·论著·

希望区域法在临床试验适应性设计中的准确性和稳健性研究*

中国人民解放军空军军医大学军事预防医学系军队卫生统计学教研室(710032)
陈垂雄　王　陵　王文文　黄曼丽　夏结来△　李　晨△

【提　要】 **目的**　评估希望区域法应用于适应性设计样本量再估计的准确性及稳健性,为该方法的适用条件提供理论参考。**方法**　以二分类资料为例,以包含期中分析的两阶段适应性试验为框架,采用 Monte Carlo 方法,在相同的组间响应率率差、同合并率等参数的模拟场景下,比较固定样本量设计、成组序贯设计和希望区域法进行样本量再估计的准确性和稳健性。**结果**　模拟研究显示希望区域法在期中检验效能落入希望区域内进行适应性样本量调整时,Ⅰ类错误与同场景下固定设计和成组序贯设计近似。初始估计样本量被低估时,希望区域法比固定设计的检验效能平均提高约 5%,比成组序贯设计提高约 8.8%,样本量比固定设计多耗费0.18倍,比成组序贯设计多 0.38 倍。高估初始估计样本量时,三种设计的检验效能相差仅在 1%左右,但希望区域法比固定设计多消耗 7.5%、比成组序贯设计多耗费 48%的样本量。**结论**　相比于固定样本量及成组序贯设计,希望区域法对于试验的整体检验效能提高不大,其平均样本量均大于相同设定场景下的固定设计及成组序贯设计,使用该设计方法时应谨慎权衡试验收益。

【关键词】 希望区域法　样本量再估计　条件检验效能　Ⅰ类错误　适应性设计
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.03.001

The Accuracy and Robustness of the Promising Zone in Adaptive Clinical Trial Designs

Chen Chuixiong, Wang Ling, Wang Wenwen, et al (*Department of Military Health Statistics, Department of Military Prevention, Air Force Military Medical University*(710032), *Xi'an*)

【**Abstract**】 **Objectives**　This study aims to evaluate the accuracy and robustness of the promising zone (PZ) for sample size re-estimation (SSR) in adaptive clinical trials, providing theoretical reference for the applicability conditions of the method. **Methods**　Using binary data as an example, within the framework of a two-stage adaptive trial with interim analysis (IA), Monte Carlo simulation was used to compare the accuracy and robustness of the fixed sample size design (Fixed), the group sequential design (GSD), and PZ for SSR, under the same simulated scenarios of between-group response rate difference and merger rate. **Results**　Simulation studies demonstrated that when the IA result was promising for SSR, the Type I error rate of PZ for SSR was comparable to Fixed and GSD. When the initial estimated sample size was underestimated, the statistical power of SSR was on average approximately 5% higher than that of Fixed designs and approximately 8.8% higher than that of GSD, with a average sample size increase of 0.18 times that of Fixed designs and 0.38 times that of GSD. When the initial estimated sample size was overestimated, the difference in power among the three designs was only about 1%, but SSR consumed 7.5% more samples than Fixed designs and 48% more samples than GSD. **Conclusion**　Compared to Fixed and GSD, PZ is only suitable for scenarios where the initial estimated sample size is underestimated. In scenarios where the initial estimated sample size is overestimated, the PZ does not significantly improve the overall power of the trial. Furthermore, the average sample size of the PZ is higher than that of Fixed and GSD under the same settings. Careful consideration of the trade-off between the benefits and costs of using the PZ in clinical trials.

【**Key words**】 Promising zone; Sample size re-estimation; Conditional power; Type I error rate; Adaptive design

临床试验中,样本量的确定依赖于预估的治疗效应,如果实际疗效远低于预期,样本量将被严重低估,从而导致试验检验效能过低而失败[1]。因此,研究者们提出根据期中疗效估计进行样本量适应性调整的策略[2-4],其中希望区域法(promising zone, PZ)是基于期中分析条件检验效能(conditional power, CP)调整样本量的非盲态样本量再估计方法(unblinded sample size re-estimation, Ub-SSR),因其易于理解便于实施而被广泛关注[5]。希望区域法作为一类适应性设计方法,主要包括 Mehta 和 Pocock 等[5-6]提出的 M&P 设计及 CPZ 设计、Jennison 和 Turnbull 等[7-8]提出的基于期望效用最大化的 J&T 设计、Mehta 和 Liu 等[9]的 Baysian 希望区域法。国内外在应用希望区域法时,以 M&P 设计应用最为广泛,目前,国外已有学者采用 Monte Carlo 模拟方法对该方法的试验效率进行评价[6,10-12],但对于Ⅰ类错误的膨胀问题鲜有涉及。本研究拟通过模拟研究观察希望区域法的 M&P 设计在适应性设计样本量再估计的Ⅰ类错误率、检验效能等操作特征(operating characteristics, OC),以评价该方法应用于适应性设计样本量再估计的准确性、稳健性及其适用条件。

*基金项目:国家自然科学基金面上项目(82273728;82273729;82373680)
△通信作者:李晨,E-mail:lc.biosta@qq.com,夏结来,E-mail:jielaixia@yahoo.com

原理与方法

1.条件检验效能

条件检验效能是假设组间实际存在差异时，基于当前数据预测试验结束时推断出这种差异的概率[13]。假设观察到期中检验统计量 z_1 等于第一阶段的检验统计量 Z_1，最终分析时拒绝 H_0 的条件概率为

$$CP_{\delta}(z_1,\tilde{n}_2)=P_{\delta}(Z_2\geqslant z_{\alpha}\mid z_1) \quad (1)$$

其中，$\tilde{n}_2$ 为适应性调整的样本量增量，z_1 为期中检验统计量，Z_2 为最终检验统计量。由于治疗效应 δ 的真值是未知的，假设期中疗效估计为真实的疗效，以期中疗效估计 $\hat{\delta}_1$ 代入式(1)，可得期中分析时的条件检验效能如下：

$$CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)=1-\Phi\left(\frac{z_{\alpha}\sqrt{n_2}-z_1\sqrt{n_1}}{\sqrt{\tilde{n}_2}}-\frac{z_1\sqrt{\tilde{n}_2}}{\sqrt{n_1}}\right) \quad (2)$$

其中，$z_1,n_1,\tilde{n}_2,n_2$ 分别为期中检验统计量，第一阶段样本量，第二阶段样本量的增量，总样本量。

2.希望区域法

希望区域法是基于期中疗效估计进行非盲态样本量再估计的一种适应性试验设计方法，其主要原理是在试验设计阶段根据预先设定的样本量增大倍数、目标检验效能、信息时间定义一个以条件检验效能为尺度的希望区域，当期中分析条件检验效能在希望区域范围内时进行适应性样本量调整，使其能增加到目标检验效能，当期中的条件检验效能落入希望区域外时，不对样本量进行调整，按原定样本量进行至试验结束，最终检验分析采用常规的统计检验，Ⅰ类错误不发生膨胀[14]。依据不同的期中条件检验效能 $CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)$ 落入的区域，有如下样本量适应性调整策略(如图1所示)：

