Using Prediction Intervals to Evaluate Putative Surrogate Measurements

Session PS6b - Challenges and Recommendations in Using Biomarker/Surrogate Endpoints for the Accelerated Approval. *Organizer(s): Yeh-Fong Chen, FDA; Aijun Gao, Covance; Min Min, FDA; Shiling Ruan, Novartis; Chair(s): Yeh-Fong Chen, FDA; Aijun Gao, Covance*

ASA Regulatory Industry Statistics Workshop

2020 Fri September 25, 3:30-4:45pm

Gene Pennello[1]

FDA/CDRH; [1]Division of Imaging, Diagnostics, & Software Reliability

This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

…

Background

- **Acronyms**
  - Steatosis or nonalcoholic fatty liver (NAFL)
  - Nonalcoholic fatty liver disease (NAFLD)
  - Nonalcoholic Steatohepatitis (NASH)
- **NAFLD disease progression:**
  - Chronic inflammation (steatohepatitis or NASH) -> fibrosis -> ultimately cirrhosis
  - A subgroup of patients with NAFL will progress to NASH and subsequent cirrhosis

Noncirrhotic Nonalcoholic Steatohepatitis With Liver Fibrosis: Developing Drugs for Treatment

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Outline

- Background
- Diagnostic biomarkers for liver disease
  - Statistical Considerations
  - Prediction Intervals
- Surrogate endpoint evaluation
  - Statistical Considerations
- Discussion

Diagnostic Biomarkers

Statistical Considerations

Biomarkers Needed for Liver Histology

- **FDA Guidance**
  - At this time, reliable diagnosis and staging of NASH can only be made by histopathological examination of a liver biopsy specimen.
  - Liver biopsy, however, is an invasive procedure that is associated with occasional morbidity and, in rare circumstances, mortality.
  - The use of liver biopsies in clinical trials poses significant logistical challenges (e.g., cost, availability of pathologists with specific expertise in NASH); in addition, some patients are reluctant or unwilling to undergo biopsy.
  - Therefore, noninvasive biomarkers are needed (including imaging biomarkers) to supplant liver biopsy and provide a comparable or superior ability to accurately diagnose and assess various grades of NASH and stages of liver fibrosis. Identification and validation of such biomarkers could significantly accelerate drug development in NAFLD.
  - FDA encourages sponsors to consider biomarker development.

Biomarkers Needed for Liver Histology

- Sanyal AJ, Brunt EM, Kleiner DE, Kowdley KV, Chalasani N, Lavine JE, Ratziu V, McCullough A. Endpoints and Clinical Trial Design for Nonalcoholic Steatohepatitis. *Hepatology* 2011 Jul; 54(1): 344-53:
  - "There is currently considerable interest in the development of noninvasive biomarkers for
    (1) the diagnosis of NASH,
    (2) the fibrosis stage, and
    (3) the effect of treatment of NASH."

Biomarkers for Disease Stage

- **Fibrosis stage**: METAVIR scoring system:
  - F0—no **fibrosis**
  - F1—portal **fibrosis**
  - F2—periportal **fibrosis**
  - F3—bridging **fibrosis**
  - F4—cirrhosis.

Ordinal Disease Stage Modeling

- Ordinal disease stages are often dichotomized before analysis of their association with biomarker, which discards information.
- For $J = 5$ categories of fibrosis score $F = 0,1,2,3,4$ and biomarker X, consider **cumulative version of the proportional odds model**:

$$logit\big(Pr(F \leq j|X = x)\big) = \alpha_j + \beta x \quad \forall j = 1,..J$$

- For a unit increase in x, the **cumulative log odds ratio**

$$\beta = \log\left[\frac{Pr(F \leq j|X = x+1)}{Pr(F > j|X = x+1)}\left(\frac{Pr(F \leq j|X = x)}{Pr(F > j|X = x)}\right)^{-1}\right]$$

is the same for $\forall$ **cutoffs** $\boldsymbol{j = 1, 2, 3, \ldots, J}$ and is

- The model holds if X is a linear predictor of an unknown, latent continuous variable Y to which $J-1$ cut-offs are applied to obtain the ordinal variable F.
- **References**: McCullagh, 1980; Agresti, 2010; Johnson and Albert, 1999; Scott and Goldberg, 1997.

Verification Bias

- To evaluate a diagnostic test (e.g., **biomarker**), the clinical reference standard (e.g., **liver biopsy**) must be available for verifying disease status (e.g., **fibrosis stage**).
- In a pivotal study, the clinical reference standard may only be available for a non-random subset of study subjects because
  - reference standard is invasive (e.g., liver biopsy).
  - patients are reluctant or unwilling to undergo the reference (e.g., liver biopsy)
- Non-random selection of subjects for verification of disease status introduces **verification bias (also call referral bias).**

Verification Bias

- Imputation strategies for disease status:
  - **Missing at Random** (MAR)
  - **Worse Case:** disagrees with test result
  - **Non-informative**: disease status imputations are not associated with test result.

Diagnostic “Catch-22”

- A test is often ordered precisely when a physician is *unsure* if a subject should be referred to the invasive reference standard procedure or not.
- Thus, the **intended use population of the test is necessarily larger than those who actually get the clinical reference standard**.
- Using verified subjects $(V = 1)$, fit

$$\Pr(D = 1 | T = t, Z = z)$$

Histologic Categorization of NAFLD

- Factors affecting quality of the histologic data:
  - manner of procurement (intraoperative techniques may induce inflammation),
  - type of biopsy (needle core vs. wedge),
  - biopsy location,
  - dimensions of the biopsy core,
- Inherent variability in subjective assessment of liver histology.

Imperfect Reference Bias

- An imperfect reference standard will sometimes misclassify disease status.
- Imperfect reference will tend to attenuate test accuracy if it is conditionally independent of test result.
- Imperfect reference will tend to inflate test accuracy if it and test are positively dependent.
  - EX. Nucleic acid amplification tests (NAAT) evaluated for accuracy using Patient Infection Status Algorithm (PISA), which depends on comparator NAAT assays (Hadgu 2012)
  - Study reader assessments evaluated for accuracy against a reference defined as expert reader consensus.

Hadgu A, Dendukuri N, Wang L. Evaluation of Screening Tests for Detecting Chlamydia trachomatis Bias Associated With the Patient-infected-status Algorithm. Epidemiology 2012;23: 72– 82.

Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol 1997 Jul 15; 146(2): 195-203.

Interchangeability

- In lieu of evaluating <u>accuracy</u> of biomarker, consider <u>individual equivalence</u>

$$IE = \sqrt{E(Y - X)^2 - E(X_1 - X_2)^2}$$

- Biomarker variable Y is <u>*interchangeable*</u> with standard variable X if variation *between* Y and X is the same as variation *within* the standard (repeated measures X_1 and X_2) .

Barnhart H, Kosinski A, Haber M. Assessing Individual Agreement *J Biopharm Stat* 2007; 17: 697–719
Obuchowski NA. Can electronic medical images replace hard-copy film? Defining and testing the equivalence of diagnostic tests. Stat Med 2001; 20:2845–2863.
Obuchowski NA, Subhas N, Schoenhagen P. Testing for Interchangeability of Imaging Tests. Acad Radiol 2014 Nov;21(11):1483-9.
Schall R, Luus HG. On population and individual bioequivalence, *Statist Med* 1993; 12: 1109-1124.

