器械中有大一类产品是诊断产品，譬如，试剂盒、超声、核磁共振、X射线等。诊断产品做临床试验主要验证的是诊断准确率，譬如，灵敏度、特异度、阳性预测率、阴性预测率、ROC曲线下面积等指标。如果一个新的产品要通过做临床试验提交药监局批准上市，需要考虑常见的几种诊断试验设计类型。

根据有无金标准，设计类型可以分为以下四种类型：

1. 有金标准的单组设计：

此时主要衡量新诊断产品准确率是否达到一定标准。譬如，通过做临床试验验证乳腺超声的灵敏度是否大于0.85，特异度是否大于0.8。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 金标准 | |  |
| 新诊断产品 |  | 有病 | 无病 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |

灵敏度=

特异度=

可以采用精确概率法或者近似正态Z检验对灵敏度和特异度是否达到目标值进行检验。

1. 有金标准的两组配对设计：

此时，主要衡量新诊断产品是否优于或者非劣于对照产品。譬如，通过做临床试验验证3D乳腺X射线的灵敏度和特异度是否优于2D乳腺射线。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 已经过金标准确诊为疾病的受试者 | | | | |
|  |  | 对照产品 | |  |
| 新诊断产品 |  | 阳性 | 阴性 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |

新诊断产品的灵敏度=

对照产品的灵敏度=

若要比较新诊断产品的灵敏度是否优于对照产品，可以采用McNemar检验（想一想为什么？）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 已经过金标准确诊为健康的受试者 | | | | |
|  |  | 对照产品 | |  |
| 新诊断产品 |  | 阳性 | 阴性 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |

新诊断产品的特异度=

对照产品的特异度=

若要比较新诊断的特异度是否优于对照产品，可以采用McNemar检验：

1. 有金标准的两组平行对照设计：

此时，主要衡量新诊断产品是否优于或者非劣于对照产品。譬如，通过做临床试验验证3D乳腺X射线的灵敏度和特异度是否优于2D乳腺射线。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 金标准 | |  |
| 新诊断产品 |  | 有病 | 无病 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |
|  |  |  | |  |
| 对照产品 | 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |

新诊断产品的灵敏度=

对照产品的灵敏度=

若要比较新诊断产品的灵敏度是否优于对照产品，可以采用卡方检验或者近似正态Z检验。

新诊断产品的特异度=

对照产品的特异度=

若要比较新诊断的特异度是否优于对照产品，可以采用卡方检验或者近似正态Z检验。

1. 无金标准时的两组配对设计：

对比新诊断产品和对照产品用PPA和NPA衡量。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 对照产品 | |  |
| 新诊断产品 |  | 阳性 | 阴性 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |

PPA=

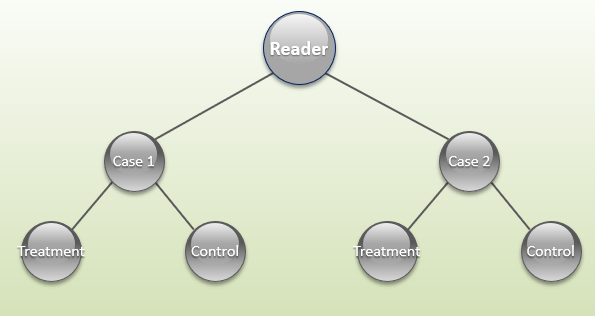
NPA=

可以采用McNemar检验对新诊断和对照产品的阳性比例/阴性比例是否有差异做检验。

影像类诊断器械一般根据是否采用多阅片人，可以分为以下三种类型（以配对设计为例说明，关于平行对照设计可以思考下采用什么分析方法）：

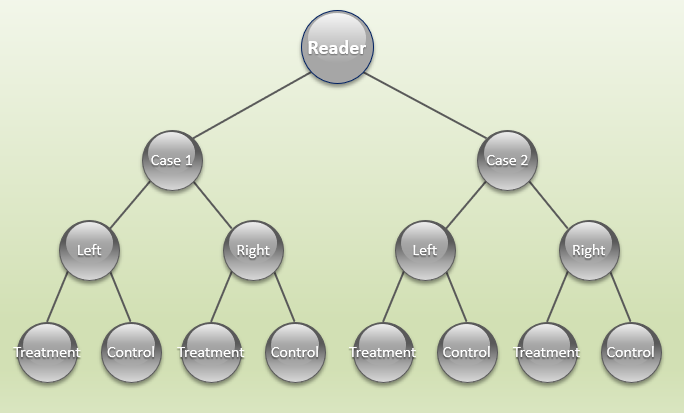
1. Matched-Pair design

不考虑阅片人之间的经验差异，用1个阅片人（实际中，往往采用两个阅片人+一个裁决者阅片；当两个阅片人判断结果不一致时，裁决者来判断结果是阳性还是阴性）采用新诊断产品和对照产品分别判断影像结果。如下图，阅片人对多个受试者分别采用新诊断产品（treatment）和对照产品（control）产生影像结果并进行判断。统计方法采用McNemar 检验或GEE模型。



1. Clustered Matched-Pair design

不考虑阅片人之间的经验差异，用1个阅片人采用新诊断产品和对照产品分别判断影像结果，但是考虑到cluster结构，譬如，乳腺X射线判断乳腺结果时考虑到左右乳房的相关性，此时采用clustered Matched pair设计。统计分析方法采用clustered McNemar 检验或GEE模型。



1. Multiple reader multiple case design考虑阅片人之间的经验差异，用多个（例如，5个阅片人）阅片人采用新诊断产品和对照产品分别判断影像结果。如下图，多个阅片人对多个受试者分别采用新诊断产品（treatment）和对照产品（control）产生影像结果并进行判断。统计方法采用MRMC 混合效应模型或者GEE模型。

