Normative Database Construction, Reliability, and Usability: Considerations in Premarket Review

Kristen Meier, Ph.D.
Mathematical Statistician, Division of Biostatistics, OSB/CDRH

FDA/American Glaucoma Society Workshop
on the Validity, Reliability, and Usability of
Glaucoma Imaging Devices
October 5th, 2012

Acknowledgments

- R. Lee Kramm, MSE, MD (former FDA Medical Officer)
- Rona Tang, PhD, Division of Biostatistics, CDRH/FDA

Outline

- Terminology
- Approaches for Reporting Results Relative to Database
- Database Construction and Stratification
- Database Reliability
- Database Usability

Examples of Normative Database Parameters (Measurements)

- RNFL (retinal nerve fiber layer)
  - average RNFL thickness, temporal, superior, nasal, inferior, clock hour segments, …
- ONH (optic nerve head)
  - average cup-to-disc (CDR) ratio, vertical/horizontal CDR, disc area, rim area, cup volume,…
- Ganglion Cell
  - average thickness, superior thickness, inferior thickness,…
- Macular Thickness
  - central subfield, inner/outer temporal, inner/outer superior, ...

Terminology: Types of Databases for a Parameter

- cross-sectional reference/normative database (NDB)
  - results from many individuals, each contributing a single result (e.g. one RNFL thickness result)
- person-specific database
  - results over time on the same individual
  - useful for patient monitoring

Combining Multiple Parameters

- Devices that use an algorithm combining multiple parameters from a subject in order to predict the subject’s health status are more complex and beyond the scope of this talk

- Consult with FDA through the pre-submission process

Terminology: “Normative” vs. “Reference” Database

- “Normal” has different connotations
  - observed distribution of results follows the statistical normal [Gaussian] distribution
  - typical for a group of individuals
  - absence of only the condition of interest
  - absence of all possible disease
- “Reference” is a better term
  - prevents ambiguities, incorrect assumption about “normal”
  - fosters correct interpretation of results by necessitating description of “reference” group

What a Cross-Sectional Database Provides

- characterizes inter-individual variability
- assists physician in interpreting a new individual’s result by comparison with normative or reference data
  - how common or unusual is the result with respect to a well-characterized reference group?

Reporting Results Relative to a Normative Database

- Normal Limit (NL) approach
  - flag or color code result depending on whether it is < or > normal limit [e.g. 5th percentile) from NDB
  - makes sharp [arbitrary?] distinction between ‘normal’ versus ‘outside normal limits’
- Percentile approach
  - Report result as percentile [e.g. 87th percentile] of NDB results

Single NL – Single Parameter

For a single parameter (e.g., RNFL thickness), based on chance alone,

•1 in 20 normal subjects (5%) will have a result that falls below the 5% NL, or is color-coded as ‘outside normal limits’

Multiple NL Limits for Multiple Parameters

- probability of at least one result falling below 5% NL is *greater than* 5%
- if results for 10 different [independent] parameters [e.g. RNFL, CDR, …] from a normal eye are each compared to their respective 5% NL, then the probability that at least one of the 10 results is below the 5% NL could be as high as ***~40%*** [=100 × ($1-0.95^{10}$)]

Percentile Approach

- report result (e.g., 84 microns) and percentile corresponding to the result (e.g. 6th percentile)
- familiar - used in pediatric height/weight growth charts, and educational test scores
- avoids [unnecessary?] dichotomy of normal versus outside normal

Normal Limits vs Decision Limits

- NLs alone do not necessarily discriminate between states of health (e.g., normal versus abnormal or diseased)
- developing *clinical decision limits* that discriminate between states of health requires distribution of results in other populations (e.g., diseased, undifferentiated subjects)

Results Relative to NDB

- Both the normal limits approach and percentile approach provide comparison of new subject result to results in NDB
- *A well-characterized NDB population is critical to assess NDB relevance for an individual*