Variability in Histological Result

- To assess **repeatability** of Histological Categorical Result,

  consider **Gini concentration index**, the probability that a pair of results disagree (Light and Margolin, 1971)

Imprecision in Categorical Measurement

- **Gini concentration index** $g(\boldsymbol{p})$ is probability that two results disagree:

$$g(\boldsymbol{p}) = 1 - \sum_{k=1}^{K} p_k^2 = \Pr(Y_1 \neq Y_2)$$

- $g(\boldsymbol{p})$ behave like a measure of variability.
  - For continuous data, variance *VY* is a function of the sum of squared pairwise differences. For categorical data, let a pairwise difference equal 1 if the two results disagree, or 0 if they agree. Then, applying the function yields $g(\boldsymbol{p})$.
- $g(\boldsymbol{p})$ is amenable to analysis of variance to determine which factors contribute the most variability among repeated measurements:
  - Light, R.J. and Margolin, B.H., "An Analysis of Variance for Categorical Data," J Amer Statist Assoc, 1971; 66: 534-44.
  - Mittlbock M, Schemper M. Explained variation for logistic regression. Statist Med 1996; 15: 1987-1997.

Gini Concentration

N	0	1+	2+	3+
10	7	1	1	1
p_k	0.7	0.1	0.1	0.1

- How do we evaluate *precision*?

$$g(\boldsymbol{p}) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 = 0.48$$

$\min[g(\boldsymbol{p})] = 0$ if $p_k = 1$ for some k

$\max[g(\boldsymbol{p})] = (K-1)/K$ if $p_k = 1/K$ for all k

Reader Concordance

- Kappa is
  - sensitive to the marginal distribution of the categorical data (Crewson PE, *AJR*: 184, May 2005).
  - may not be very interpretable clinically because it is a scaled, unitless measure.
- Positive and negative percent agreements (PPA, NPA) are conditional probabilities of agreement of on one reader with the result from the other reader.
- Neither reader is a reference. Thus PPA and NPA conditional probabilities are not preferred.

Reader Concordance

- **Dice similarity index (DSI)**
  - Called **ppos** or **pneg** in Cicchetti and Feinstein.
  - It is the average of the two conditional agreement probabilities weighted by their respective marginal distributions.
  - In CDRH, we call them *average positive* and *average negative* agreement (APA, ANA).
  - APA or ANA standardized by the probability of random agreement equals kappa (Fleiss)

Dice LR. Measures of the amount of ecologic association between species. *Ecology.* 1945;26(3):297-302.
Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol.* 2004;11(2):178-189.
Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics.* 1975;31(3):651-659.
Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol.* 1990;43(6):551-558.

Average Negative Agreement (ANA)

	$R2$		
$R1$	$-$	$+$	
$-$	p_{11}	p_{12}	$p_{1\bullet}$
$+$	p_{21}	p_{22}	$p_{2\bullet}$
	$p_{\bullet 1}$	$p_{\bullet 2}$	1

$$ANA = \frac{2p_{11}}{p_{1\bullet} + p_{\bullet 1}}$$

$$= w\frac{p_{11}}{p_{1\bullet}} + (1-w)\frac{p_{11}}{p_{\bullet 1}}$$

$$= wNPA_{2|1} + (1-w)NPA_{1|2},$$

$$w = \frac{p_{1\bullet}}{p_{1\bullet} + p_{\bullet 1}}$$

Prediction

- **EX 1**. Can arterial partial pressure of carbon dioxide (**PaCO2**) replace intramucosal **pH**, used to assist in decision making for critically ill patients?
- **EX 2**. Can **glycated albumin** replace **fructosamine** as a marker of hyperglycemia?
- Consider **prediction interval** for new value of the standard measurement given the value of the proposed replacement measurement.

EX 1. PaCO2 vs. pH

Statistics Notes

Calculating correlation coefficients with repeated observations: Part 1—correlation within subjects

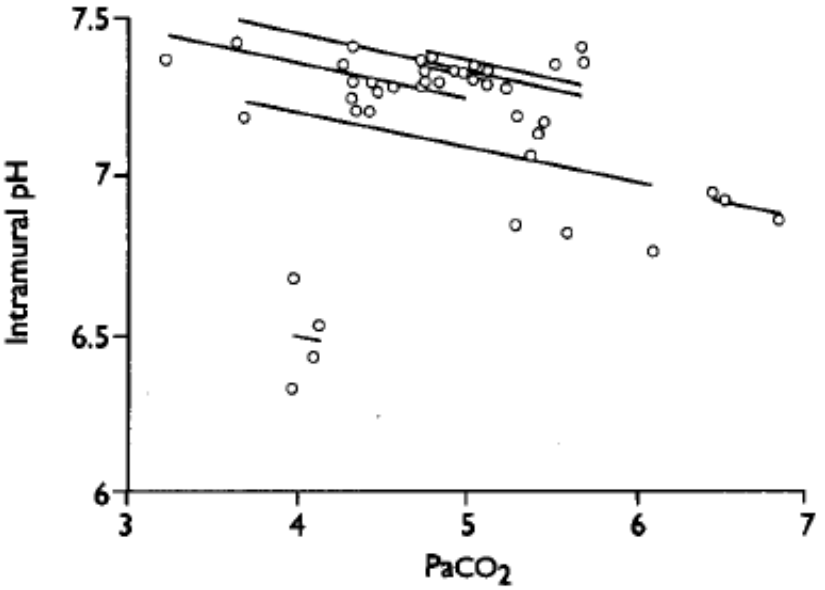
J Martin Bland, Douglas G Altman

In an earlier *Statistics Note*[1] we commented on the analysis of paired data where there is more than one observation per subject, as shown in table I. We pointed out that it could be highly misleading to analyse such data by combining repeated observations from several subjects and then calculating the correlation coefficient as if the data were a simple sample. This note is a response to several letters about the appropriate analysis for such data.

TABLE I—*Repeated measurements of intramural pH and $Paco_2$ for eight subjects*[2]

Subject	pH	$Paco_2$	Subject	pH	$Paco_2$
1	6·68	3·97	5	7·30	4·32
1	6·53	4·12	5	7·37	3·23
1	6·43	4·09	5	7·27	4·46
1	6·33	3·97	5	7·28	4·72
2	6·85	5·27	5	7·32	4·75
2	7·06	5·37	5	7·32	4·99
2	7·13	5·41	6	7·38	4·78
2	7·17	5·44	6	7·30	4·73
3	7·40	5·67	6	7·29	5·12
3	7·42	3·64	6	7·33	4·93
3	7·41	4·32	6	7·31	5·03
3	7·37	4·73	6	7·33	4·93
3	7·34	4·96	7	6·86	6·85
3	7·35	5·04	7	6·94	6·44
3	7·28	5·22	7	6·92	6·52
3	7·30	4·82	8	7·19	5·28
3	7·34	5·07	8	7·29	4·56
4	7·36	5·67	8	7·21	4·34
4	7·33	5·10	8	7·25	4·32
4	7·29	5·53	8	7·20	4·41
4	7·30	4·75	8	7·19	3·69
4	7·35	5·51	8	6·77	6·09
5	7·35	4·28	8	6·82	5·58
5	7·30	4·44			

which shows how the variability in pH can be partitioned into components due to different sources. This method is also known as analysis of covariance and is equivalent to fitting parallel lines through each subject's data (see figure). The residual sum of squares



pH against $Paco_2$ for eight subjects, with parallel lines fitted for each subject

in table II represents the variation about these lines. We remove the variation due to subjects (and any other nuisance variables which might be present) and express the variation in pH due to $Paco_2$ as a proportion of what's left:

$$\frac{\text{Sum of squares for } Paco_2}{\text{Sum of squares for } Paco_2 + \text{ residual sum of squares}}$$

The choice of analysis for the data in table I depends on the question we want to answer. If we want to know whether subjects with high values of intramural pH also tend to have high values of $Paco_2$ we are interested in whether the average pH for a subject is related to the subject's average $Paco_2$. We can use the correlation between the subject means, which we shall describe in a subsequent note. If we want to know whether an increase in pH within the individual was associated with an increase in $Paco_2$ we want to remove the differences between subjects and look only at changes within.