根据试验目的选好试验设计类型后，接下来重要的一步是计算样本量，那么对于有金标准的两组配对设计，样本量怎么计算呢？

诊断试验中有金标准的配对设计主要衡量新诊断产品的诊断准确率（灵敏度，特异度、ROC曲线下面积）是否优于或者非劣于对照产品。譬如，通过做临床试验验证3D乳腺X射线的灵敏度和特异度是否优于2D乳腺射线。关于配对设计，需要考虑在产品之间诊断结果的相关性，然后计算样本量；如果不通过相关性来计算样本量，也可以通过四格表中的a, b, c, d来计算样本量。下面介绍下常见的几种样本量计算方法，并且对比这几种方法在相关性从0到1变化时的表现。配对设计中，我们对疾病受试者总数和健康受试者总数分别计算。本文以疾病受试者样本量为例，介绍常见的配对设计样本量计算方法。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 已经过金标准确诊为疾病的受试者 | | | | |
|  |  | 对照产品 | |  |
| 新诊断产品 |  | 阳性 | 阴性 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 已经过金标准确诊为疾病的受试者 | | | | |
|  |  | 对照产品 Y0 | |  |
| 新诊断产品Y1 |  | 阳性 | 阴性 |  |
| 阳性 |  |  |  |
| 阴性 |  |  |  |
|  |  |  |  | 1 |

其中，

新产品的灵敏度=

对照产品的灵敏度=

**新诊断产品和对照产品之间的相关性大小：**

=

关于的具体推导参见（Conner 1987, appendix）。

**Miettinen 样本量公式 (Miettinen 1968)：**

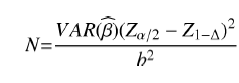
**Conner样本量公式 (Conner 1987)：**

**独立两样本率比较的样本量公式（药监局-医疗器械临床试验设计指导原则）：**

**GEE (logit function)样本量公式(Zhang 2014)：**

**GEE (identity function) 样本量公式 (Liu 1997)：**

**GLMM (Generalized linear mixed model)样本量计算公式(Dang 2008)：**



****

**表1 相关系数对样本量的配对率比较的样本量影响**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 一类错误/二类错误 |  |  |  |  |  | Miettinen | Conner | GEE  (Identity/logit) | 独立两样本 |
| 0.05 / 0.2 | 0.9 | 0.8 | 0.1 | 0 | 0.667 | 56 | 77 | 71 / 75 | 197 |
|  |  |  | 0.11 | 0.01 | 0.583 | 77 | 92 | 87 / 92 | 197 |
|  |  |  | 0.12 | 0.02 | 0.5 | 96 | 108 | 103 / 108 | 197 |
|  |  |  | 0.13 | 0.03 | 0.417 | 114 | 124 | 118 / 125 | 197 |
|  |  |  | 0.14 | 0.04 | 0.333 | 131 | 139 | 134 / 141 | 197 |
|  |  |  | 0.15 | 0.05 | 0.250 | 148 | 155 | 150 / 158 | 197 |
|  |  |  | 0.16 | 0.06 | 0.1667 | 164 | 171 | 165 / 175 | 197 |
|  |  |  | 0.17 | 0.07 | 0.083 | 181 | 186 | 181 / 191 | 197 |
|  |  |  | 0.18 | 0.08 | 0 | 197 | 202 | 197 / 208 | 197 |
|  |  |  | 0.19 | 0.09 | -0.083 | 213 | 218 | 212 / 224 | 197 |
|  |  |  | 0.2 | 0.1 | -0.167 | 229 | 234 | 228 / 241 | 197 |

通过该表可以发现：

* Conner方法算得的样本量大于 Miettinen方法算得的样本量
* Conner方法算得的样本量跟GEE (logit)方法算得的样本量接近。
* 当相关性低于0.5时， GEE (identity)方法算得的样本量跟Miettine方法算得的样本量接近。
* GEE (logit)方法算得的样本量大于GEE (identity)算得的样本量
* 当相关性大于0.1时，独立两样本方法算得的样本量最大。但是，当相关性小于0.1尤其是小于0时，独立两样本方法并不保守，算得的样本量小于其他几种方法。

到底选择哪个样本量公式, 以下提供几点建议：

* 在试验设计阶段，通过模拟比较或者理论方法比较选择合适的分析方法，譬如到底是选择GEE模型、还是GLMM模型还是McNemar检验，可以在设计阶段通过模拟或者理论比较来选择。然后根据选定的方法选择相应的样本量估计方法。譬如，如果选择GEE(identity)分析方法，那么就选择对应的样本量估计方法；如果选择McNemar分析方法，那么选择对应的比较保守的Conner方法。
* 根据以往数据来估计相关性的大小，在相关性大于0的时候，选择保守的独立两样本方法估计样本量也是可以的。

**参考文献：**

1. Miettinen OS. The matched pairs design in the case of all-or-none responses. Biometrics. 1968;24 (2): 339–352.
2. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. Biometrics. 1987; 43 (1): 207–211.
3. Zhang S, Cao J, Ahn C. A GEE approach to determine sample size for pre- and post-intervention experiments with dropout. Computational Statistics and Data Analysis. 2014; 69: 114–121.
4. Liu G, Liang KY. Sample size calculations for studies with correlated observations. Biometrics. 1997; 53: 937–947.
5. Dang TY, Mazumdar S, Houck PR. Sample Size and Power Calculations Based on Generalized Linear Mixed Models with Correlated Binary Outcomes. Computer methods programs biomedicine. 2008; 91(2): 122-127.
6. 医疗器械临床试验设计指导原则