FDA Question 1

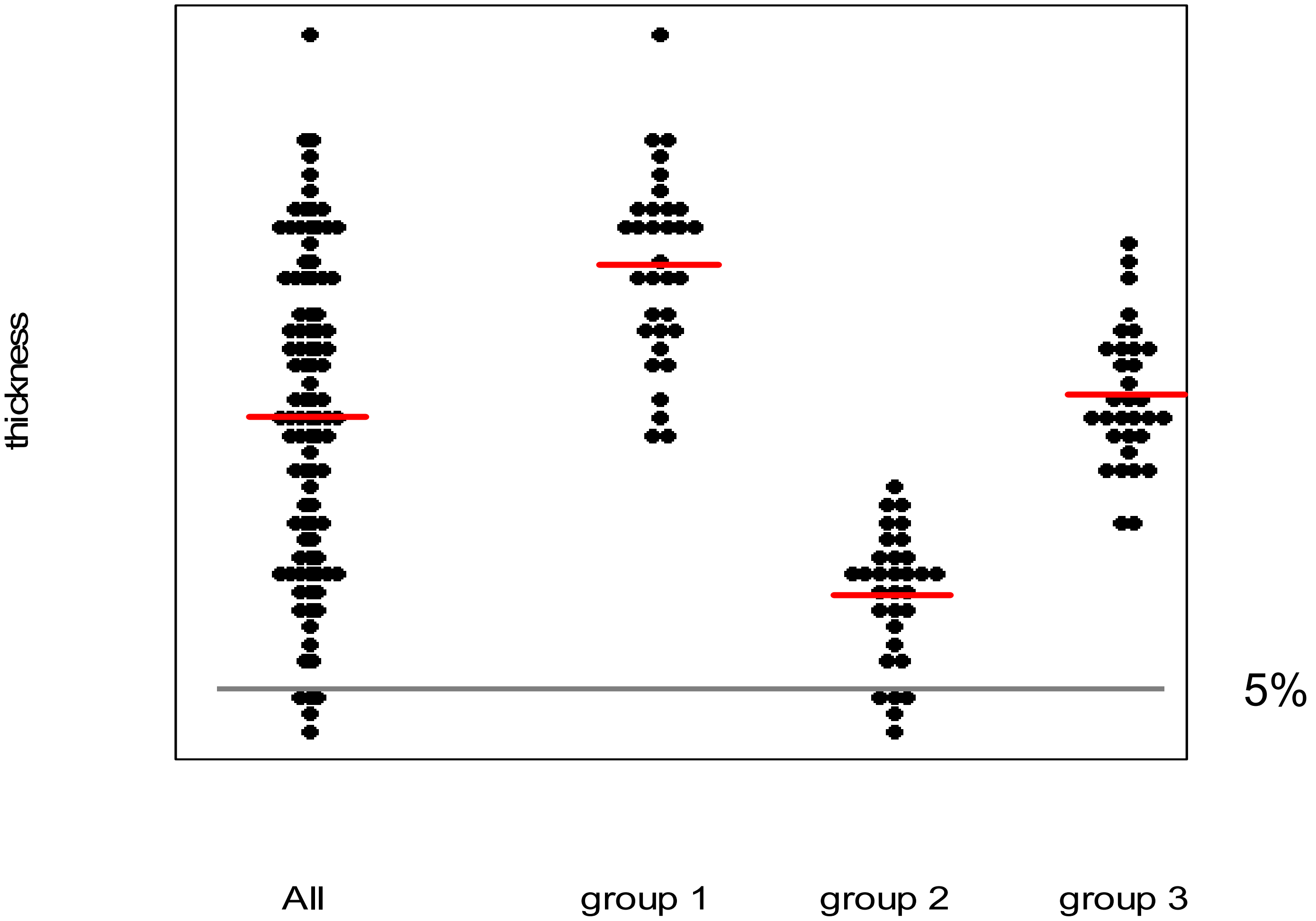
- *What clinical work-up should be done to establish a “normal” population? Who can be defined as not having disease? What patient characteristics are important to represent in the definition of a normal population?*