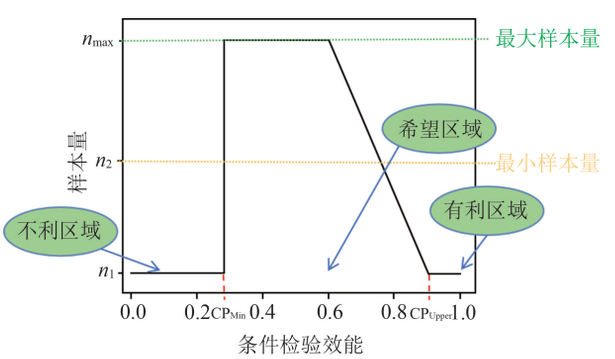

图1 希望区域法适应性调整策略示意图

不利区域：$CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)<CP_{Min}$，表示期中疗效估计不理想，不值得增加过多的样本量以支持最终分析足够的检验效能。试验以原样本量 n_2 继续进行；

有希望区域：$CP_{Upper}>CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)\geqslant CP_{Min}$，期中疗效估计尚可，但未达到或超过试验预先设定的检验效能。试验总的样本量适应性调整为 n_2^*；

有利区域：$CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)\geqslant CP_{Upper}$，期中估计疗效达到预期，在当前数据趋势下，到最终分析时有很大的概率得出药物有效的结果。试验以原样本量 n_2 继续进行。

3.希望区域的构建

假设期中检验统计量 $Z_1=z_1$，样本增量由 $\tilde{n}_2$ 调整到 $\tilde{n}_2^*$，$\tilde{n}_2^*$ 的计算公式如下：

$$P_0(Z_2^*>b(z_1,\tilde{n}_2^*))=\alpha \quad (3)$$

其中，Z_2^* 是最终检验的检验统计量。b 是由 K.K.Gordon Lan 和 Janet Wittes 基于布朗运动的特性提出的一种方法，其基本思想是：通过 Z 统计量的函数构造服从布朗运动的统计量 B，并根据 Z 统计量的分布特性及两者关系，推导出 B 值的分布[15]。根据 Gao 等人[16]的推导，得到如下式子：

$$b(z_1,\tilde{n}_2^*)=(n_2^*)-0.5\left[\sqrt{\frac{\tilde{n}_2^*}{\tilde{n}_2}}(z_{\alpha}\sqrt{n_2}-z_1\sqrt{n_1})+z_1\sqrt{n_1}\right] \quad (4)$$

如果样本量没有调整，则 $\tilde{n}_2^*=\tilde{n}_2$，$Z_2^*=Z_2$，$\hat{b}(z_1,\tilde{n}_2^*)=z_{\alpha}$，说明在期中分析时改变样本量，最终检验使用常规的检验时，只需要将检验界值 z_{α} 替换为 $b(z_1,\tilde{n}_2^*)$。当使用 $Z_2^*\geqslant z_{\alpha}$ 来拒绝 H_0 时，则使Ⅰ类错误不发生膨胀的希望区域定义如下：

$$P=\{CP_{\hat{\delta}_1}(z_1,\tilde{n}_2):b(z_1,\tilde{n}_2^*(z_1))<z_{\alpha}\} \quad (5)$$

其中 $\tilde{n}_2^*(z_1)$ 由 z_1 计算得到，然而希望区域的设定是在期中揭盲前确定的，此处定义样本量只有在期中观测值 $Z_1=z_1$ 使得 $CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)\in P$ 时才进行适应性调整，否则样本量维持原来的设定。则有：

$$\alpha=P_0(Z_2^*)\geqslant b(z_1,\tilde{n}_2^*(z_1))\geqslant P_0(Z_2^*\geqslant z_{\alpha}) \quad (6)$$

由上述，定义希望区域的范围为 $CP_{Min}\leqslant CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)<CP_{Upper}$，$CP_{Upper}$ 根据可接受的检验效能选取，一般取 $1-\beta$。CP_{Min} 为希望区域的下界，其取值由试验设计时预设的样本量增大上限、信息时间、目标检验效能确定，具体计算过程如下：

预先设定期中条件检验效能 $CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)\in(0,1)$ 由等式(2)得到该范围内的 $CP_{\hat{\delta}_1}$ 值对应的期中检验统计量 z_1，由等式(8)~(10)得到适应性调整后的总的样本量 n_2^*，将得到的 z_1,n_2^* 代入公式(4)，可得到 $b(z_1,\tilde{n}_2^*)$，最终得到每一个 $CP_{\hat{\delta}_1}$ 值对应的 b 值 $CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)$ *versus* $b(z_1,\tilde{n}_2^*(z_1))$，则在 $b(z_1,\tilde{n}_2^*(z_1))\leqslant z_{\alpha}$ 范围内最小的 $CP_{\hat{\delta}_1}(z_1,\tilde{n}_2)$，即为希望区域的下界 CP_{Min}。同时 $CP_{\hat{\delta}_1},z_1,\frac{\hat{\delta}_1}{\delta_1}$ 可用式(7)相互转换，以表示不同尺度的希望区域：

$$\frac{\hat{\delta}_1}{\delta_1}=\left[\frac{z_1}{z_{\alpha}+z_{\beta}}\right]\sqrt{\frac{n_2}{n_1}} \quad (7)$$

4.样本量再估计

当期中条件检验效能在希望区域范围内时，样本量增加 $\tilde{n}_2^*$，使试验的检验效能达到目标检验效能，即样本量适应性调整后，满足下式：

$$CP_{\hat{\delta}_1}(z_1,\tilde{n}_2^*)=1-\beta \quad (8)$$

其中 $\hat{\delta}_1$ 是期中分析时观察到的治疗效应，z_1 是期中检验统计量。由式(1)和式(3)，得到达到目标条件检验效能的样本量的增量为：

$$\tilde{n}_2^*(z_1)=\left[\frac{n_1}{z_1^2}\right]\left[\frac{z_\alpha\sqrt{n_2}-z_1\sqrt{n_1}}{\sqrt{n_2-n_1}}+z_\beta\right]^2 \quad (9)$$

其中 $z_u=\Phi^{-1}(1-u)$ 为正态分布累积分布函数。调整后的样本量为变量 $n_2^*=n_1+\tilde{n}_2^*$。实际临床试验中，为避免由样本量的增量逆推期中估计疗效，造成意外破盲，可规定样本量增加到指定的倍数或阶梯倍数[17]。同时，考虑到受试者资源、研究经费等因素，样本量适应性调整的增大上限可取原样本量的 1.5 或 2 倍等，保守起见调整后的样本量不低于初始估计值，则适应性调整后的样本量一般可确定为：

$$n_2^*(z_1)=\min(\max(n_2^*,n_2),n_{max}) \quad (10)$$

模拟研究

1.模拟试验设计

本文以信息时间为 0.5 的两阶段双臂临床试验为例，在二分类资料中进行固定样本量设计(Fixed)、希望区域法(PZ)及成组序贯设计(group sequential design，GSD)的比较研究。设定 PZ 法的期中分析仅判断是否需要进行样本量的适应性调整，适应性调整样本量上限为初始估计样本量的两倍。GSD 设计仅考虑因有效而提前终止，使用经典 O′Brien-Fleming 设计确定初始样本量及期中分析的检验界值。采用 SAS 9.4 随机产生模拟研究数据进行模拟，模拟次数为单场景 10000 次。模拟均在大样本下进行，使用 Z 检验。模拟假定某双臂试验，试验组与对照组随机化配比为 1∶1，试验组与对照组的总体反应率分别为 π_t，π_c，分别服从 $B(N,\pi_t)$ 和 $B(N,\pi_c)$ 的二项分布，取单侧 $\alpha=0.025$ 的检验假设为 H_0：$\pi_t\leqslant\pi_c$ VS H_1：$\pi_t>\pi_c$，组间率差的估计值为 δ_0，则每组需要的样本量为：

$$N_0=\frac{2(z_{1-\alpha/2}+z_{1-\beta})^2\bar{\pi}(1-\bar{\pi})}{(\pi_t-\pi_c)^2} \quad (11)$$

