基于小样本临床指标数据的引产预测模型构建

秦雅莉[1]，姚莉萍[2]，袁玲[3]，陈胜[1]

1. 上海理工大学 光电信息与计算机工程学院（上海 200093）

2. 上海市第一妇婴保健院 超声科（上海 201204）

3. 上海市第一妇婴保健院 产科（上海 201204）

附件 1 效应量计算公式和取值依据

Supplement 1 Formulas for calculating the effects and basis of values

效应大小是统计学中用来量化变量之间关系强度的指标[1-2]。一般来说，效应量分为小、中、大三种，分别约等于 0.2、0.5 和 0.8[3-4]。效应量为 0.6 介于中效应量和大效应量之间，表明预测模型在区分引产结果方面具有很强的鉴别能力。此外，在医学研究中，效应量为 0.6 表明预测变量对引产结果的影响是显著的，具有临床意义。本研究采用 t 检验对样本量进行先验分析，效应大小的计算公式如下：

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{pooled}} \tag{1}$$

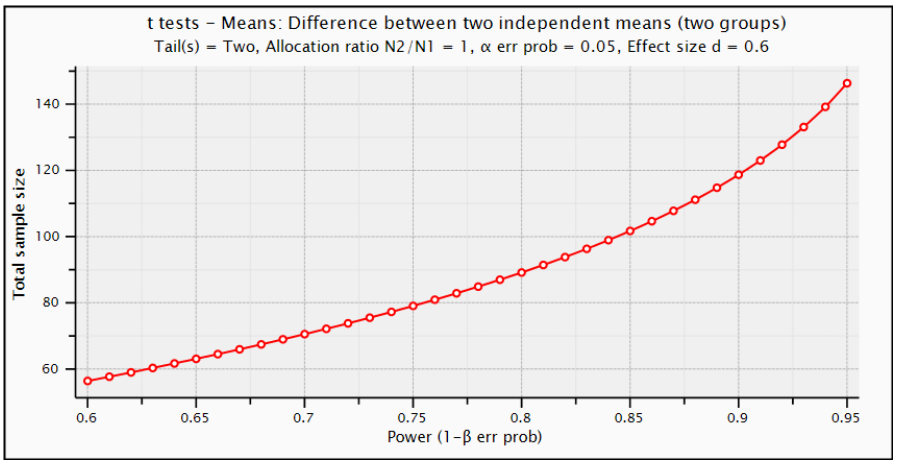
其中，$\overline{X}_1$ 和 $\overline{X}_2$ 为两组的均值，S_{pooled} 为两组的综合标准偏差，计算公式为:

$$S_{pooled} = \sqrt{\frac{(n_1-1)s_1^{\,2} + (n_2-1)s_2^{\,2}}{n_1+n_2-2}} \tag{2}$$

其中，s_1 和 s_2 是两组的标准差，n_1 和 n_2 是两组相应的样本量。

附件 2 样本容量先验分析

Supplement 2 Sample size a priori analysis



附件 3 SVM 模型中基于 Hinge Loss 的损失函数

Supplement 3　Hinge Loss-based loss function in the SVM model

SVM 采用的损失函数公式为：

$$J(\omega,b)=\lambda\|\omega\|^2+C\sum_{i=1}^{n}\max\left[0,1-y_i(\omega^T x_i+b)\right] \tag{3}$$

其中，$\|\omega\|^2$ 为 L2 范数的平方，L2 范数公式见式（4）。在优化问题中，为了便于计算梯度和求解，通常会将 L2 范数的平方除以 2，即 $\lambda=\frac{1}{2}$。$\max\left[0,1-y_i(\omega^T x_i+b)\right]$ 为 Hinge Loss 函数。ω 是模型的权重向量，b 为偏置项，C 是正则化参数，n 为样本数量，(x_i,y_i) 是训练集中的样本点，x_i 为特征向量，y_i 为对应的类别标签。