To look at variation within the subject we can use multiple regression. We make one of our variables, pH or $Paco_2$, the outcome variable and the other variable and the subject the predictor variables. Subject is treated as a categorical factor using dummy variables [3 4] and so has seven degrees of freedom. We use the analysis of variance table [3 4] for the regression (table II),

TABLE II—*Analysis of variance for the data in table I*

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability
Subjects	7	2·9661	0·4237	48·3	<0·0001
$Paco_2$	1	0·1153	0·1153	13·1	0·0008
Residual	38	0·3337	0·0088		
Total	46	3·3139	0·0720		

The magnitude of the correlation coefficient within subjects is the square root of this proportion. For table II this is:

$$\sqrt{\frac{0{\cdot}1153}{0{\cdot}1153+0{\cdot}3337}}=0{\cdot}51$$

The sign of the correlation coefficient is given by the sign of the regression coefficient for $Paco_2$. Here the regression slope is −0·108, so the correlation coefficient within subjects is −0·51. The P value is found either from the F test in the associated analysis of variance table, or from the *t* test for the regression slope. It doesn't matter which variable we regress on which; we get the same correlation coefficient and P value either way.

If we incorrectly calculate the correlation coefficient ignoring the fact that we have 47 observations on only 8 subjects, we get −0·07, P=0·7. Hence the correct analysis within subjects reveals a relation which the incorrect analysis misses.

1 Bland JM, Altman DG. Correlation, regression, and repeated data. *BMJ* 1994;308:896.

2 Boyd O, Mackay CJ, Lamb G, Bland JM, Grounds RM, Bennett ED. Comparison of clinical information gained from routine blood-gas analysis and from gastric tonometry for intramural pH. *Lancet* 1993;341:142-6.

3 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991.

4 Armitage P, Berry G. *Statistical methods in medical research.* 3rd ed. Oxford: Blackwell, 1994.

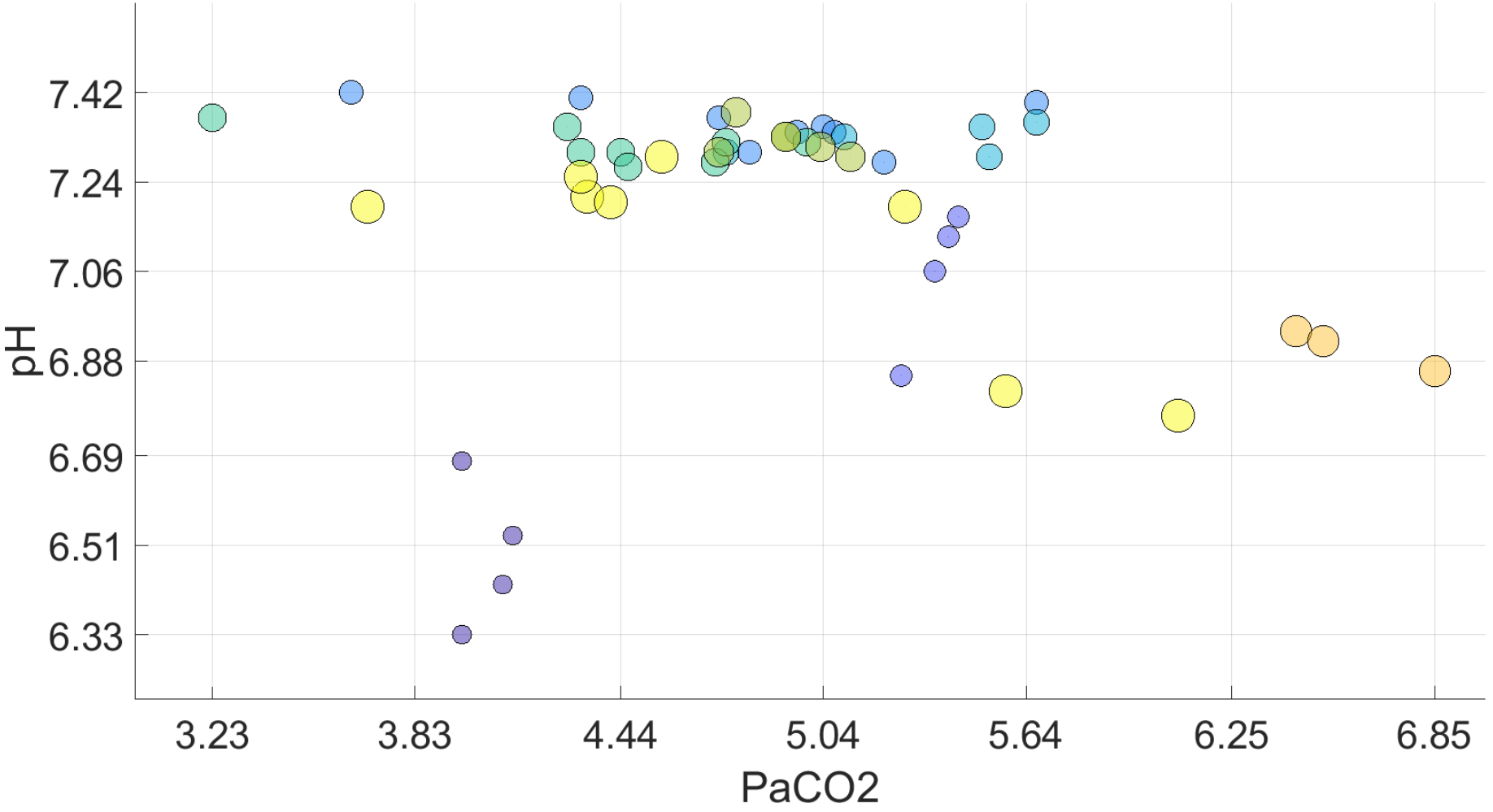
TABLE I—*Repeated measurements of intramural pH and* $Paco_2$ *for eight subjects*[2]

Subject	pH	$Paco_2$	Subject	pH	$Paco_2$
1	6·68	3·97	5	7·30	4·32
1	6·53	4·12	5	7·37	3·23
1	6·43	4·09	5	7·27	4·46
1	6·33	3·97	5	7·28	4·72
2	6·85	5·27	5	7·32	4·75
2	7·06	5·37	5	7·32	4·99
2	7·13	5·41	6	7·38	4·78
2	7·17	5·44	6	7·30	4·73
3	7·40	5·67	6	7·29	5·12
3	7·42	3·64	6	7·33	4·93
3	7·41	4·32	6	7·31	5·03
3	7·37	4·73	6	7·33	4·93
3	7·34	4·96	7	6·86	6·85
3	7·35	5·04	7	6·94	6·44
3	7·28	5·22	7	6·92	6·52
3	7·30	4·82	8	7·19	5·28
3	7·34	5·07	8	7·29	4·56
4	7·36	5·67	8	7·21	4·34
4	7·33	5·10	8	7·25	4·32
4	7·29	5·53	8	7·20	4·41
4	7·30	4·75	8	7·19	3·69
4	7·35	5·51	8	6·77	6·09
5	7·35	4·28	8	6·82	5·58
5	7·30	4·44			

Bland JM, Altman DG. Calculating correlation coefficients with repeated observations: Part I -correlation within subjects. *BMJ* 1994;310:446.

Boyd 0, Mackay CJ, Lamb G, Bland JM, Grounds RM, Bennett ED. Comparison of clinical information gained from routine blood-gas analysis and from gastric tonometry for intramural pH. *Lancet* 1993;341:142-6.