Database Stratification / Adjusting Normal Limits

- When results vary across covariates (e.g. different age groups, racial/ethnic groups or for different levels of image quality), the NLs may need to be ***stratified*** or ***adjusted*** for those covariates (e.g. covariate-specific NLs are needed)
- Different statistical models and approaches can be used to estimate covariate-specific NLs (e.g. linear regression, quantile regression, see references)

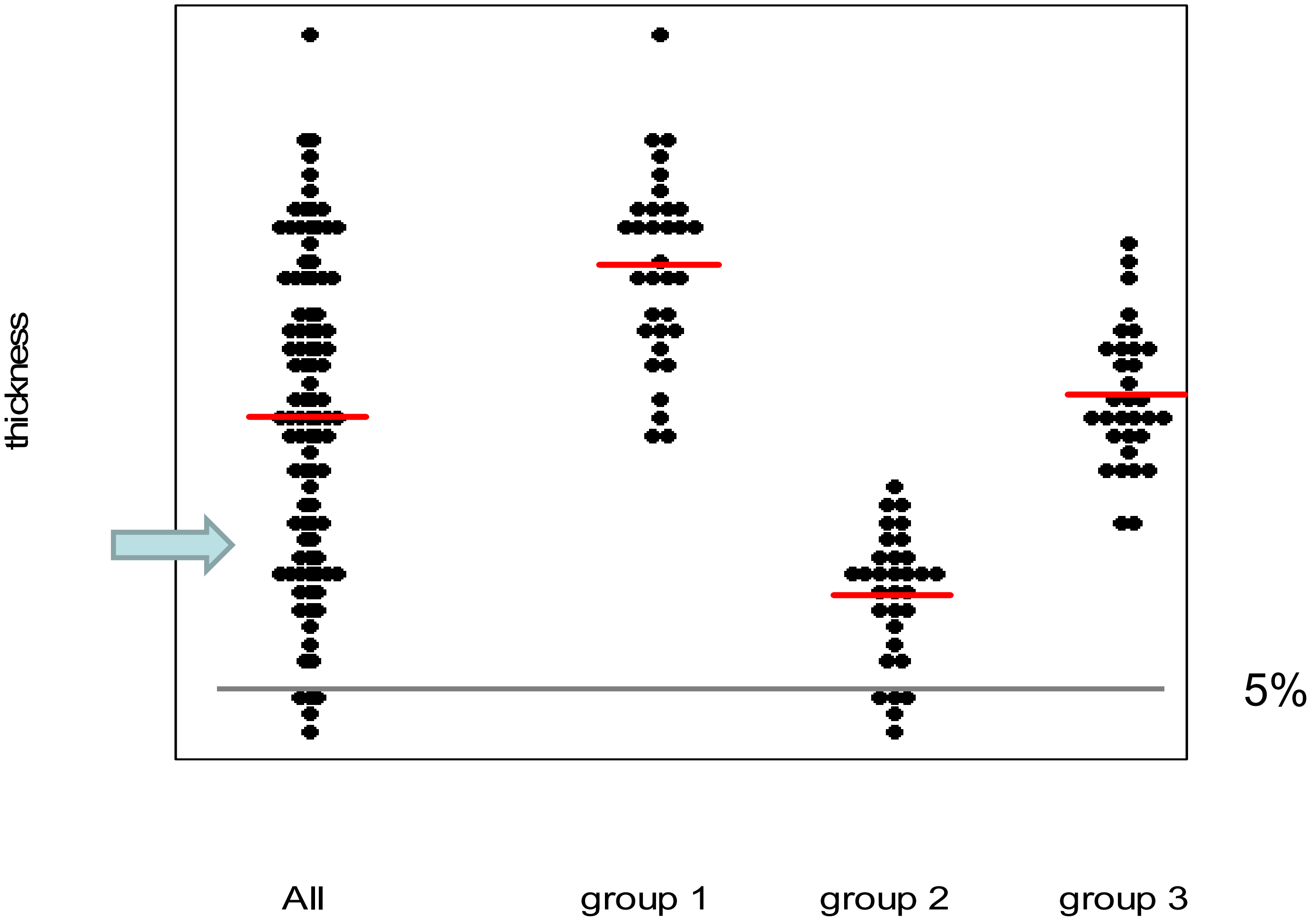
Example: Subgroup Heterogeneity

Figure. Hypothetical dot plots of thickness measurements for an entire cohort and stratified by subgroup; red line indicates mean



