均方误差倒数法预测效果较好。

如表4所示,恶性肿瘤死亡率组合模型预测精度>90%,均高于上述任一单项预测方法。其中基

表3 组合预测方法拟合2001—2013年金昌队列恶性肿瘤死亡率(/10万)

年份	死亡率	组合预测11	组合预测12	组合预测13	组合预测14	组合预测21	组合预测22	组合预测23	组合预测24
2001	161.86	-	-	-	-	-	-	-	-
2002	134.85	166.68	167.89	167.29	168.44	168.62	168.52	168.57	167.50
2003	193.27	163.42	167.40	165.44	169.23	179.00	178.94	178.97	178.37
2004	182.95	181.83	183.75	182.80	184.62	189.60	189.59	189.60	189.40
2005	201.72	189.51	191.92	190.73	193.02	200.45	200.47	200.46	200.59
2006	194.73	201.29	203.37	202.34	204.32	211.56	211.59	211.58	211.95
2007	176.61	207.47	210.62	209.06	212.07	222.93	222.98	222.96	223.48
2008	301.86	208.24	213.83	211.07	216.40	234.58	234.64	234.61	235.21
2009	289.18	248.54	247.67	248.10	247.28	246.53	246.59	246.56	247.13
2010	315.32	265.56	263.70	264.62	262.84	258.79	258.83	258.81	259.25
2011	228.94	284.29	281.19	282.72	279.76	271.38	271.40	271.39	271.60
2012	246.87	271.28	274.34	272.83	275.75	284.31	284.30	284.30	284.18
2013	291.43	275.45	280.92	278.22	283.44	297.60	297.54	297.57	296.99

表4 恶性肿瘤死亡率(/10万)组合预测方法拟合效果评价

预测方法	MSE	MAE	MAPE	拟合精度
组合预测11	200.63	12.67	6.25	93.75
组合预测12	302.79	15.56	7.67	92.33
组合预测13	249.78	14.14	6.97	93.03
组合预测14	356.68	16.89	8.33	91.67
组合预测21	2.66	1.42	0.66	99.34
组合预测22	2.83	1.46	0.68	99.32
组合预测23	2.74	1.44	0.67	99.33
组合预测24	4.73	1.89	0.88	99.12

于4种权重计算法的GM(1,1)与线性回归的组合方式MSE、MAE、MAPE均小于ARIMA(1,0,0)与GM(1,1)组合方式(基于4种权重系数算法)。GM(1,1)与线性回归的组合方式预测精度>99%,均高于ARIMA(1,0,0)与GM(1,1)组合方式。其中以基于算术平均法及均方误差倒数法的权重计算[GM(1,1)与线性回归]组合模型预测效果最好。

3. 恶性肿瘤死亡率组合预测:组合预测中,以GM(1,1)与线性回归的组合模型拟合效果最好,故用基于算术平均法、均方误差倒数法计算的权重系数的组合模型[GM(1,1)与线性回归]预测2014—2018年金昌队列人群恶性肿瘤死亡率,预测结果见表5、图1。

表5 组合模型预测2014—2018年金昌队列人群恶性肿瘤死亡率(/10万)

方法	2014年	2015年	2016年	2017年	2018年
算术平均法	311.27	325.35	339.84	354.78	370.18
均方误差倒数法	311.22	325.26	339.71	354.60	369.95

讨论

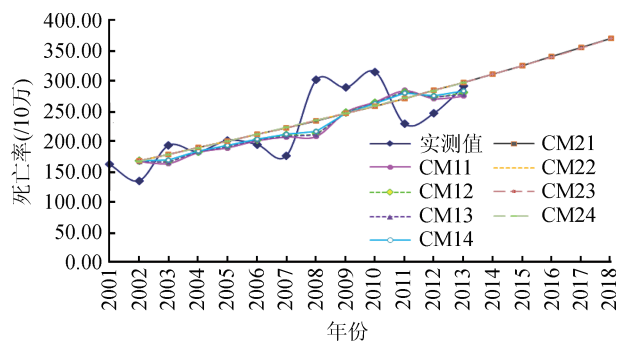
近年来,国内外有许多学者在恶性肿瘤死亡率的预测中使用多种单项预测方法。如动态数列平均增长速度^[1]、线性回归^[2]、曲线拟合^[9]、指数平滑^[3,10]、

ARIMA模型^[4,10]、GM(1,1)^[5,11-12]等,取得了较好的预测效果。其中动态数列平均增长速度、线性回归是恶性肿瘤趋势分析常用的方法,指数平滑、GM(1,1)、ARIMA模型、Joinpoint回归分析却较少应用^[6,13-15]。

本研究Joinpoint回归拟合精度达到87.64%,线性回归、动态数列、GM(1,1)、指数平滑、ARIMA模型(1,0,0)预测精度分别为87.32%、86.99%、86.25%、85.72%、81.98%。队列恶性肿瘤死亡率数据波动大,故在预测中指数平滑预测效果优于ARIMA模型(1,0,0),且因隐