Scatterplot



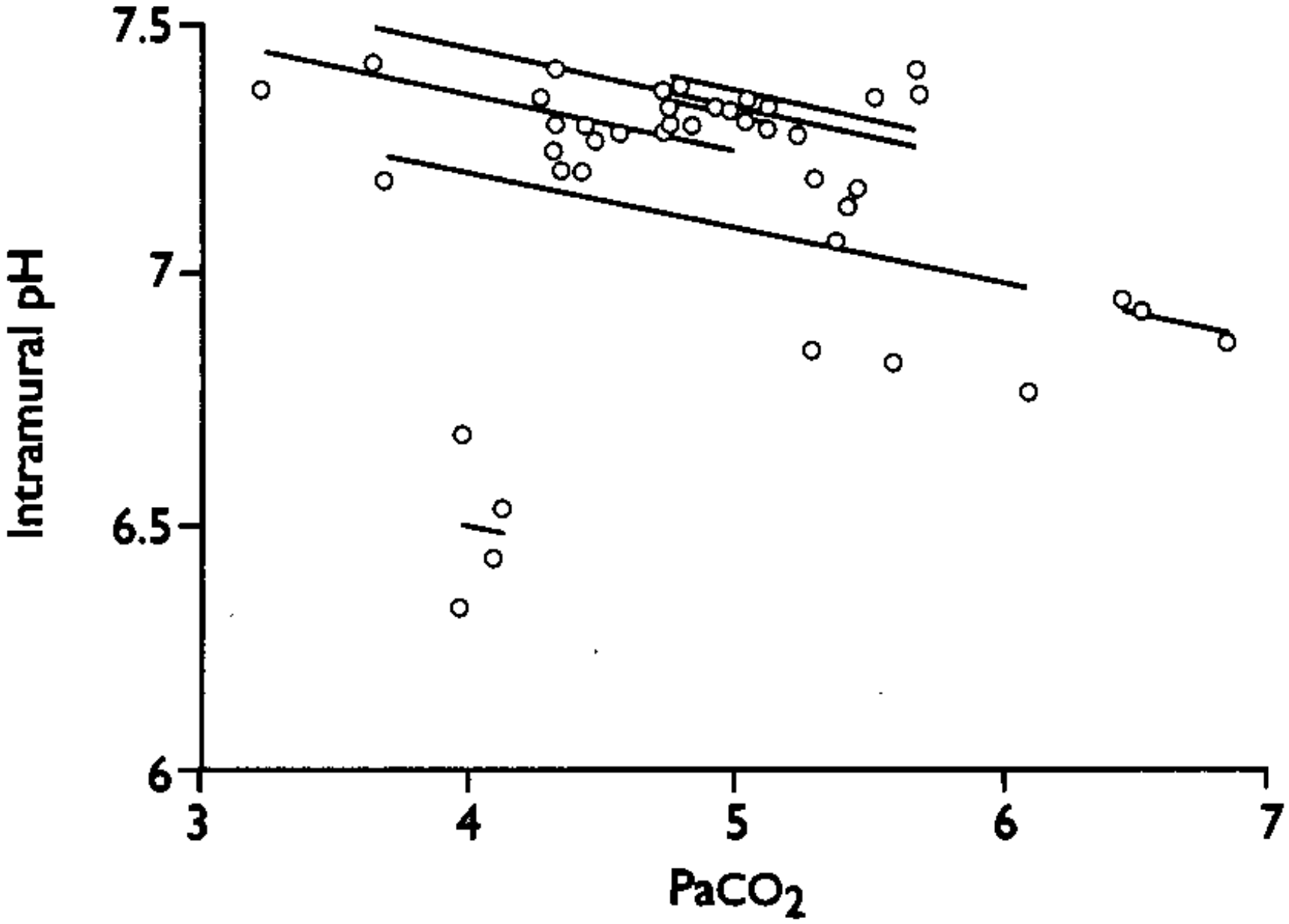
Pooled Analysis

If observations are pooled across subjects, correlation between $PaCO_2$ and pH is

$$r = -0.07 \ (p \ value \ 0.7)$$

If use ANCOVA to remove subject effects on *pH*, *partial correlation* between $PaCO_2$ and pH is

$$r = -0.51 \ (p \ value \ 0.0008)$$



pH against $Paco_2$ for eight subjects, with parallel lines fitted for each subject

ANCOVA Model

- **Notation**

$Y =$ standard quantity used in practice

$Z =$ new quantity proposed to replace Y

$X =$ dummy variable matrix of subject ids:

$$X = BD\left(\mathbf{1}_{n_i}\right),$$

$n_i =$ number observations for subject i

- **Model**

$$EY = X\beta + Z\gamma$$

$$VY = \sigma^2$$

Least Squares Estimates

$$\widehat{\boldsymbol{\gamma}} = \boldsymbol{s_{yz}s_{zz}^{-1}} = \boldsymbol{\Sigma_{i=1}^{I} w_i \widehat{\gamma}_i},$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \bar{\boldsymbol{y}}_{\boldsymbol{i}\bullet} - \bar{\boldsymbol{z}}_{\boldsymbol{i}\bullet}\widehat{\boldsymbol{\gamma}}$$

$$s_{yz} = \Sigma_{i=1}^{I}\, s_{yzi},$$

$$s_{yzi} = \Sigma_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\bullet}\right)\left(z_{ij} - \bar{z}_{i\bullet}\right)$$

$$s_{zz} = \Sigma_{i=1}^{I} s_{zzi},$$

$$s_{zzi} = \Sigma_{j=1}^{n_i} \left(z_{ij} - \bar{z}_{i\bullet}\right)^2$$

$$w_i = s_{zzi}(s_{zz})^{-1},$$

$$\hat{\gamma}_i = s_{yzi}(s_{zzi})^{-1}$$

Inference

- For a new observation on subject i, the prediction is

$$\hat{y}_0 = \bar{y}_{i\bullet} + (z_0 - \bar{z}_{i\bullet})\hat{\gamma}$$

- For inference, assume

$$Y \sim N(X\beta + Z\gamma, \sigma^2 I)$$

$1-\alpha$ Prediction Interval (PI)

A $1-\alpha$ level confidence interval for a new observation y_0 on subject i is called a *prediction interval* and is