其中，两样本的平均率 $\bar{\pi}=\frac{\pi_t+\pi_c}{2}$，$N_0$ 为单组样本量。

2.模拟场景设置

模拟一观察不同试验设计的Ⅰ类错误、检验效能及平均样本量。模拟二观察不同期中决策区域下不同试验设计的Ⅰ类错误、检验效能及平均样本量。在观察Ⅰ类错误的场景中，设置真实治疗效应 $\delta=0$；在观察检验效能的场景中，设置真实治疗效应 $\delta=0.15$、0.30、0.45，组间响应率率差估计值 δ_0 分别为 δ 的 1.2 倍及 0.8 倍。

3.模拟试验流程

模拟试验流程如图 2 所示，招募第一阶段受试者，当入组受试者数量达到预设的信息时间时进行第一次期中分析，根据期中决策，模拟下一阶段试验，在最终阶段进行假设检验，评价指标如下：

(1)经验Ⅰ类错误(empirical alpha)

$$\alpha_{em}=\frac{\sum_{l=1}^{N_{Sim}}I_l(P\leqslant\alpha)}{N_{Sim}}$$

其中 I_i 为指示函数(如果 $P\leqslant\alpha$，则取值为 1，反之 $P>\alpha$，则取值为 0)，α 为预先指定的Ⅰ类错误，N_{Sim} 为模拟次数。

(2)检验效能(the ture power)

$$p_{\text{power}}=\frac{\sum_{l=1}^{N_{Sim}}I_l(P\leqslant\alpha)}{N_{Sim}}$$

(3)平均样本量(average sample size)

$$N_{ASS}=\frac{\sum_{i=1}^{N_{Sim}}N_i}{N_{Sim}}$$

其中 N_i 为单次模拟时的最终样本量，N_{Sim} 为模拟次数。

4.模拟参数设置

模拟参数设置见表 1。

表 1　模拟参数设置

参数	具体意义	取值范围
π_C	对照组响应率	0.1、0.3、0.7
π_T	试验组响应率	模拟Ⅰ类错误时取 $\pi_T=\pi_C$； 模拟检验效能时 $\pi_T=\pi_C+\delta$
δ	两组间的实际响应率率差	模拟Ⅰ类错误时取 $\delta=0$； 模拟检验效能时取 $\delta=0.15$、0.30、0.45
δ_0	两组间估计的响应率率差	模拟Ⅰ类错误时取 $\delta_0=0.15$、0.3、0.45； 模拟检验效能时取 $\delta_0=R.\delta(R=0.8、1.2)$
t	信息时间	1/2
sim	模拟次数	10000
α	总Ⅰ类错误(单侧检验)	0.025
$1-\beta$	目标检验效能(power)	0.9

结　果

1.模拟一

表 2、图 3 展示了不同试验设计的Ⅰ类错误和平均样本量的变化情况。当真实治疗效应 $\delta=0$ 时，在不同参数场景下三种试验设计的Ⅰ类错误在 0.025 波动，PZ 法的Ⅰ类错误均比同场景下的 Fixed 设计要低，与 GSD 设计的Ⅰ类错误近似，相差仅在 0.2%左右，与 GSD 设计近似，比 Fixed 设计的样本量平均多消耗 7%，比 GSD 设计多消耗 13%。在 $\pi_{c0}=0.1$ 时，随

着 δ_0 的增加，三种设计在Ⅰ类错误的变化上与整体表现一致，PZ 法平均样本量与 GSD 设计相比，增加幅度随着 δ_0 的增加不断变大，这说明当试验样本量的初始估计值越小时，PZ 法适应性调整样本量的变化幅度越大，三种设计的Ⅰ类错误及平均样本量在 $\pi_{c0}=0.3$、$\pi_{c0}=0.7$ 内与在 $\pi_{c0}=0.1$ 内表现一致。