$$\|R\|_2=\sqrt{\sum_{i=1}^{n}r_i^2} \tag{4}$$

L2 范数公式中的 r_i 是向量 R 的第 i 个元素，n 是向量的维度。

附件 4 FCNN 模型运算过程

Supplement 4　The computation process of the FCNN model

FCNN 模型运算过程如下：

第 l 层输入层到隐藏层的传播为：

首先，进行线性变换，将输入 $a^{[l-1]}$ 与权重矩阵 $W^{[l]}$ 相乘并加上偏置向量 $b^{[l]}$，即：

$$z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]} \tag{5}$$

然后，将第 l 层的线性变换结果 $z^{[l]}$ 输入到 ReLU 激活函数，得到：

$$a^{[l]} = \mathrm{ReLU}(z^{[l]}) = \max(0, z^{[l]}) \tag{6}$$

最后，使用 Dropout 层对第 l 层的输出结果 $a^{[l]}$ 进行操作，并应用保留率 p，即：

$$a_{\mathrm{dropout}}^{[l]} = \mathrm{dropout}(a^{[l]}, p) \tag{7}$$

其中，l 表示神经网络中的不同层，由 1 到 L，本研究中的 $L = 1$。

输出层的正向传播过程为：

首先，输出层的线性变换是通过将第 L 层的输出 $a_{\mathrm{dropout}}^{[l]}$ 与输出层的权重矩阵 $W^{[L+1]}$ 相乘，并且加上输出层的偏置向量 $b^{[L+1]}$ 实现的：

$$z^{[L+1]} = W^{[L+1]} a_{\mathrm{dropout}}^{[L]} + b^{[L+1]} \tag{8}$$

最后，将输出层的线性变换结果 $z^{[L+1]}$ 输入到 Sigmoid 激活函数 $\sigma(\cdot)$ 中，得到输出层的预测概率 $\hat{y}$：

$$\hat{y} = \sigma(z^{[L+1]}) = \frac{1}{1 + e^{-z^{[L+1]}}} \tag{9}$$

附件 5 55 维临床特征的 MIC 分值

Supplement 5　MIC scores for 55-dimensional clinical features

特征	MIC-Score	特征	MIC-Score
HR-前	0.403	无 FGR	0.035
EOS-前	0.352	FGR	0.035
HR-后	0.300	无 ARM	0.033
CL-前	0.297	ARM	0.033
CL-后	0.255	HDP	0.031
ECI-前	0.251	无 HDP	0.031
IOS-前	0.242	宫颈位置-前	0.026
IOS/EOS Ratio-后	0.206	宫颈容受-前	0.025
IOS/EOS Ratio-前	0.203	12-27+6 周引产次数	0.018
分娩孕周	0.192	AID-CID	0.018
先露高低-后	0.177	无 ICP	0.018
孕前 BMI	0.172	ICP	0.018
引产孕周	0.170	Bishop 评分-前	0.016
EOS-后	0.156	AID-VTE	0.009
ECI-后	0.151	无 PROM	0.009
IOS-后	0.141	PROM	0.009
早孕人流次数	0.131	先露高低-前	0.007
产次	0.111	无 GDM	0.007
宫腔镜手术次数	0.101	GDM	0.007
年龄/岁	0.078	宫颈质地-前	0.006
Bishop 评分-后	0.076	PA	0.006
孕次	0.066	无 PA	0.006
宫颈位置-后	0.039	宫颈容受-后	0.005
无 APS	0.035	注射催产素	0.002
APS	0.035	未注射催产素	0.002
AID-SS	0.035	无 AID	0.002
无血小板少	0.035	AID-UCTD	0.000
血小板少	0.035		

注：“-前”表示球囊放置之前的指标，“-后”表示球囊放置之后的指标

附件 6 连续变量的 *t* 检验分析

Supplement 6 *t*-test analysis of continuous variables

变量（连续）	引产失败（*n*=42）	引产成功（*n*=48）	*P* 值
年龄	32.07±3.910	30.92±4.191	0.182
孕前 BMI	20.943±1.995	21.861±2.842	0.084
引产孕周	39.111±1.118	39.797±1.032	0.003*
分娩孕周	39.365±1.132	39.971±1.024	0.009*
CL-前	33.455±6.299	28.779±7.063	0.001*
ECI-前	3.010±1.277	3.494±0.990	0.046*
HR-前	58.780±18.831	45.041±15.424	≤0.001*
IOS-前	0.350±0.118	0.392±0.104	0.078
EOS-前	0.304±0.102	0.392±0.107	≤0.001*
IOS/EOS Ratio-前	1.165（0.230）	1.040（0.370）	0.022*
CL-后	24.800（6.600）	21.525（15.900）	0.004*
ECI-后	4.226±1.180	4.154±1.309	0.786
HR-后	40.765（13.110）	33.125（21.590）	0.004*
IOS-后	0.406±0.088	0.438±0.118	0.162
EOS-后	0.396±0.104	0.407±0.105	0.631
IOS/EOS Ratio-后	1.143±0.356	1.082±0.262	0.350