$$\hat{y}_0 \pm t_{1-\alpha/2,n_\bullet-p}\hat{\sigma}[\bullet]^{1/2}$$

$t_{\alpha,f} = \alpha$th quantile from the Student's t dist'n with dof f

$$\hat{\sigma}^2 = s_{yy} - s_{yz}^2 s_{zz}^{-1} \equiv s_{yy}(1-r_*^2)$$

$$[\bullet] = 1 + n_1^{-1} + s_{zz}^{-1}(z_0 - \bar{z}_{1\bullet})^2$$

$$p = I + 1$$

r_* = partial correlation of Y and Z after subjects effects are removed

1st observation on each subject

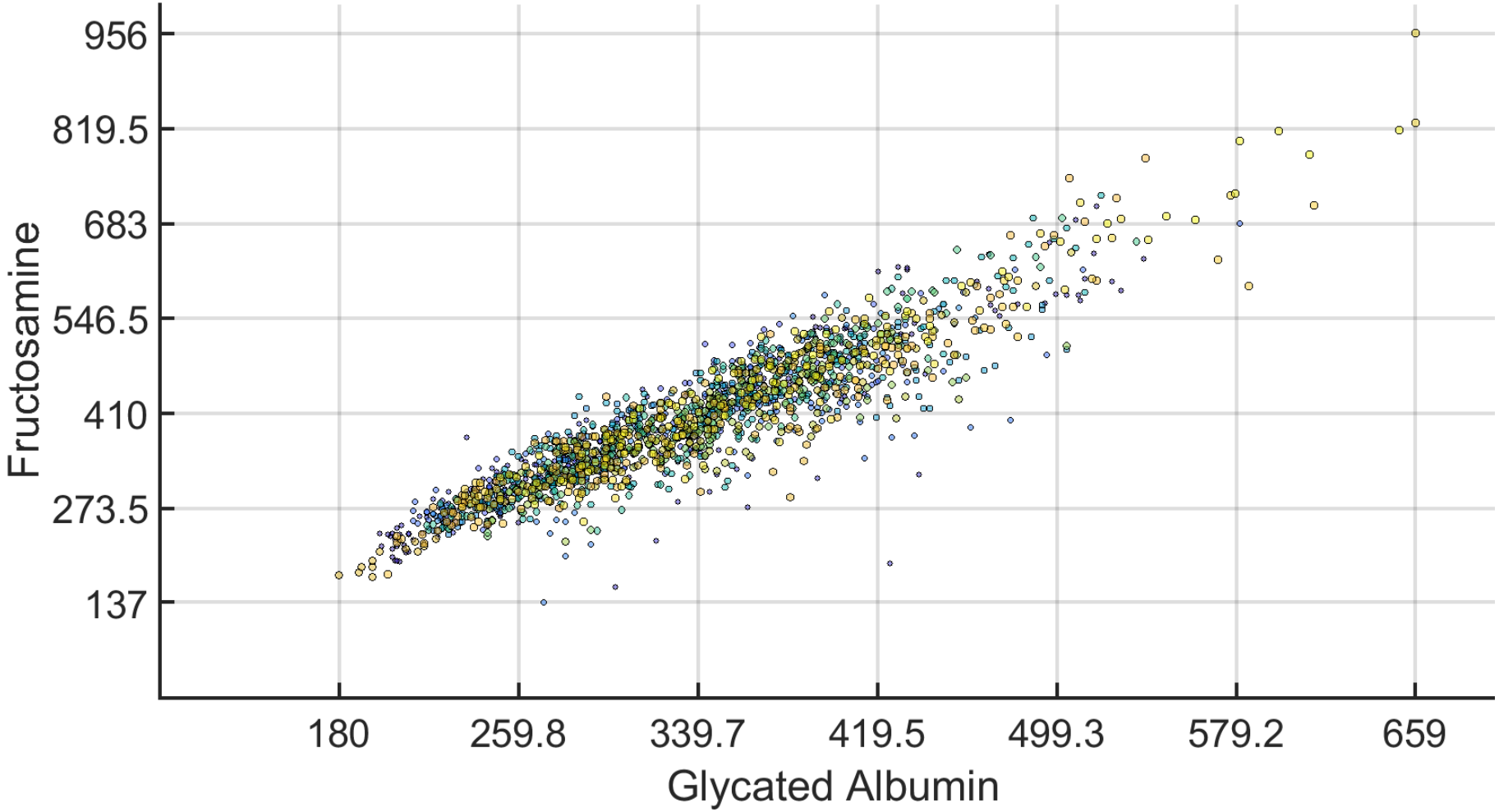
S	#Obs	pH	PaCO2	PaCO2 mean	Diff	#sds	Z Var term
1	4	6.7	4.0	4.0	-0.1	-0.7	0.0
2	4	6.8	5.3	5.4	-0.1	-1.1	0.0
3	9	7.4	5.7	4.8	0.8	9.0	0.1
4	5	7.4	5.7	5.3	0.4	3.8	0.0
5	8	7.3	4.3	4.4	-0.1	-1.3	0.0
6	6	7.4	4.8	4.9	-0.1	-1.5	0.0
7	3	6.9	6.8	6.6	0.2	2.6	0.0
8	8	7.2	5.3	4.8	0.5	5.3	0.0

1st observation on each subject

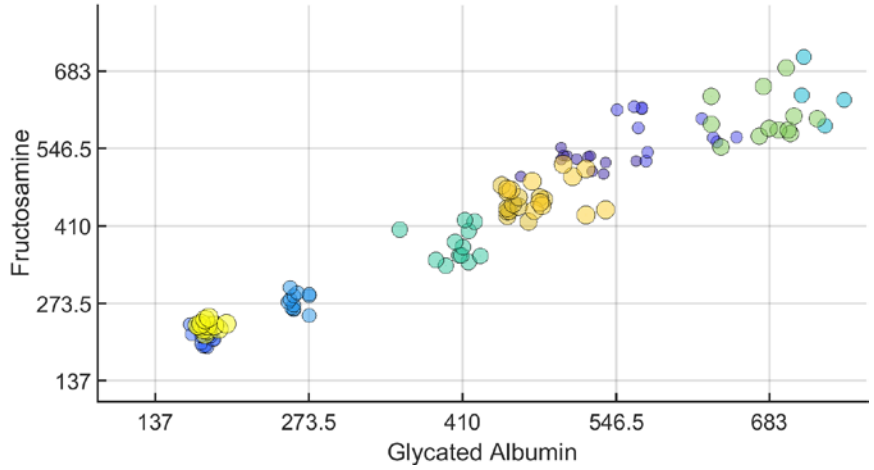
S	#Obs	pH	pH Pred	Diff	95% PI	Width	pH Cov'd?
1	4	6.7	6.4	0.3	6.1, 6.6	0.5	0
2	4	6.8	7.1	-0.2	6.9, 7.3	0.4	0
3	9	7.4	7.3	0.1	7.1, 7.5	0.4	1
4	5	7.4	7.3	0.0	7.1, 7.5	0.4	1
5	8	7.3	7.2	0.1	7.0, 7.4	0.4	1
6	6	7.4	7.3	0.1	7.1, 7.5	0.4	1
7	3	6.9	7.1	-0.2	6.8, 7.3	0.5	1
8	8	7.2	7.1	0.1	6.9, 7.3	0.4	1
							6/8

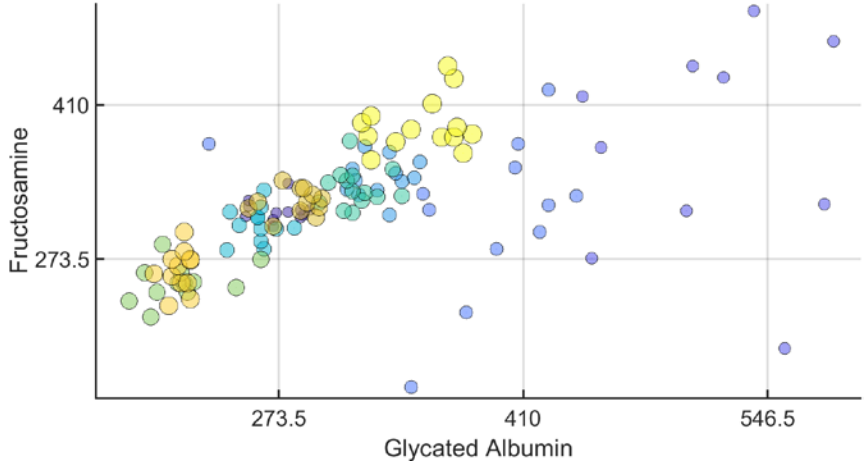
Glycated albumin proposed to replace fructosamine as a marker of hyperglycemia