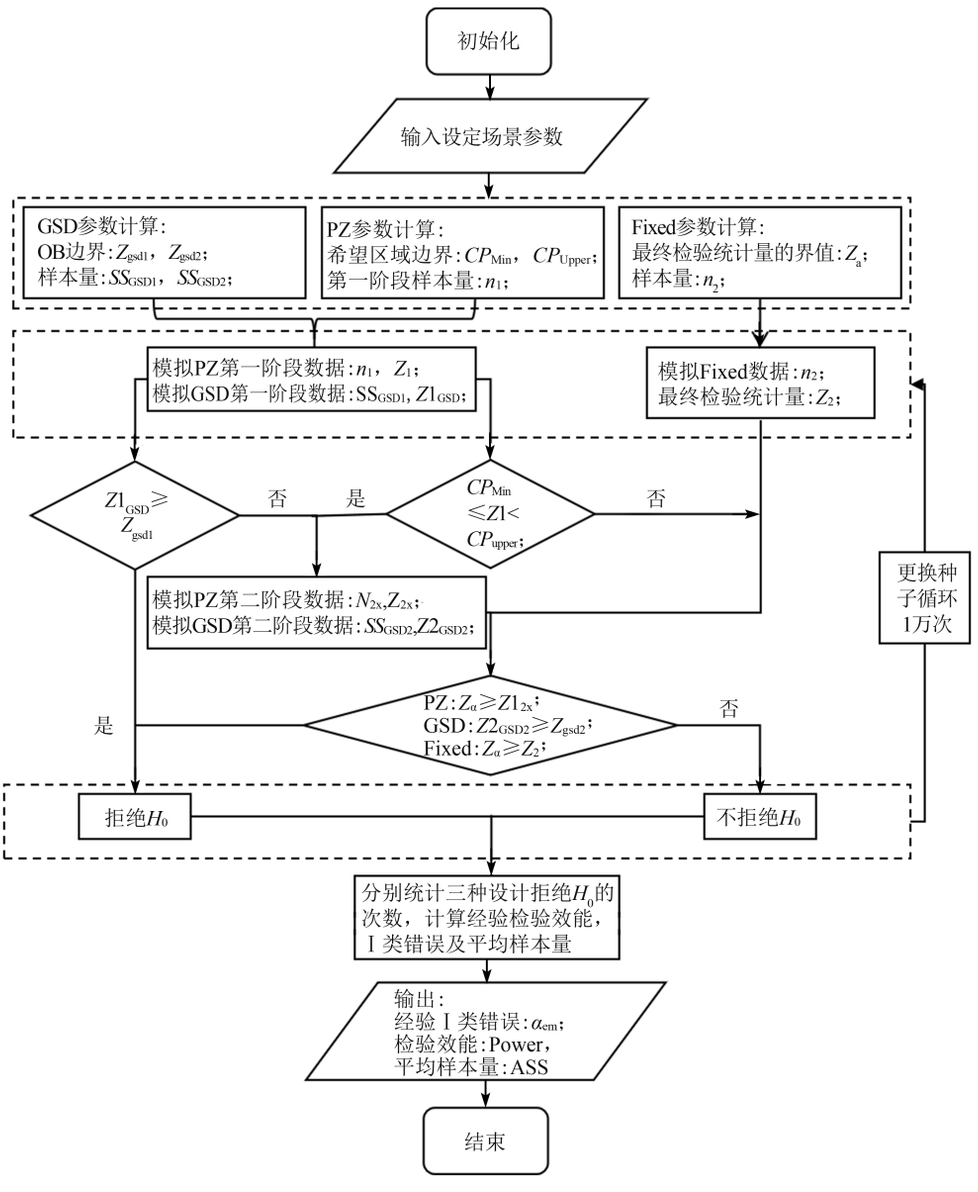

图 2 模拟试验流程图

表 2 不同试验设计的Ⅰ类错误及平均样本量

π_{c0}	δ_0	n_1	n_2	ASS_{GSD}	ASS_{PZ}	α_1	α_{GSD1}	α_{Fixed}	α_{GSD}	α_{PZ}
	0.15	134	266	261.71	286.02	0.0260	0.0022	0.0256	0.0238	0.0249
0.1	0.30	42	84	77.98	90.85	0.0220	0.0004	0.0267	0.0278	0.0237
	0.45	22	42	36.00	45.24	0.0057	0.0000	0.0223	0.0188	0.0193
	0.15	218	434	431.48	462.40	0.0254	0.0024	0.0266	0.0257	0.0245
0.3	0.30	56	112	105.84	119.32	0.0249	0.0030	0.0258	0.0263	0.0257
	0.45	24	48	41.96	52.15	0.0253	0.0019	0.0276	0.0252	0.0263
	0.10	392	784	783.22	833.85	0.0263	0.0020	0.0261	0.0257	0.0243
0.7	0.15	162	322	317.62	343.82	0.0262	0.0024	0.0229	0.0245	0.0228
	0.20	82	164	159.78	174.68	0.0249	0.0028	0.0218	0.0224	0.0200

注：π_{c0}为对照组初始估计响应率；δ_0为初始估计治疗效应；n_1为第一阶段样本量；n_2为初始估计样本量；ASS_{GSD}：成组序贯设计的平均样本量；ASS_{PZ}为希望区域法再估计后的平均样本量；α_1为第一阶段所犯的Ⅰ类错误；α_{GSD1}为成组序贯设计第一阶段的Ⅰ类错误；α_{Fixed}为固定样本量设计的Ⅰ类错误；α_{GSD}为成组序贯设计的Ⅰ类错误；α_{PZ}为希望区域法的Ⅰ类错误。

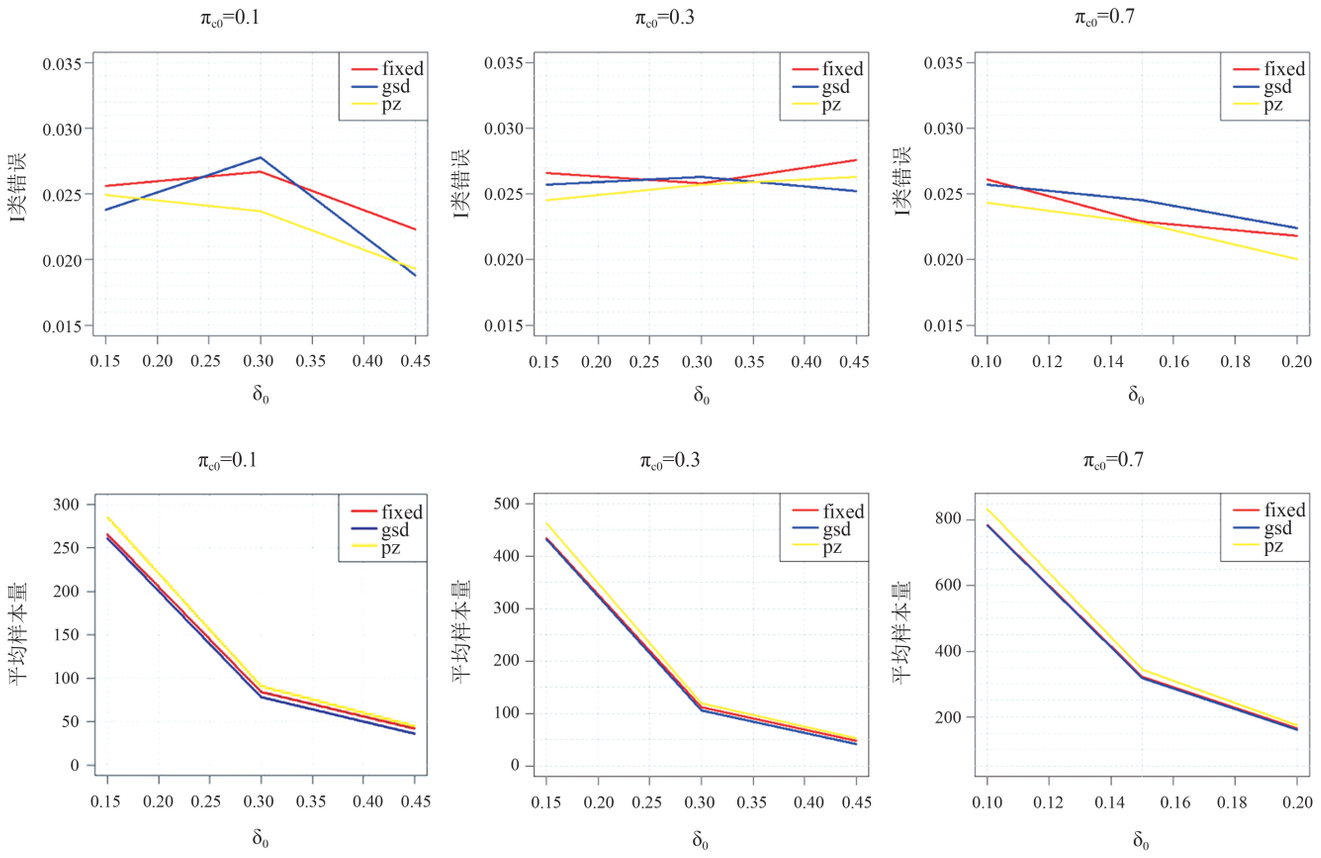

图 3　不同试验设计的 I 类错误及平均样本量

表 3、图 4 展示了 $\delta_0 = 1.2\delta$ 时不同试验设计的检验效能及平均样本量，初始估计样本量被低估时，PZ 法比 Fixed 设计的检验效能平均提高 5%，样本量多耗费 0.18 倍，比 GSD 设计的检验效能平均提高 8.8%左右，多耗费约 0.38 倍的平均样本量；在 $\pi_{c0} = 0.1$ 时，随着 δ_0的增加，PZ 法比 Fixed 设计检验效能和平均样本量的增加量分别在 4%和 0.17 倍左右，相对于 GSD 设计，PZ 法平均效能的增加量从 5%～15%不断增大，平均样本量增加量的幅度亦随之不断加大，为 0.32～0.53倍，在 π_{c0}分别为 0.3、0.7 时，与 0.1 时的变化趋势类似。同时，此场景的模拟中，PZ 法的平均检验效能最高为 85%，达不到目标效能 90%。

