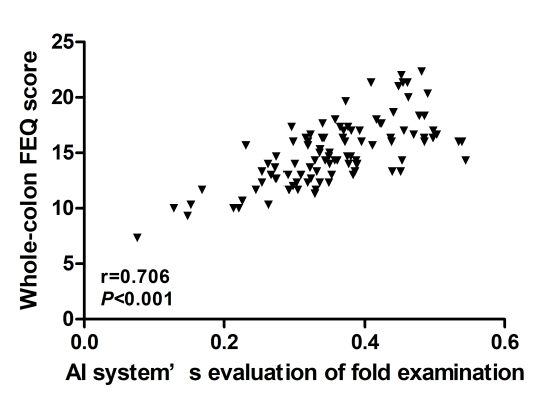
1、研究内容，前期通过视频拆分成图片通过深度学习研发了一个评价消化内镜视频操作质量的AI模型。

2、研究目的，通过AI对视频测试，同时专家组对视频进行人为评价打分，然后做相关性分析评价该模型的效果，

2、研究方法，纳入4个机构11个医生，每个医生操作约10个视频，共103个视频。

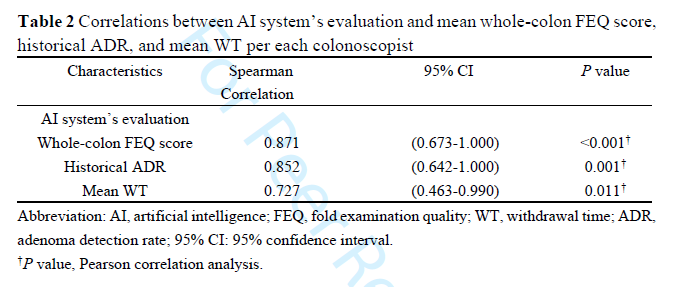
4、首先从每个视频（共103）的角度分析，分析AI测试结果与专家打分做相关性分析（Pearson's Correlation Tests）。结果下图



**Fig. 3** Correlations between AI system’s evaluation of FEQ and Whole-colon FEQ from experts of each video. AI system’s evaluation is significantly associated with whole colon fold examination quality (FEQ) (r=0.706; *P* < 0.001, Pearson correlation analysis) per each video clips (n=103).

5、然后从医生角度再次进行相关性分析，思路如下，以每个医生为单位分析，每个医生操作的视频约10个，通过对这10个视频的AI测试以及专家打分，得到AI和专家打分2组得分，计算AI和专家打分的平均值，作为医生的特征，包括（1）AI评价结果，及（2）专家打分，同时收集该医生的（3）既往退镜时间（WT），以及（4）既往的息肉检出率（ADR）。

然后做（1）AI评价结果与（2）专家打分（3）WT、（4）ADR的Pearson's Correlation Tests，得出下一个表格



统计审稿人提出问题如下：The authors aimed to validate their AI system using 11 endoscopists from 4 institutions. The authors used a very simple analysis approach, Pearson correlation completely ignoring that there was an added variation due to the endoscopists being from 4 institutions and each endoscopist assessed many images which again introduces statistical dependencies. A correlation is not a proper method for validation of an AI system. The authors should have used a linear mixed model for repeated measures with a random term for the Institution variable. Moreover, the study is not sufficiently powered given there were only 11 endoscopists.

他们建议通过a linear mixed model for repeated measures with a random term for the Institution variable，对这个问题不太理解，请各位老师指导。