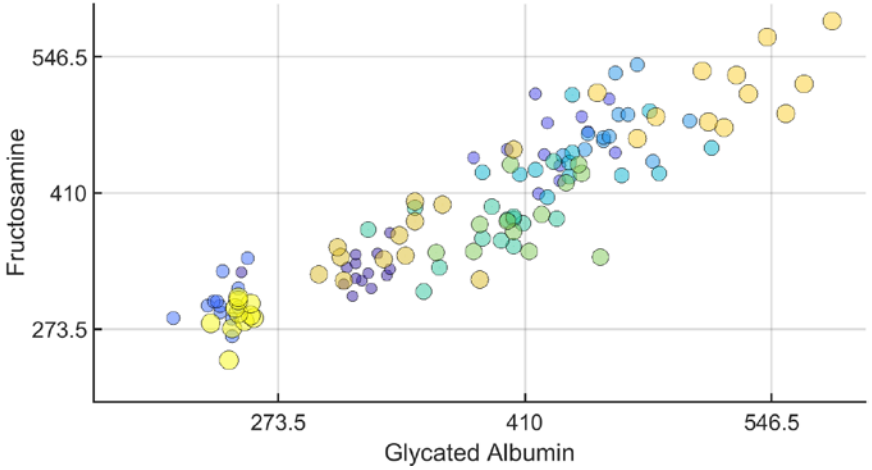
Scatterplot

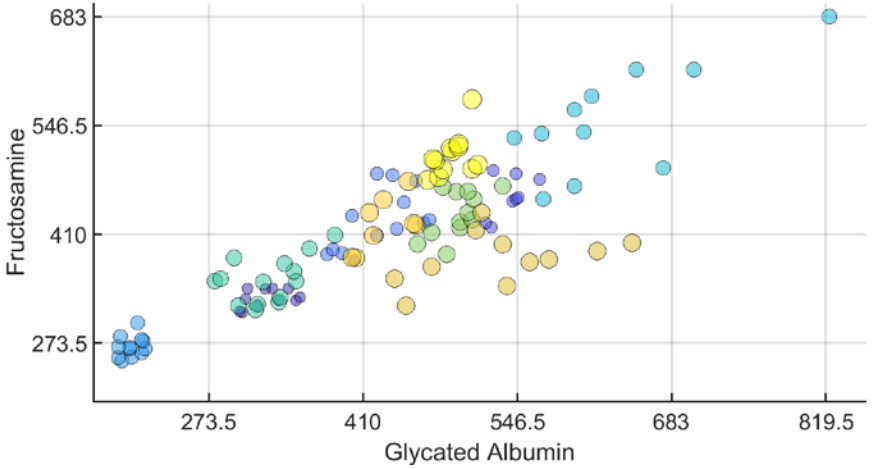


Scatterplots









1st observation on each subject

S	#Obs	FA	GA	Mean GA	Diff	#SDs	Z Var Term
1	12	497	370	401.0	-31.0	1.2	0.0
2	11	524	435	445.9	-10.9	0.4	0.0
3	12	196	207	205.7	1.3	0.1	0.0
4	12	252	260	253.5	6.5	0.2	0.0
5	4	586	528	524.8	3.3	0.1	0.0
6	12	340	331	336.6	-5.6	0.2	0.0
7	12	549	474	497.3	-23.3	0.9	0.0
8	12	417	374	368.0	6.0	0.2	0.0
9	11	429	404	386.6	17.4	0.7	0.0
10	12	219	206	207.4	-1.4	0.1	0.0
11	12	302	258	260.2	-2.2	0.1	0.0
12	10	194	425	399.8	25.2	1.0	0.0
13	12	160	303	323.4	-20.4	0.8	0.0
14	12	312	296	293.1	2.9	0.1	0.0
15	12	281	243	253.0	-10.0	0.4	0.0
16	14	314	284	285.1	-1.1	0.0	0.0
17	12	222	218	227.8	-9.8	0.4	0.0
18	12	302	258	265.3	-7.3	0.3	0.0
19	12	232	224	227.4	-3.4	0.1	0.0
20	14	361	290	306.4	-16.4	0.6	0.0

1st observation on each subject

S	#Obs	FA	Pred FA	Diff	95% PI LB	95% PI UB	Width	FA Cov'd?
1	12	497	484.8	12.2	430.7,	538.8	108.1	1
2	11	524	568.2	-44.2	514.0,	622.5	108.4	1
3	12	196	211.3	-15.3	157.3,	265.4	108.0	1
4	12	252	285.3	-33.3	231.3,	339.3	108.0	1
5	4	586	645.3	-59.3	587.3,	703.4	116.1	0
6	12	340	368.3	-28.3	314.3,	422.4	108.0	1
7	12	549	570.8	-21.8	516.7,	624.8	108.1	1
8	12	417	455.3	-38.3	401.3,	509.4	108.0	1
9	11	429	490.9	-61.9	436.7,	545.1	108.4	0
10	12	219	232.0	-13.0	178.0,	286.0	108.0	1
11	12	302	313.2	-11.2	259.2,	367.2	108.0	1
12	10	194	404.8	-210.8	350.3,	459.2	108.9	0
13	12	160	290.3	-130.3	236.2,	344.3	108.1	0
14	12	312	348.9	-36.9	294.9,	402.9	108.0	1
15	12	281	291.1	-10.1	237.1,	345.2	108.0	1
16	14	314	335.1	-21.1	281.4,	388.8	107.4	1
17	12	222	240.5	-18.5	186.4,	294.5	108.0	1
18	12	302	314.6	-12.6	260.6,	368.6	108.0	1
19	12	232	258.6	-26.6	204.5,	312.6	108.0	1
20	14	361	371.8	-10.8	318.0,	425.5	107.5	1
								16/20

Surrogate Endpoints

Statistical Considerations

Accelerated Approval in Drugs

- 21 CFR (314 and 601) Accelerated Approval Rule
  - for serious or life-threatening illness
  - it allows the use of **surrogate or non-ultimate clinical endpoints** when the effect on a surrogate end point is "reasonably likely" to predict clinical benefit
  - post-market data is required "to verify and describe the drug's clinical benefit and to resolve the remaining uncertainty as to the relation of the surrogate endpoint up on which approval was based to clinical benefit, or the observed clinical benefit to ultimate outcomes."

Guidance for Industry and FDA Staff

Qualification Process for Drug Development Tools

Qualification of Medical Device Development Tools

Guidance for Industry, Tool Developers, and Food and Drug Administration Staff

Document issued on: August 10, 2017

The draft of this guidance document was issued on November 14, 2013.

Qualification of Medical Device Development Tools (MDDTs)

- Some examples of the specific roles for MDDTs in device development include:
  - For selection of clinical study subjects:
  - To stratify patient population by predicted risk;
  - For study population enrichment;
  - For an **intermediate endpoint**;
  - For a **surrogate endpoint**

Noncirrhotic Nonalcoholic Steatohepatitis With Liver Fibrosis: Developing Drugs for Treatment

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Liver Disease Endpoints

- **Clinical Endpoints** NASH treatment should slow, halt, or reverse disease progression and improve clinical outcomes, i.e.,
  - prevent progression to cirrhosis and cirrhosis complications,
  - reduce the need for liver transplantation, and
  - improve survival
- NASH has slow progression to cirrhosis or death
- **Surrogates**. Liver histological improvements reasonably likely to predict clinical benefit to support accelerated approval:
- — Resolution of steatohepatitis (absent fatty liver disease or isolated or simple steatosis without steatohepatitis and a NAS score of 0–1 for inflammation, 0 for ballooning, and any value for steatosis) AND no worsening of NASH CRN fibrosis score.

OR

- — Improvement in liver fibrosis greater than or equal to one stage (NASH CRN fibrosis score) and no worsening of steatohepatitis (defined as no increase in NAS for ballooning, inflammation, or steatosis);

OR

- — Both

18 of the 36 cancer drugs that were approved by the FDA from 20 2012 on the basis of a surrogate endpoint, typically tumor shrinka PFS. Post-marketing studies did not indicate a survival benefit.

- Kim and Prasad JAMA Intern Med. 2015;175(12):1992-1994.

Editor's Note

Improving the Accelerated Pathway to Cancer Drug Approvals

The US Food and Drug Administration (FDA) must balance the need to bring potentially lifesaving drugs to market with the need to ensure the safety and effectiveness of these drugs. To balance these competing goals, the FDA has increasingly used the accelerated pathway, which is meant for drugs that treat serious conditions and fill an unmet medical need. Approval is based on a surrogate or an early clinical endpoint and is conditional on the completion of confirmatory trials, which are planned prior to the approval process.

Once granted, accelerated drug approvals are subject to withdrawal if "a postmarketing clinical study fails to verify clinical benefit."[1] The FDA defines *clinical benefit* as prolonging life or improving the quality of life (QoL). Withdrawal of approval is rare. The only drug for which the FDA withdrew approval—as a result of failure of confirmatory data—was bevacizumab for metastatic breast cancer in 2011. However, Medicare and other major insurers still cover bevacizumab for this indication, despite the FDA ruling or the drug's lack of clinical benefit.