表 3　$\delta_0 = 1.2\delta$ 时不同试验设计的检验效能及平均样本量

π_{c0}	δ	δ_0	n_1	n_2	ASS_{GSD}	ASS_{PZ}	AP_1	AP_{GSD1}	AP_{Fixed}	AP_{GSD}	AP_{PZ}
0.1	0.15	0.18	98	196	174.22	230.90	0.5196	0.1852	0.8118	0.7989	0.8572
	0.30	0.36	32	62	52.21	72.64	0.5570	0.1355	0.8155	0.7678	0.8580
	0.45	0.54	16	30	22.82	34.94	0.5317	0.0983	0.8051	0.6918	0.8432
0.3	0.15	0.18	154	306	272.57	358.04	0.5007	0.1962	0.7830	0.7752	0.8263
	0.30	0.36	40	78	66.59	93.99	0.4644	0.1504	0.7829	0.7260	0.8370
	0.45	0.54	16	32	25.16	37.58	0.5342	0.0699	0.7609	0.6811	0.8003
0.7	0.10	0.12	266	530	479.29	626.67	0.4680	0.1783	0.7633	0.7564	0.8146
	0.15	0.18	106	212	192.63	250.38	0.4591	0.1478	0.7457	0.7330	0.7981
	0.20	0.24	52	104	90.88	123.96	0.4748	0.1483	0.7521	0.7221	0.8059

注：π_{c0}为对照组初始估计响应率；δ 为真实治疗效应；δ_0为初始估计治疗效应；n_1为第一阶段样本量；n_2为初始估计样本量；ASS_{GSD}：成组序贯设计的平均样本量；ASS_{PZ}为希望区间法再估计后的平均样本量；AP_1 为希望区域法第一阶段的检验效能；AP_{GSD1}为成组序贯设计第一阶段的检验效能；AP_{Fixed}为固定样本量设计的检验效能；AP_{GSD}为成组序贯设计的检验效能；AP_{PZ}为希望区域法的检验效能。

表 4、图 5 展示了 $\delta_0 = 0.8\delta$ 时不同试验设计的检验效能及平均样本量，当初始估计样本量被高估时，整体来看，PZ 法的平均检验效能均高于 Fixed 设计和 GSD 设计，但三者差别不大，约在 0.1%左右，PZ 法比 Fixed 设计平均多消耗 7.5%的样本量，比 GSD 设计多耗费 0.48 倍的样本量。在 $\pi_{c0} = 0.1$ 时，随着 δ_0 的增加，相对于 Fixed 设计，检验效能由 1%提高到 1.7%，平均样本量多花费 8%左右，与 GSD 设计对比，提高了 0.9%～1.8%的检验效能，但多消耗 0.45～0.5 倍的样本量，$\pi_{c0} = 0.3$ 及 0.7 时表现类似。这说明此

场景下，PZ 法检验效能的提高耗费的样本量较多，对比 Fixed 设计，提高 1% 的效能约需要多花费 8% 的样本量，与 GSD 设计相比，提高 1% 多消耗 0.45 倍的样本量。

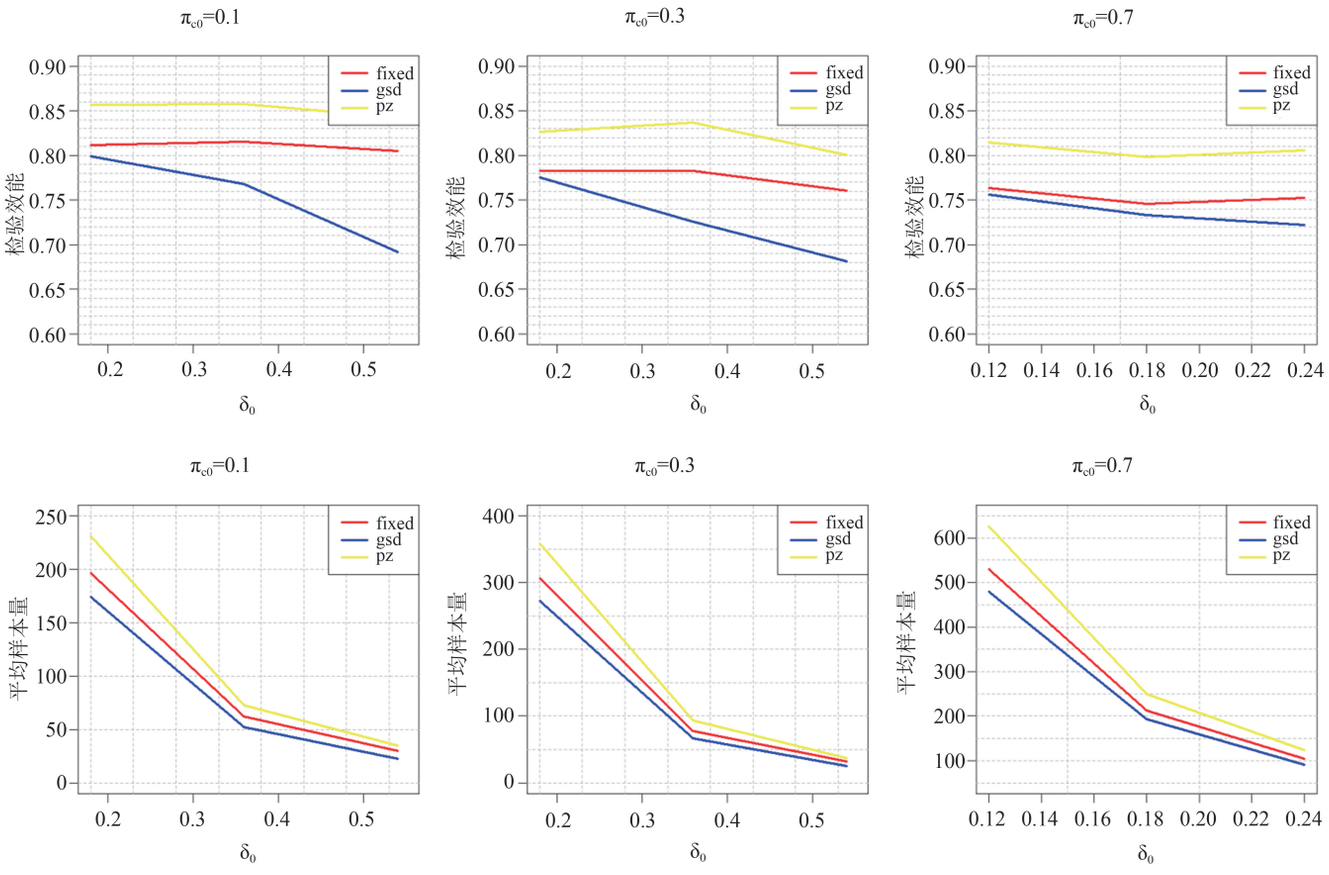

图 4 $\delta_0 = 1.2\delta$ 时不同试验设计的检验效能及平均样本量

表 4 $\delta_0 = 0.8\delta$ 时不同试验设计的检验效能及平均样本量

π_{c0}	δ	δ_0	n_1	n_2	ASS_{GSD}	ASS_{PZ}	AP_1	AP_{GSD1}	AP_{Fixed}	AP_{GSD}	AP_{PZ}
0.1	0.15	0.12	196	390	291.44	422.36	0.4925	0.8047	0.9784	0.9759	0.9856
	0.30	0.24	62	122	90.09	131.87	0.4468	0.8121	0.9817	0.9742	0.9869
	0.45	0.36	32	62	44.31	66.64	0.4175	0.8394	0.9843	0.9713	0.9890
0.3	0.15	0.12	336	670	496.36	725.55	0.5139	0.8108	0.9795	0.9784	0.9858
	0.30	0.24	88	174	124.16	187.59	0.5457	0.8084	0.9835	0.9800	0.9877
	0.45	0.36	40	78	55.21	84.07	0.4663	0.8253	0.9866	0.9761	0.9927
0.7	0.10	0.08	630	1260	915.82	1349.89	0.5495	0.8262	0.9837	0.9833	0.9891
	0.15	0.12	266	530	382.37	564.68	0.5482	0.8425	0.9861	0.9860	0.9913
	0.20	0.16	140	278	197	294.60	0.5662	0.8599	0.9909	0.9895	0.9933