注：*表示双尾 *P* 值小于 0.05；“-前”表示球囊放置之前的指标，“-后”表示球囊放置之后的指标。

由结果可知，共计 9 项临床指标具有统计学意义：引产孕周、分娩孕周，球囊放置前的 CL、ECI、HR、IOS/EOS ratio、EOS，以及球囊放置后的 CL、HR。

附件 7 分类变量的 χ^2 检验分析

Supplement 7 χ^2 *-test* analysis of categorical variables

变量（分类）		引产失败（*n*=42）	引产成功（*n*=48）	*P* 值
GDM	无	39（92.9%）	41（85.4%）	0.327
	有	3（7.1%）	7（14.6%）	
HDP	无	42（100%）	0（0%）	0.058
	有	0（0%）	5（10.4%）	
ICP	无	36（85.7%）	46（95.8%）	0.139
	有	6（14.3%）	2（4.2%）	
PA	无	42（100%）	47（97.9%）	1.000
	有	0（0%）	1（2.1%）	
APS	无	39（92.9%）	3（7.1%）	0.098
	有	48（100%）	0（0%）	
血小板少	无	39（92.9%）	3（7.1%）	0.098
	有	48（100%）	0（0%）	
AID	无	33（78.6%）	40（83.3%）	0.114
	UCTD	3（7.1%）	4（8.3%）	
	CID	0（0%）	3（6.3%）	
	VTE	3（7.1%）	1（2.1%）	
	SS	3（7.1%）	0（0%）	
FGR	无	39（92.9%）	48（100%）	0.098
	有	3（7.1%）	0（0%）	
ARM	无	15（35.7%）	7（14.6%）	0.020*
	有	27（64.3%）	41（85.4%）	
催产素	无	3（7.1%）	2（4.2%）	0.661
	有	39（92.9%）	46（95.8%）	
PROM	无	39（92.9%）	47（97.9%）	0.336
	有	3（7.1%）	1（2.1%）	

注：*表示双尾 *P* 值小于 0.05；*“-前”表示球囊放置之前的指标，“-后”表示球囊放置之后的指标

由结果可知，分类变量中的 ARM 具有统计学意义。

附件 8 有序变量的 Mann-Whitney *U* 检验分析

Supplement 8　Mann-Whitney *U*-test analysis of ordered variables

变量（有序）		引产失败（*n*=42）	引产成功（*n*=48）	*P* 值
宫颈容受-前	0	0（0%）	2（4.2%）	1.000
	1	42（100%）	44（91.7%）	
	2	0（0%）	2（4.2%）	
宫颈质地-前	0	0（0%）	0（0%）	0.287
	1	18（42.9%）	26（54.2%）	
	2	24（57.1%）	22（45.8%）	
先露高低-前	0	39（92.9%）	41（85.4%）	0.265
	1	3（7.1%）	7（14.6%）	
宫颈位置-前	0	3（7.1%）	1（2.1%）	0.056
	1	39（92.9%）	44（91.7%）	
	2	0（0%）	3（6.3%）	
宫颈容受-后	0	0（0%）	0（0%）	0.373
	1	6（14.3%）	4（8.3%）	
	2	36（85.7%）	44（91.7%）	
先露高低-后	0	39（92.9%）	18（37.5%）	≤0.001*
	1	3（7.1%）	30（62.5%）	
宫颈位置-后	0	0（0%）	0（0%）	0.009*
	1	9（21.4%）	23（47.9%）	
	2	33（78.6%）	25（52.1%）	
Bishop 评分-前	2	3（7.1%）	1（2.1%）	0.809
	3	9（21.4%）	13（27.1%）	
	4	30（71.4%）	33（68.8%）	
	5	0（0%）	1（2.1%）	
	6	0（0%）	0（0%）	
	7	0（0%）	0（0%）	
Bishop 评分-后	2	0（0%）	0（0%）	0.001*
	3	0（0%）	0（0%）	
	4	0（0%）	0（0%）	
	5	12（28.6%）	4（8.3%）	
	6	30（71.4%）	37（77.1%）	
	7	0（0%）	7（14.6%）	
孕次	1	18（42.9%）	24（50%）	0.917
	2	18（42.9%）	12（25.0%）	