In this issue of *JAMA Internal Medicine*, Rupp and Zuckerman[2] examine 18 cancer drugs that received accelerated FDA approval but were found in postmarketing confirmatory trials to have no overall survival (OS) benefit.[3] Less than half of these drugs had been studied using QoL outcomes. Although 6 drugs lack OS or QoL benefit, all but 1 (bevacizumab) have retained their approval and are still on the market.

We suggest 3 improvements to the accelerated pathway for cancer drug approvals. First, confirmatory postmarketing studies for accelerated drug approvals should include both OS and QoL outcomes because these are the 2 facets of clinical benefit currently being used by the FDA. Second, preapproved QoL measures should be published for specific drug classes. Third, anticipated or clinically significant changes in OS and in QoL measures should be defined a priori to facilitate the identification of drugs whose "postmarketing clinical study fails to verify clinical benefit."

In following the principle of "first, do no harm," the FDA should promptly withdraw approval for cancer drugs that are proven to have no clinical benefit. Removing these drugs, each of which costs between $20 000 and $170 000 per year, from the market will improve the quality and value of cancer care.

Scott R. Bauer, MD, ScM
Rita F. Redberg, MD, MSc

Corresponding Author: Scott R. Bauer, MD, ScM, Division of General Internal Medicine, University of California, San Francisco, 1545 Divisadero St, San Francisco, CA 94115 (scott.bauer@ucsf.edu).

Conflict of Interest Disclosures: None reported.

1. Food and Drug Administration, US Dept of Health and Human Services. 21 CFR §601.43. Withdrawal procedures. https://www.gpo.gov/fdsys/pkg/CFR-2014-title21-vol7/xml/CFR-2014-title21-vol7-sec601-43.xml. Accessed November 23, 2016.

2. Rupp T, Zuckerman D. Quality of life, overall survival, and costs of cancer drugs approved based on surrogate endpoints [published online November 29, 2016]. *JAMA Intern Med*. doi:10.1001/jamainternmed.2016.7761

3. Kim C, Prasad V. Cancer drugs approved on the basis of a surrogate end point and subsequent overall survival: an analysis of 5 years of US Food and Drug Administration approvals. *JAMA Intern Med*. 2015;175(12):1992-1994.

Prentice Criteria for Statistical Surrogate

- With respect to treatment indicator Z , S is a statistical surrogate endpoint for true endpoint T if
  1. $S|Z \neq S$ (treatment has effect on S)
  2. $T|Z \neq T$ (treatment has effect on T)
  3. $T|S \neq T$ (S is associated with T)
  4. $T|S, Z = T|S$ (i.e., $T \perp\!\!\!\perp Z|S$)
- Note that if condition 4 holds, then
  - no treatment on the surrogate $(S|Z = S)$ **implies**
  - no treatment effect on the true endpoint $(T|Z = T)$:

$$T|Z = \sum_S T|S, Z \bullet S|Z = \sum_S T|S \bullet S = T$$

Quantitative Measures of Surrogacy

Prentice (1989), Freedman *et al* (1992)

	Quantity	Estimate	Test
1	Effect of Z on T	β	$(T\|Z) \neq (T)$
2	Effect of Z on S	α	$(S\|Z) \neq (S)$
3	Effect of S on T	γ	$(T\|S) \neq (T)$
4	Effect of Z on T, given S	β_S	$(T\|Z,S) = (T\|S)$

↓

Proportion explained

$$PE = \frac{\beta - \beta_S}{\beta}$$

↙ ↘

Relative Effect

$$RE = \frac{\beta}{\alpha}$$

Adjusted Association

$$\rho_Z = \text{Corr}(S, T|Z)$$

Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. Statistical Validation of Surrogate Endpoints: Problems and Proposals, *Drug Information Journal* 2000; , 34, 447-454.

Surrogates

- Fleming, T. and DeMets, D. (1996). Surrogate end points in clinical trials: Are we being misled? *Ann. Int. Med.* **125**:605-613.
- “A correlate does not a surrogate make”.

Causal Surrogate

- Under Prentice criteria, $S \perp\!\!\!\perp Z$ implies $T \perp\!\!\!\perp Z$. But independence does not imply no causation.
- **Principal Surrogate** (Frangakis and Rubin, 2002): For all s,

$$Pr(T(1)|S(1) = S(0) = s) = Pr(T(0)|S(1) = S(0) = s)$$

- **Causal necessity** (Frangakis and Rubin, 2002): Causal effect of treatment Z on the true end point T can occur only if a causal effect of Z on the surrogate S has occurred.
- **Consistent surrogate** (Chen, Geng, Jia, 2007). A non-positive (non-negative) causal effect of Z on S implies a non-positive (non-negative) causal effect of Z on T.

Direct Acyclic Graphs

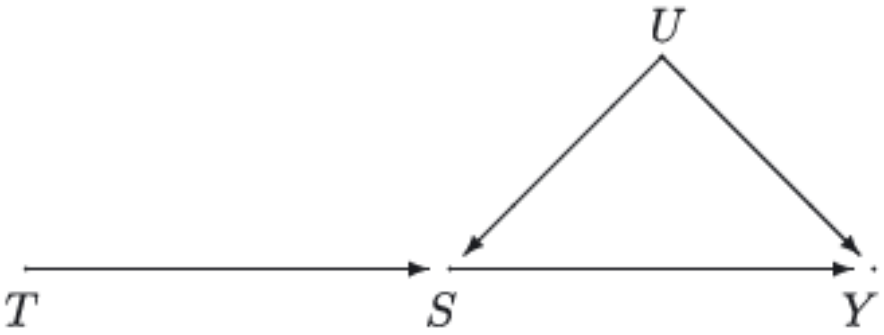


Fig. 1. *S* is a strong surrogate for the true end point *Y*

- Lauritzen, 2004; Chen, Geng, and Jia, 2007; Joffe and Greene, 2009.

ENAR

ENAR 2020 SPRING MEETING

Novel methods to evaluate surrogate endpoints

Organizer:

Ludovic Trinquart, Boston University School of Public Health

Chair:

Michael LaValley, Boston University School of Public Health

Speakers:

Layla Parast, RAND Corporation

Isabelle Weir, Boston University School of Public Health

Emily Roberts, University of Michigan

Ariel Alonso Abad, Katholieke Universiteit Leuven

Biomarkers Consortium - Workshop: Defining an Evidentiary Criteria Framework for Surrogate Endpoint Qualification

The Foundation for the National Institutes of Health (FNIH) Biomarkers Consortium, in partnership with the Food and Drug Administration's (FDA) Center for Drug Evaluation and Research, hosted a public meeting entitled Framework for Defining Evidentiary Criteria: Surrogate Endpoint Qualification Workshop on July 30th and 31st, 2018.

https://fnih.org/what-we-do/biomarkers-consortium/programs/biomarkers-consortium-workshop-defining-evidentiary-criteria-framework-surrogate-endpoint

Table 2 Criteria for validating a surrogate

(i) Define patients, treatments, and clinical endpoints for which the potential surrogate applies.
(ii) A strong statistical association between the surrogate and the clinical outcome of interest.
(iii) Strong, consistent evidence of treatment differences in the surrogate for each trial.
(iv) Treatment difference in clinical outcome within each trial is statistically explained by the surrogate.
(v) Across trials, magnitudes of treatment difference in the surrogate and in the clinical outcome are closely linked.

Weintraub WS, Lüscher TF, Pocock S. The perils of surrogate endpoints Eur Heart J 2015 Sep 1; 36(33): 2212-8.

Discussion

Diagnostic Biomarker Issues

- Ordinal Disease Stage
- Verification Bias
- Imperfect Reference Standard
- Interchangeability
- Reader Concordance
- Prediction intervals were presented, assuming measurements are normally distributed.
  - Quantitative values are necessarily > 0 and thus not normally distributed.
  - Consider basing prediction interval on log normal or gamma regression (generalized linear model).