注：π_{c0} 为对照组初始估计响应率；δ 为真实治疗效应；δ_0 为初始估计治疗效应；n_1 为第一阶段样本量；n_2 为初始估计样本量；ASS_{GSD}：成组序贯设计的平均样本量；ASS_{PZ} 为希望区域法再估计后的平均样本量；AP_1 为希望区域法第一阶段的检验效能；AP_{GSD1} 为成组序贯设计第一阶段的检验效能；AP_{Fixed} 为固定样本量设计的检验效能；AP_{GSD} 为成组序贯设计的检验效能；AP_{PZ} 为希望区域法的检验效能。

2.模拟二

图 6 展示了不同期中区域下不同设计的期中决策概率、Ⅰ类错误及平均样本量，不同初始治疗效应估计值的期中决策概率相近，有利区域、希望区域和不利区域分别在 0.02、0.1、0.88 波动。在有利区域，三种设计的Ⅰ类错误均发生膨胀，PZ 法的Ⅰ类错误比 GSD 设计少膨胀约 0.02～0.12，样本量平均比 GSD 设计多花费 4%～18%；在希望区域部分，三种设计的Ⅰ类错误均发生膨胀，PZ 法的Ⅰ类错误比 GSD 设计少膨胀约 0～0.04，比 Fixed 设计少 0.001～0.026，比 GSD 设计的平均样本量多消耗 0.72～1 倍，比 Fixed 设计多花费约 0.71～0.8 倍样本量；在不利区域，三种设计的Ⅰ类错误均低于 0.008，GSD 设计的Ⅰ类错误比 PZ 法多 0.0006～0.0035，样本量平均比 PZ 法少花费 0.4%～14%，PZ 法与 Fixed 设计的检验效能和平均样本量在不利及有利区域相同。