Example: Subgroup Heterogeneity

For a subject with a result near blue arrow, whether they are inside or outside NL depends on which subgroup they belong



Covariate-Specific NLs: Factors to Consider

- consider magnitude, direction and clinical significance of subgroup differences or covariate effect
  - assess statistical significance, but don’t use as sole criterion (depends on sample size)
- is covariate identified and established in scientific literature?
- is covariate used in similar devices on market?

FDA Questions 2 and 3

- *What magnitude of difference between subgroups (e.g. in microns or percent difference) is clinically important to warrant covariate-specific normal limits (i.e., stratified databases or an adjustment of the normal limits)?*
- *For which covariates are covariate-specific limits needed?*
- *What range/spectrum of these covariates should be included in a database (e.g., what age range and which race/ethnic groups)?*

Database Construction

- prospectively plan the study
- develop subject selection criteria
  - develop clinical work-up (e.g. perimetry)
  - specify recruitment, inclusion/exclusion criteria
    - recruit subjects from multiple clinical sites
    - use one eye per subject (or adjust data analysis for correlation between eyes within subject)
- specify testing protocols
- record and consider impact of image quality

Database Construction (cont.)

- consider need for covariate-specific NLs
- evaluate distribution plots of results (e.g., histograms, Q-Q plots)
- determine statistical models used
- assess statistical model fit and underlying assumptions (e.g., assess residual plots for lack-of-fit and homogeneous variability)
- calculate 95% confidence intervals for limits (can you distinguish between 1st and 5th percentiles?)

Transferability of NDBs

- Transferring a database from one device model to another requires *interchangeability* of results between models
- It may be possible to transfer a NDB with an appropriate calibration if difference between model results is well characterized
  - Refer to principles in CLSI documents C28-A3 (Section 10) for transference and EP9-A2 for bias estimation
  - Consult with FDA through the pre-submission process

Database/NL Reliability

Estimated normal limits are subject to variation due to different sources:

- characteristics of subjects in NDB
- sampling (sample size) variability
  - characterized by confidence intervals for NLs
- statistical models used
  - different statistical models for estimating covariate-specific NLs can lead to different limits; verifying adequacy of model is critical
- reproducibility (within-model between-device/operator variability)
  - measurements collected using one device/operator may not coincide with measurements collected using a different device/operator of the same model

Database Usability – What to Convey to User

Convey/display information in Labeling (User’s Manual) or software interface so that the user understands

•NDB population characteristics

•how variability in estimation of normal limits affects interpretation of an individuals’ result

•how total variability of measurement (reproducibility) affects interpretation of an individuals’ result

Usability: Characterizing a NDB Population

- clinical definition of “normal”/clinical workup
- recruitment and inclusion/exclusion criteria
- number of subjects, description of sites, subject demographics
- device model and settings, software version
- testing protocol
- how scans were deemed acceptable for use
- limitations of use (including covariates for which the database accounts/does not account)

Examples of Types of Labeling Disclaimers Related to NDBs

- Normative data colors will not appear if the patient is less than X years old. For patients under X years old, the legend and color coding is not displayed. Data was not collected from patients under X years old.
- Results in patients Y years of age or older should be interpreted with caution since only Z subjects were included in the normative database who were Y years of age or older.
- The deviation map shows when a particular region of an eye is thinner than the same region in a population of normal subjects, but such deviation is not always due to pathological loss of RNFL, for any of the following reasons…

Variability in Estimation of NL: Example of Reporting NLs and 95% Confidence Intervals

ONH Parameter	Age (years)	Disc Area	mean	1% NL [95% CI]	5% NL [95% CI]
Rim area	50	1.5			
		2.0			
		2.5			
	70	1.5			
		2.0			
		2.5			

Total Variability of Measurement: Example of Reporting Result with Reproducibility

RNFL Parameters	Result – 2×reproducibility [percentile]	Result in microns [percentile result]	Result + 2×reproducibility [percentile]
Average thickness	72 4%	**80** **12%**	86 20%
Superior thickness	70 <1%	**94** **4%**	118 15%
Etc…			

FDA Question 4

- What does the user need to know about total variability (including within-eye, between-operator, and between-device) of the measurement?