Surrogate endpoints

- Progress has been made on a causal estimand framework for evaluating a surrogate endpoint.
- Data type (continuous, ordinal, nominal) may be different for surrogate than true clinical endpoint.
- Evaluating the adequacy of a surrogate endpoint is facilitated
  - not by scaled summaries (e.g., proportion of variation explained),
  - but by summaries in the units of the measurements (e.g., prediction interval for causal treatment effect).

Other Statistical Issues

- **Clinical Benefit of a Diagnostic**
- **Regression to the Mean** due to enrollment criterion of $NAS \geq 4$, for example:
  - Variable that is extreme on its first measurement will tend to be closer to the center of the distribution for a later measurement.
- **See Supplemental Slides**

Thank You!

Acknowledgements. Thanks to:

- Organizers for the invitation
- FDA Division of Imaging, Diagnostics, and Software Reliability for helpful comments during dry run

Supplemental

Regression to the Mean (RTM)

- NAFLD Activity Score (NAS) ranges from 0-8:
  - NAS = sum of scores of steatosis (0-3), lobular inflammation (0-3) and hepatocyte ballooning (0-2).
  - NAS ≥ 5 correlated with Dx of “definite NASH”
  - NAS ≤ 3 correlated with Dx of “not NASH”
- $NAS \geq 4$ is a frequently used inclusion criterion.

Regression to the Mean (RTM)

- $NAS \geq 4$ is a frequently used inclusion criterion.
- A baseline $NAS \geq 4$ value may tend to decrease upon repeat measurement if
  - NAS mean is < 4 and
  - NAS is subject to measurement variability.
- RTM may lead to misinterpretation of study results within treatment arms.
- Increase precision of estimated treatment effect by using ANCOVA to adjust for baseline value.

Regression to the Mean

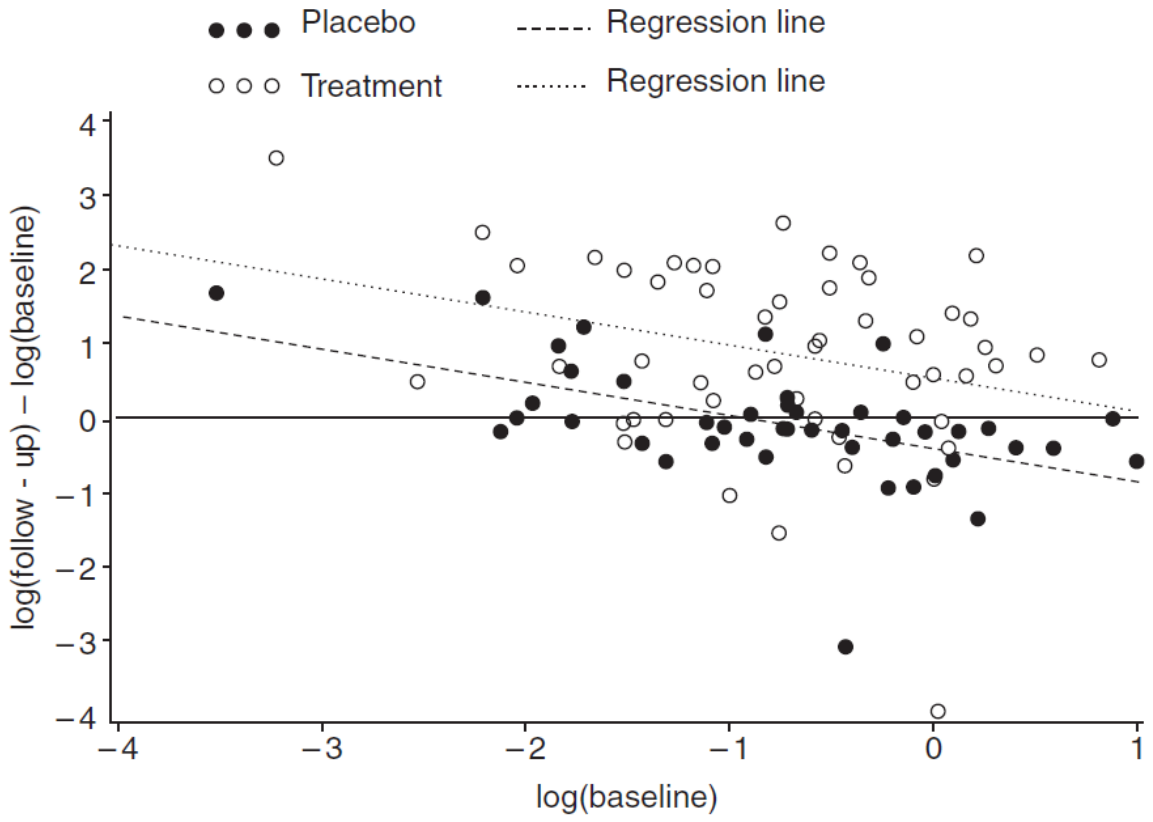


Figure 3 Scatter-plot of n = 96 paired and log-transformed betacarotene measurements showing change (log(follow-up) minus log(baseline)) against log(baseline) from the Nambour Skin Cancer Prevention Trial. The solid line represents perfect agreement (no change) and the dotted lines are fitted regression lines for the treatment and placebo groups

RTM References

Barnett AG, van der Pols JC, Dobson AG. Regression to the mean: what it is and how to deal with it. *Internat J Epidemiol* 2005; 34: 215–220.

Barnett AG, van der Pols JC, Dobson AG. Correction to: Regression to the mean: what it is and how to deal with it International Journal of Epidemiology, *Internat J Epidemiol* 2015; 44: 1748–1748.

Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 1976 Nov; 104(5): 493-8.

Senn S. Francis Galton and regression to the mean. *Significance,* 2011.

Twisk, Jos W. R. *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide, 2nd ed.* Cambridge, 2013.

Clinical Benefit of a Diagnostic

- **Clinical Benefit** - does the test support clinical decisions for patient management such as effective treatment or preventive strategies?
- **Patient Outcome Efficacy** (Fryback-Thornbury level 6)
- **Clinical Utility**. The degree to which actual use of the corresponding test in healthcare is associated with changing health outcomes, such as preventing death and restoring or maintaining health (Bossuyt et al 2012).

Bossuyt PMM, Reitsma JB, Linnet K, Moons KGM. Beyond Diagnostic Accuracy: The Clinical Utility of Diagnostic Tests Clin Chem, 1 Dec 2012, 58 (12): 1636–43.

Clinical Benefit of a Diagnostic

Table 1.1 *Criteria for a useful diagnostic/screening test*

(1) Disease should be serious or potentially so
(2) Disease should be relatively prevalent in the target population
(3) Disease should be treatable
(4) Treatment should be available to those who test positive
(5) The test should not harm the individual
(6) The test should accurately classify diseased and non-diseased individuals

Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford, 2003

Biases in Evaluating Clinical Benefit

- Lead time bias
  - Earlier detection by screening may erroneously appear to indicate beneficial effects on the outcome of a progressive disease
- Stage migration bias
  - Stage migration due to different methods of cancer staging can artifactually inflate cancer survival rates by shifting patients with a marginal prognosis out of a better prognosis group into a worse prognosis group.

Table 4. Effects of Stage Migration on Six-Month Survival Rates in the 1977 Cohort.*

Old-Data TNM Stage *	Stage Migration	New-Data TNM Stage *
	six-month survival	
I: 32/42 (76)	→ I: 22/24 (92)	I: 22/24 (92)
	↗ II: 1/1 (100)	
	↘ III: 9/17 (53)	
II: 17/25 (68)	→ II: 12/17 (71)	II: 13/18 (72)
	↘ III: 5/8 (63)	
III: 23/64 (36)	→ III: 23/64 (36)	III: 37/89 (42)
	Total 72/131 (55)	

*TNM denotes tumor, nodes, and metastases.[16] Values are numbers of patients, with percentages in parentheses.

Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *NEJM* June 1985; **312** (25): 1604–8