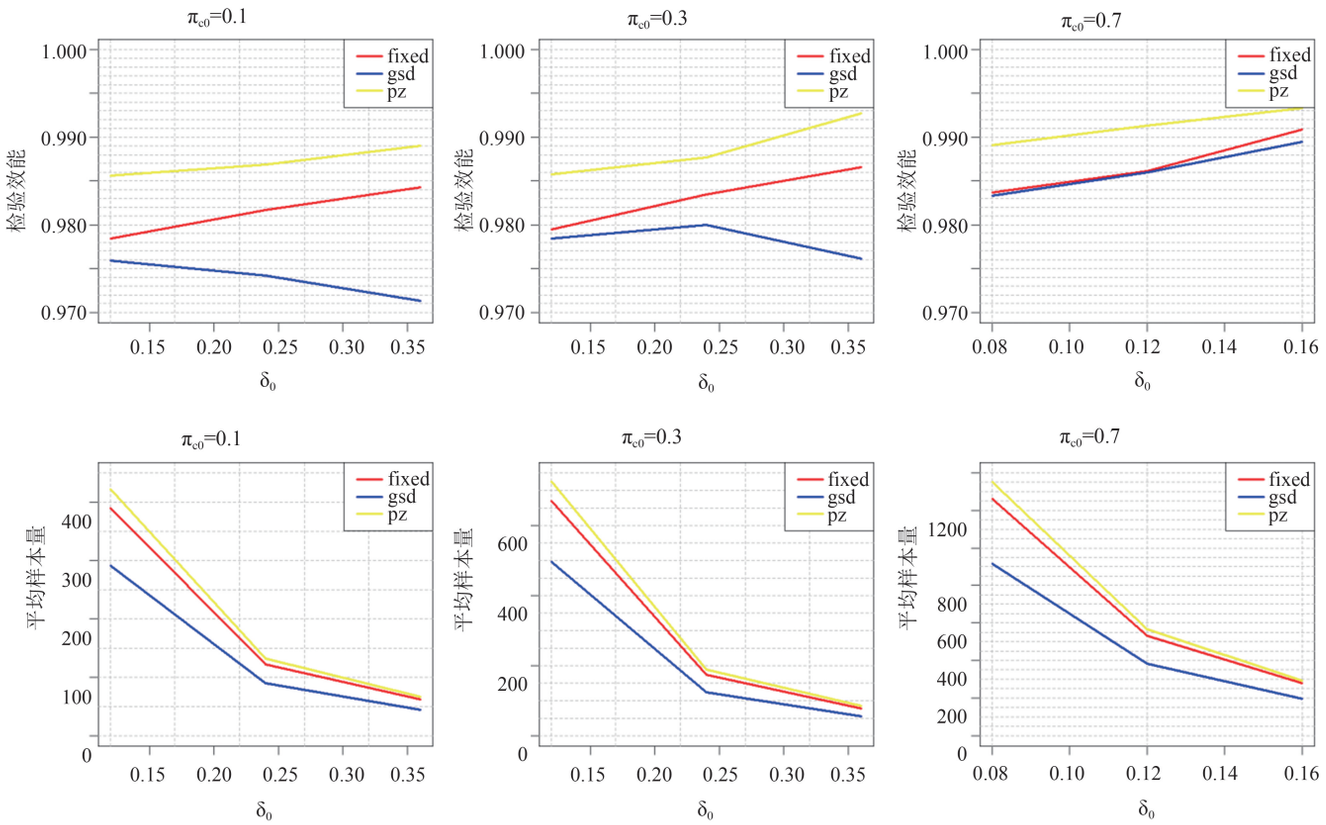

图 5　$\delta_0 = 0.8\delta$ 时不同试验设计的检验效能及平均样本量

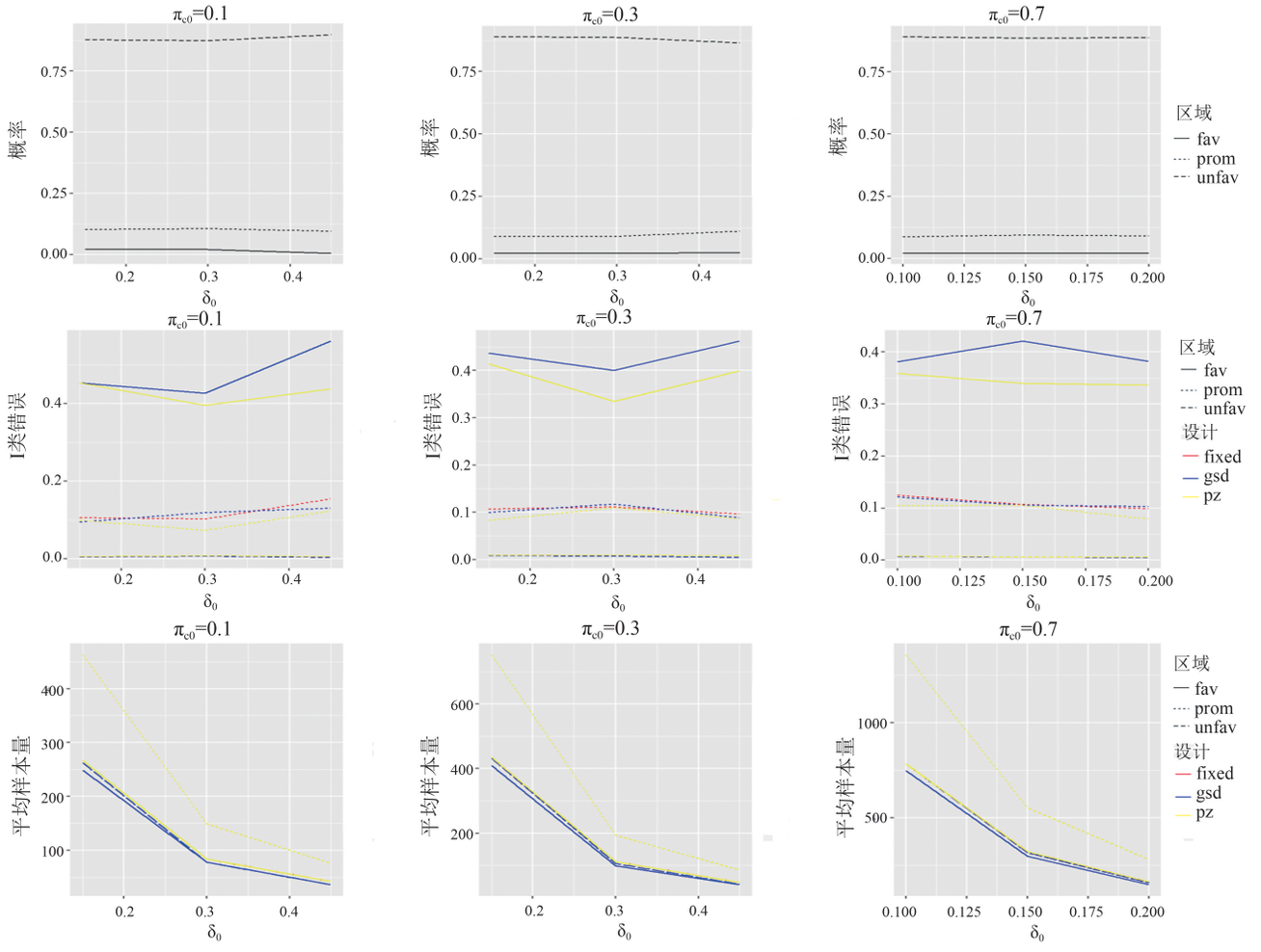

图 6　不同期中区域下不同设计的 I 类错误及平均样本量

图 7 展示了当疗效差异的初始估计值为真实值的 1.2 倍 $\delta_0 = 1.2\delta$ 时，不同期中区域下不同设计的期中决策概率，初始估计样本量被低估时，有利区域、希望区域和不利区域在不同真实疗效的期中结果概率分别围绕 0.45、0.3、0.25 波动。有利区域部分，PZ 法的检验效能最低比 GSD 设计低 0.0033，最高比 GSD 设计高 0.0262，平均样本量比 GSD 设计多消耗 23.2%～32.7%；在希望区域部分，PZ 法的平均样本量比 GSD 设计多消耗 59.7%～100%，检验效能比 GSD 设计提高 15.6%～23.5%，比 Fixed 设计多花费 56.8%～60.5%的平均样本量，检验效能提高约 13.4%～16.9%。在不利区域，PZ 法的检验效能比 GSD 设计高 1.9%～27%，平均样本量比 GSD 设计多耗费 1%～23%；希望区域法与固定设计在不利及有利区域部分，检验效能和平均样本量相同。

图 8 展示了当 $\delta_0 = 0.8\delta$ 时，不同期中区域下不同设计的期中决策概率，初始估计样本量较高时，期中决策概率在 0.80，0.15，0.05 波动。有利区域部分，PZ 法的检验效能与 GSD 设计相差不大，约在 0.1%左右，平均样本量比 GSD 设计多消耗 49.3%～54.5%；在希望区域部分，PZ 法的检验效能比 GSD 设计多花费 2.1%～7.6%，比 Fixed 设计提高 1.9%～4.4%，然而 PZ 法平均比 GSD 设计多耗费 48.6%～63.7%的样本量，比 Fixed 设计多耗费 0.48～0.51 倍的样本量。在不利区域，PZ 法的检验效能比 GSD 设计提高 0.5%～14.22%，平均比 GSD 设计的多消耗 0.1 到 0.11 倍的样本量；不利区域和有利区域部分，PZ 法的检验效能和平均样本量与 Fixed 设计相同。

