关于文章提到的数据集合搜集办法：

1. 打开国泰安处罚数据，共5275条数据
2. 筛选promulgator=证监会、上交所、深交所，共2133条数据
3. 筛选declaredate在2001-2008的数据，共651条数据
4. 删去违规类别为：“其他”，“推迟披露”，“擅自改变资金用途”，“违规购买股票”以及这四种的结合，共354条数据
5. 按照股票代码查重，去掉重复数据，共253条

\*如果第三步不使用declaredate而是使用desposaldate，数据数目略有不同，3、4、5步分别剩余数据量为：683条、367条、261条

这个数据集合和我们之前做的数据集合数目差异较大的原因主要在于我们之前的标准和国泰安的标准不一样，具体表现在：

1. 类别问题：国泰安的“推迟披露”中有一部分是关联交易的披露不及时，我们认为被证监会和沪深两市处理的关联交易披露问题算作财务造假（操纵利润等），这里删去了；国泰安的“虚假记载（误导性陈述）”中有相当一部分没有纳入我们之前的数据集合，原因在于误导性陈述中，有很大一部分是业绩快报和审计后报表差异过大，也有一部分是其他因素的误导性陈述，比如非法或者未披露的股份交易等。
2. 我们的数据集合考虑了是否为年报造假、是否为公开发行、非公开发行、重组文件等造假，这一部分占据了不小的权重；除去我们的数据集合外还有几类造假：
3. 季报造假
4. 公司公告造假（主要是相关信息没有及时披露，如借贷关系、关联方交易等）
5. 公开发行、非公开发行、重组文件的造假（这些文件也会有财务报表）