含指数规律,故GM(1,1)预测效果较好。本研究Joinpoint回归模型拟合精度大于GM(1,1)、指数平滑法,与线性回归法拟合精度相当。Joinpoint回归分析是将年份死亡率曲线分为若干有统计学意义的区段,并对死亡率进行分段分析,分析开始于最小的折点,检验是否有一个或者更多个折点有统计学意义,双侧检验P<0.05认为折点有统计学意义。在2001—2013年金昌队列人群恶性肿瘤死亡率分析中,有0个折点,年份死亡率曲线均未被分段。国内外少数学者将Joinpoint回归分析用于癌症死亡率及发病率的研究^[16-17]。Joinpoint回归也提供了年百分比的变化(APC)、平均年百分比变化(AAPC)等指标。此外,还考虑年龄对死亡率或发病率的影响,将直线按年龄分段计算斜率或APC。因此,Joinpoint分析比传统线性回归更有实际意义。

本研究组合预测中,GM(1,1)与线性回归组合方式优于ARIMA模型(1,0,0)与GM(1,1)组合方式。不同曲线类型组合而成的模型精度高于同类型曲线的组合模型精度。GM(1,1)与线性回归组合、ARIMA模型(1,0,0)与GM(1,1)的组合中,算术平



注:CM为组合预测

图1 组合模型用于2001—2013年金昌队列恶性肿瘤死亡率的拟合及预测

均法与均方误差倒数法权重的计算预测效果较好^[18]。在以往恶性肿瘤预测中,组合模型应用较少^[9-10,19],且组合模型权重的计算多用Saaty权重法^[10]。其他领域组合模型研究表明^[19-20],组合模型精度高的先决条件,首先是参与组合的模型不是越多越好,模型数量为2~5个。故本研究选取了2种单项模型进行组合。均为GM(1,1)(指数曲线)与线性回归(线性)的组合优于其他组合方式。另外,组合模型权重系数的计算也很关键。本研究恶性肿瘤死亡率预测中,组合模型算术平均法与均方误差倒数法误差较小,预测效果优于单项预测法。

总之,本研究遴选出拟合精度较好的预测方法仅适合于金昌队列2001—2013年这个时段,因恶性肿瘤的发生和死亡受多种因素的影响,且有长期变化趋势,故随着时间变化,还需探讨不同的预测模型。