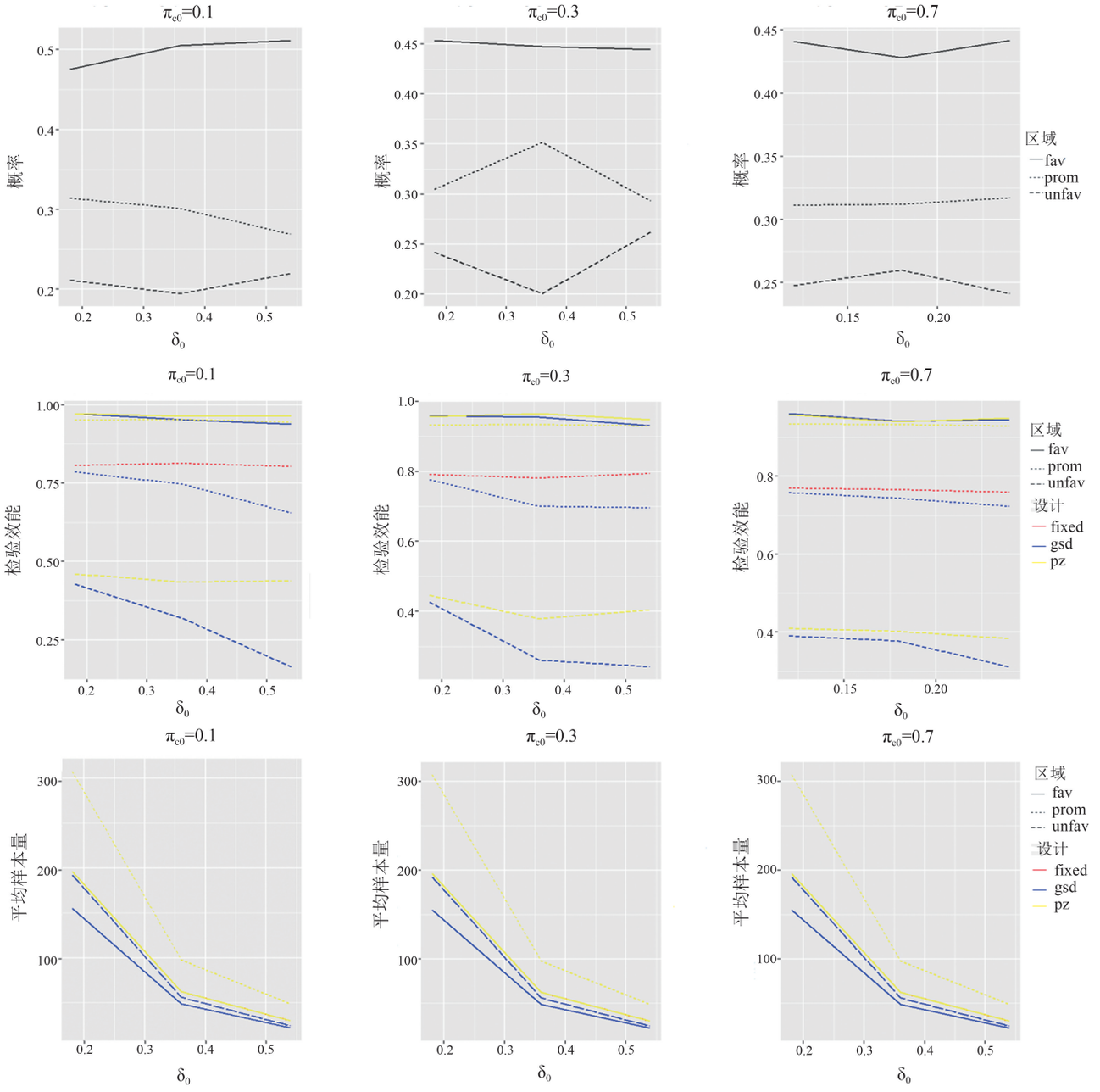

图 7 $\delta_0 = 1.2\delta$ 时不同期中区域下不同设计的检验效能及平均样本量

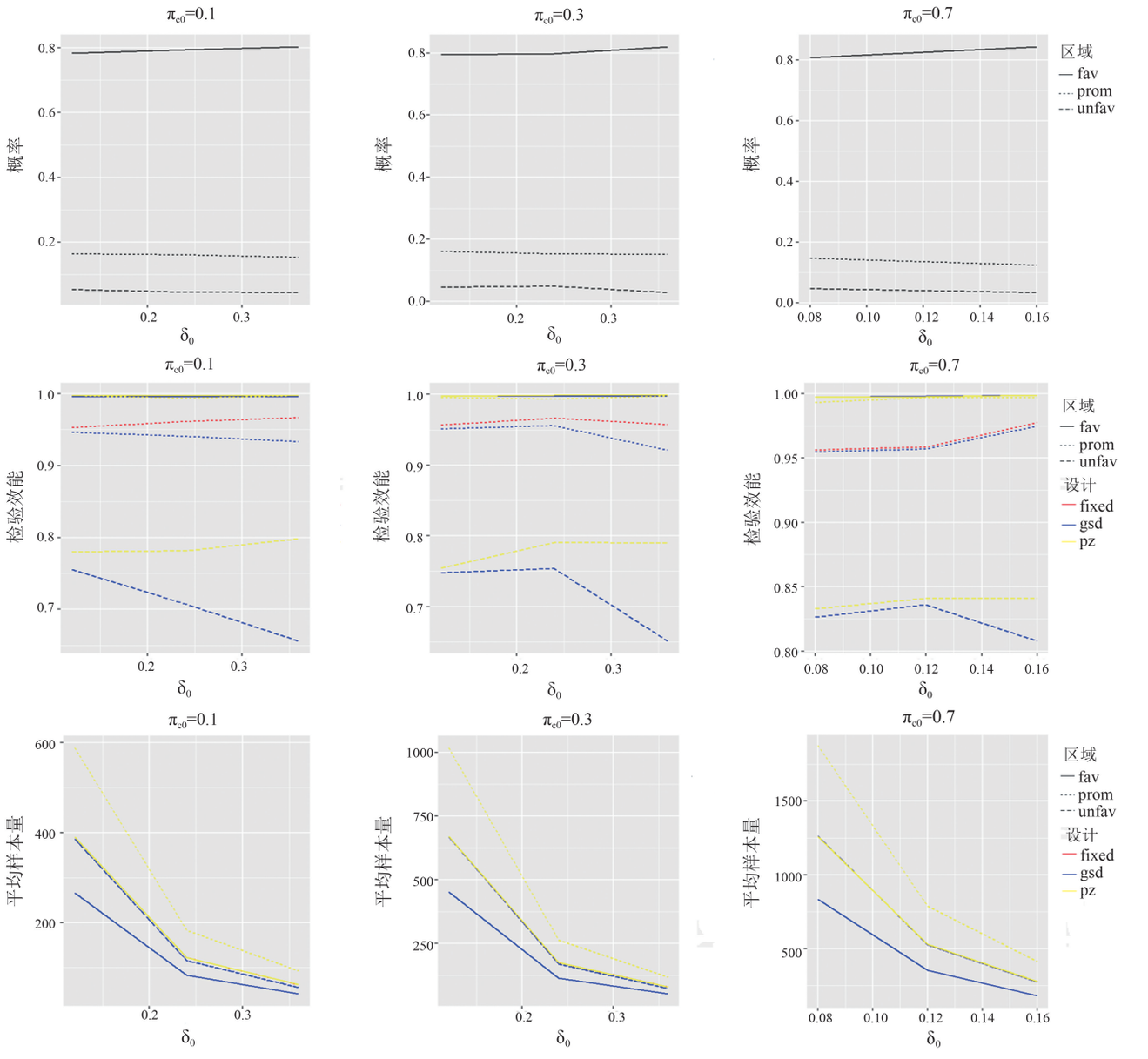

图 8　$\delta_0 = 0.8\delta$ 时不同期中区域下不同设计的检验效能及平均样本量折线图

讨　论

本研究以二分类资料为例，通过 PZ 法、GSD 设计和 Fixed 设计的模拟研究探索了 PZ 法应用于适应性设计试验样本量再估计的准确性及稳健性。结果发现设置了最低希望区域下限的 PZ 法在 Ⅰ 类错误上，与 GSD 设计及 Fixed 设计近似；当低估试验所需样本量时，PZ 法能够提高试验的检验效能，但会比 GSD 设计消耗更多的样本量，同时由于有样本量增加倍数上限的限制，当期中进行适应性样本量调整时，即使增加样本量也可能无法达到目标检验效能。从 PZ 法的期中决策区域上看，PZ 法的检验效能在不利和有利区域与 GSD 设计差别不大，但在有利区域上平均样本量比 GSD 设计平均多 50%。在有希望区域，PZ 法的检验效能比 GSD 设计平均高 20%，多耗费 0.5 倍的样本量；当高估试验所需样本量时，PZ 法与 GSD 设计在检验效能上差别不大，最高提高 0.72% 的检验效能，然而在平均样本量上最低也比 GSD 设计多花费 45%，这说明此种场景下 PZ 法的收益不高。从 PZ 法的期中决策区域上看，在不利区域，PZ 法的检验效能与 GSD 设计差别不大，但平均样本量比 GSD 设计平均多 50%。有希望区域部分，PZ 法带来的检验效能比 GSD 设计平均多 3.3%，平均样本量比成组序贯设计平均多 36%。

综上所述，我们认为 PZ 法能控制 Ⅰ 类错误，其应用于低估试验初始样本量时较好，但其检验效能与付出的样本量相比，整体收益较低。我们认为 PZ 法不适用于高估初始样本量的情况，前期试验药物的疗效较好时，应使用成组序贯试验。若采用 PZ 法，应在试验设计阶段，预先考虑多种场景，进行多次模拟，使试验的实施及决策得到保证。本研究仅针对二分类资料下信息时间为 0.5 的两阶段临床试验进行模拟数据研

（下转第 338 页）