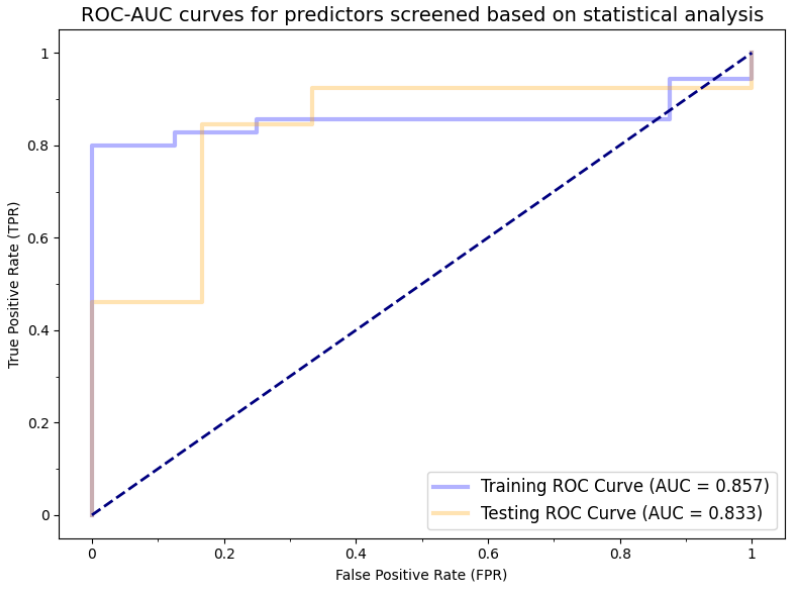
变量（有序）		引产失败（*n*=42）	引产成功（*n*=48）	*P* 值
	3	0（0%）	6（12.5%）	
	4	6（14.3%）	4（8.3%）	
	5	0（0%）	1（2.1%）	
	6	0（0%）	1（2.1%）	
12-27^{+6}周引产次数	0	42（100%）	45（93.8%）	0.101
	1	0（0%）	2（4.2%）	
	3	0（0%）	1（2.1%）	
宫腔镜手术次数	0	30（71.4%）	46（95.8%）	0.002*
	1	12（28.6%）	1（2.1%）	
	2	0（0%）	1（2.1%）	
早孕人流次数	0	18（42.9%）	29（60.4%）	0.141
	1	18（42.9%）	12（25.0%）	
	2	0（0%）	6（12.5%）	
	3	6（14.3%）	1（2.1%）	
产次	0	3（7.1%）	0（0%）	≤0.001*
	1	39（92.9%）	36（75.0%）	
	2	0（0%）	12（25.0%）	

注：*表示双尾 *P* 值小于 0.05；*“-前”表示球囊放置之前的指标，“-后”表示球囊放置之后的指标

由结果可知，在有序变量中，宫腔镜手术次数、产次，以及球囊放置后的先露高低、宫颈位置、Bishop 评分，共计 5 项临床指标具有统计学意义。

附件 9 基于统计分析筛选预测因子的 ROC 曲线

Supplement 9 ROC curves for predictors screened based on statistical analysis



参考文献

[1] Verma J P, Verma P. Determining sample size in experimental studies, F, 2020.

[2] Verma P, Verma J P. Determining sample size and power in research studies, F, 2020.

[3] Greene T. Randomized controlled trials 6: Determining the sample size and power for clinical trials and cohort studies. Methods in molecular biology, 2021, 2249: 281-305.

[4] Serdar C C, Cihan M, Yücel D, et al. Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia medica, 2021, 31(1): 010502.