利益冲突 无

参 考 文 献

- [1] 鲍俊屹,白亚娜,尚慧,等.某职业队列人群中老年肺癌患者分布特征10年变化趋势[J].中国老年学杂志,2014,34(21):6166-6168. DOI:10.3969/j.issn.1005-9202.2014.21.104.
Bao JY, Bai YN, Shang H, et al. 10-year trend of the distribution characteristics of an occupational cohort of elderly patients with lung cancer [J]. Chin J Gerontol, 2014, 34 (21) : 6166-6168. DOI:10.3969/j.issn.1005-9202.2014.21.104.
- [2] 高姗.上海市区原发性肝癌的流行病学研究[D].上海:复旦大学,2011.
Gao S. Epidemiological studies of primary liver cancer in urban Shanghai [D]. Shanghai: Fudan University, 2011.
- [3] 杨昌跃.大姚县青石棉污染区肺癌和间皮瘤死亡率的调查及其趋势预测[D].成都:四川大学,2005.
Yang CY. Study on mortality and its trends of lung cancer and mesothelioma in the crocidolite contaminated area in Dayao county [D]. Chengdu: Sichuan University, 2005.
- [4] 何庭辉,柳青.广州市荔湾区恶性肿瘤死亡率时间序列预测模型分析[J].疾病控制杂志,2006,10(6):630-632. DOI:10.3969/j.issn.1674-3679.2006.06.025.
He TH, Liu Q. Analysis on the predictive model of the time series of cancer mortality in Liwan district, Guangzhou [J]. Chin J Dis Control Prev, 2006, 10 (6) : 630-632. DOI: 10.3969/j.issn.1674-3679.2006.06.025.
- [5] 洪光烈,宋红兵,汪丛峰,等.基于灰色数列GM(1,1)模型的恶性肿瘤死亡趋势的预测及分析[J].中国肿瘤,2011,20(7):503-505.
Hong GL, Song HB, Wang CF, et al. A study of GM(1,1) model for predicting the mortality trends of malignant tumors [J]. China Cancer, 2011, 20(7) : 503-505.
- [6] Ilic M, Ilic I. Gynaecological cancer mortality in Serbia, 1991-2010: a joinpoint regression analysis [J]. Asian Pacific J Cancer Prev, 2015, 16(1) : 157-162. DOI: 10.7314/APJCP.2015.16.1.157.
- [7] 蒋传进.基于模型遴选规则的自适应组合预测研究[D].上海:东华大学,2014.
Jiang CJ. The research of adaptive combination forecasting based on rule guideline of selecting models [D]. Shanghai: Donghua University, 2014.
- [8] 戴华娟.组合预测模型及其应用研究[D].长沙:中南大学,2007.
Dai HJ. Combination forecasting model and application research [D]. Changsha: Central South University, 2007.
- [9] 韩兢,王洁贞,胡平,等.恶性肿瘤死亡率预测模型的探讨[J].中国卫生统计,1998,15(5):5-8.
Han J, Wang JZ, Hu P, et al. An inquiry of the forecast model of malignant tumor mortality [J]. Chin J Health Stat, 1998, 15(5) : 5-8.
- [10] 陈勇.启东1972—2001年恶性肿瘤发病趋势分析及预测模型研究[D].南京:南京医科大学,2005.
Chen Y. Trend analysis and prediction model of malignant tumor in Qidong 1972-2001 [D]. Nanjing: Nanjing Medical University, 2005.
- [11] 孙彬,何穗智.灰色系统残差与Verhulst模型预测恶性肿瘤死亡[J].预防医学情报杂志,2011,27(8):604-607.
Sun B, He SZ. Application of the grey system residual error model and grey Verhulst model on forecasting malignant tumor death [J]. J Prev Med Inf, 2011, 27(8) : 604-607.
- [12] 潘凤莲,陈哲.GM(1,1)模型在子宫颈癌死亡率预测中的应用[J].海南医学院报,2013,19(1):92-95. DOI:10.13210/j.cnki.jhmu.2013.01.031.
Pan FL, Chen Z. Application of GM(1,1) model in predicting mortality rate of cervical cancer [J]. J Hainan Med Univ, 2013, 19 (1) : 92-95. DOI: 10.13210/j.cnki.jhmu.2013.01.031.
- [13] Cayuela A, Rodríguez-Domínguez S, López-Campos JL, et al. Joinpoint regression analysis of lung cancer mortality, Andalusia 1975-2000 [J]. Ann Oncol, 2004, 15 (5) : 793-796. DOI: 10.1093/annonc/mdh170.
- [14] 王庆生,陈万青,郑荣寿,等.癌症年龄别发病率的Joinpoint线性回归分析及其在癌症防控中的意义[J].中国肿瘤,2013,22(3):180-185. DOI:10.11735/j.issn.1004-0242.2013.03.A201302046.
Wang QS, Chen WQ, Zheng RS, et al. Joinpoint linear regression analysis of cancer age-specific incidence rate and its significance for cancer control [J]. Chin Cancer, 2013, 22(3) : 180-185. DOI: 10.11735/j.issn.1004-0242.2013.03.A201302046.
- [15] 陈万青,郑荣寿,曾红梅,等.中国恶性肿瘤发病趋势分析和预测[J].中华预防医学杂志,2012,46(7):581-586. DOI:10.3760/cma.j.issn.0253-9624.2012.07.002.
Chen WQ, Zheng RS, Zeng HM, et al. Trend analysis and prediction of cancer incidence in China [J]. Chin J Prev Med, 2012, 46 (7) : 581-586. DOI: 10.3760/cma.j.issn.0253-9624.2012.07.002.
- [16] Dehkordi ZF, Tazhibi M, Babazade S. Application of joinpoint regression in determining breast cancer incidence rate change points by age and tumor characteristics in women aged 30-69 (years) and in Isfahan city from 2001 to 2010 [J]. J Educ Health Promot, 2014, 3 : 115. DOI: 10.4103/2277-9531.145917.
- [17] López-Campos JL, Ruiz-Ramos M, Soriano JB. Mortality trends in chronic obstructive pulmonary disease in Europe, 1994-2010: a joinpoint regression analysis [J]. Lancet Respir Med, 2014, 2 (1) : 54-62. DOI:10.1016/S2213-2600(13)70232-7.
- [18] 唐小我.最优组合预测方法及其应用[J].数理统计与管理,1992,11(1):31-35. DOI:10.13860/j.cnki.sljt.1992.01.007.
Tang XW. The optimal combined forecasting method and its application [J]. Appl Stat Manag, 1992, 11 (1) : 31-35. DOI: 10.13860/j.cnki.sljt.1992.01.007.
- [19] 陈亦新,田辉.应用组合模型分析预测恶性肿瘤死亡率[J].黑龙江医学,2000(5):24-25. DOI:10.3969/j.issn.1004-5775.2000.05.024.
Chen YX, Tian H. Analysis and prediction malignant tumor mortality by combination model [J]. Heilongjiang Med J, 2000 (5) : 24-25. DOI: 10.3969/j.issn.1004-5775.2000.05.024.
- [20] 姜明辉.商业银行个人信用评估组合预测方法研究[D].哈尔滨:哈尔滨工业大学,2006.
Jiang MH. Study on combining forecast methods for personal credit evaluation in commercial banks [D]. Harbin: Harbin Institute of Technology, 2006.

(收稿日期:2016-04-21)

(本文编辑:张林东)