- What does the user need to know about variability in the estimated normal limits (i.e. 95% confidence intervals for percentiles or limits)?

- Should the device report the measurement results as a specific percentile of the normal population in addition to designating into a classification (e.g. >5th percentile)?

Summary of Factors to Consider

- Characteristics of NDB population
- Need for covariate-specific NLs
- Appropriate choice of statistical model for estimating NLs
- Characterize variability of result and impact of variability on comparison of result to NDB

FDA Questions Related to Normative Databases

1. What clinical work-up should be done to establish a “normal” population? Who can be defined as not having disease? What patient characteristics are important to represent in the definition of a normal population?

FDA Questions Related to Normative Databases (cont.)

2. Measurements within the normal population may be heterogeneous, e.g., there may be differences in mean measurements for different age groups, for different race/ethnic groups, or for subjects with different image quality, or for other covariates.
    a. What magnitude of difference between subgroups (e.g. in microns of thickness or percent difference) is clinically important to warrant covariate-specific normal limits (i.e. stratified databases or an adjustment of the normal limits?)
    b. For which covariates are covariate-specific normal limits needed?

FDA Questions Related to Normative Databases (cont.)

3. For significant covariates identified in response to Question 2, how should individuals be selected with respect to the covariates to construct the database? Specifically, what range/spectrum of these covariates should be included in a database (e.g., what age range and which race/ethnic groups)?

FDA Questions Related to Normative Databases (cont.)

4. What information related to the NDB and the comparison of an individual's result to the database should be included in the labeling, device printout, and/or software interface to improve usage of these devices and to facilitate an informed decision about how to apply the output of the device?
   a. What does the user need to know about total variability (including within-eye, between-operator, and between-device) of the measurement?
   b. What does the user need to know about variability in the estimated normal limits (i.e. 95% confidence intervals for normal limits or estimated percentiles)?
   c. Should the device report the measurement results as a specific percentile of the normal population in addition to designating into a classification (e.g. >5th percentile)?

References

- Bland JM and Altman DG. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, **8:**135-160.
- CLSI. Defining, establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline-Third Edition. CLSI document C28-A3. Wayne, PA: Clinical and Laboratory Standards Institute
- CLSI. Method Comparison and Bias Estimation Using Patient Samples: Approved Guideline – Second Edition. CLSI document E09-A2. Wayne, PA: Clinical and Laboratory Standards Institute

References (cont.)

- Harris, E. and Boyd, JC. Statistical Bases of Reference Values in Laboratory Medicine. New York: Marcel Dekker, 1995.
- Hahn, GJ. and Meeker, WQ. Statistical Intervals: A Guide for Practitioners. New York:John Wiley & Sons, 1991.
- Peterson, PH and Henny, J, eds. (2004). Special issue on Reference Values and Reference Intervals. *Clinical Chemistry and Laboratory Medicine*, 42(7):685-876.
- Royston P. (1991). Constructing time-specific reference ranges. *Statistics in Medicine* 10; 675-690.

Thank you!

Extra Slides

Reporting Variability of Measurement

- Reproducibility
  - includes repeatability *plus* among-operator variability *plus* within-model-among-device variability
  - pertains to a *single* measurement

- 95% Reproducibility ‘coefficient’ or limit
  - pertains to the *difference* between repeated measurements [Bland and Altman,1999]
  - calculated as 1.96 ×√2×Reproducibility = 2.77×Reproducibility
  - *differences* between repeated measurements on the same eye with different operators on different devices within a model will be within ± reproducibility limit 95% of the time