Liver Biopsy Reads: Statistical Issues

Amrik Shah

Karma Statistics, LLC

Disclosures

- Consultant to:
  - Yaqrit Discovery Ltd
  - HighTide Therapeutics
  - Akero Therapeutics

- Stock ownership in:
  - Intercept Pharmaceuticals

Liver Biopsy Reads - Statistical Implications

- Limitations of Kappa (inter- and intra-)
  - Sensitivity is “the” relevant metric for quality (accuracy) of reads
  - Published Kappas and equivalent Sensitivity values

- Underestimation of Treatment Effect Size

Kappa and its Misdeeds

- Kappa is a measure of agreement between 2 readers (adjusted for chance agreement)
- Kappa does not take into account the accuracy of the read
  - “agreement” on incorrect values also adds to the kappa
- Kappa “depends” upon # categories of response
  - Kappa of 0.6 for ballooning (0-2) is ~ equivalent to 0.7 (1-4) for Fibrosis
- Many Kappas – Cichetti-Allison, Shrout-Fleiss, Fleiss-Cohen etc

Hence, kappa values more often tend to mislead/misinform with regard to quality/accuracy of reads

Published w-Kappa Values

Parameter	Kleiner 2019 (N=446)	Kleiner 2005 (N=32)	Davison 2020 (N=339)	Newsome 2021[a] (N=320)
Fibrosis	0.75	0.84	0.44	0.61 – 0.65
Lobular Inflammation	0.46	0.45	0.33	0.38 – 0.39
Ballooning	0.54	0.56	0.52	0.41 – 0.61
Steatosis	0.77	0.79	0.61	0.63 – 0.76

[a] The range is based on 2 values from Baseline and Week 72 slides.

Sources: Kleiner 2005, Kleiner 2019, Davison 2020, Newsome 2021.

Sensitivity & Kappa – Simulated Data

Sensitivity is – probability of a "correct read", i.e.

- *Prob of reading F2 if true fibrosis stage of slide is F2*

or

- *Prob of reading B2 if true ballooning stage is B2*

Example:

- Consider **300 slides, 100 each with ballooning "true" value 0, 1 and 2.**
- Reader will read 300 slides twice, say 3 months apart

- Assume the following:
    - Sensitivity is the same (0.7 for B0, B1 and B2 slides) for all grades of ballooning.
    - No read score can be more than 1 stage/grade wrong (for simplicity).
    - Prob (Under-read) = 0.2, Prob (Over-read) = 0.1
    - For "true" B0, there is no under-read, hence Prob(Over-read) = 0.3
    - For "true" B2, there is no over-read, hence Prob(Under-read) = 0.3

Simulated Data: Sensitivity & Kappa

	BALLOONING (0-2)		FIBROSIS (1-4)	
SENSITIVITY	w-KAPPA*	AGREEMENT	w-KAPPA*	AGREEMENT
0.7	0.45	56.7%	0.61	56.0%
0.8	0.61	67.4%	0.72	67.2%
0.9	0.79	81.9%	0.85	81.8%

* Landis, J.R.; Koch, G.G. (1977)

Underestimation of Treatment Effect Size

- Reading Error always dilutes Treatment Effect size

  - only ”accurately read” slides contribute to effect size

- From published/observed kappas,

  Fibrosis sensitivity of 0.7 is reasonable

Example: NASH trial setting focusing on Fibrosis endpoint

Endpoint is binary, BUT “improve”, “stable” and “worsen” buckets must be considered when assessing impact of reading error

Implications in a NASH Trial – Setting the Stage

Consider a 2-arm study: active vs control.

- 200 F2 subjects read into ITT (N=100 in each arm)

For now, assume baseline reads are accurate.

Some High Level Assumptions for EOT reads:

- Error in reading cannot exceed 1-stage
- Maximum change (from baseline) at EOT cannot exceed 1-stage

Fibrosis Endpoint: Dilution of Effect Size

TRT ARM	Expected Stage at EOT	TRUE Stage at EOT	Stage Read by Pathologist	OBSERVED Responders
ACTIVE (N = 100)	30% improve	F1 N=30	F0/F1 N=21; F2 N=9	21
	55% stable	F2 N=55	F1 N=11; F2/F3 N=44	11
	15% worsen	F3 N=15	F2/F3 N=15	0
CONTROL (N = 100)	10% improve	F1 N=10	F0/F1 N=7; F2 N=3	7
	60% stable	F2 N=60	F1 N=12; F2/F3 N=48	12
	30% worsen	F3 N=30	F2/F3 N=30	0

Observed Results: **32% vs 19% [13% △) vs**
Hypothesized Results: **30% vs 10% [20% △]**

Primary Endpoint: Incorporating NAS components (~80% joint accuracy for responder), then result is of order: **25.6% vs 15.2% [Delta = 10.4%]**

Key Takeaway Learnings

- Endpoint based on Biopsy Reads has severe limitations and not appropriate for assessing drug efficacy.

- Impact of reading error CANNOT be overcome by increasing sample size

  -Doubling sample size still yields same % Delta

- If forced to stick with Biopsy Reads, the dilution of effect size MUST be considered for Benefit-Risk assessment.

  - Dilution may range from 30% - 60%