A semiparametric Bayesian model for examiner agreement in periodontal research

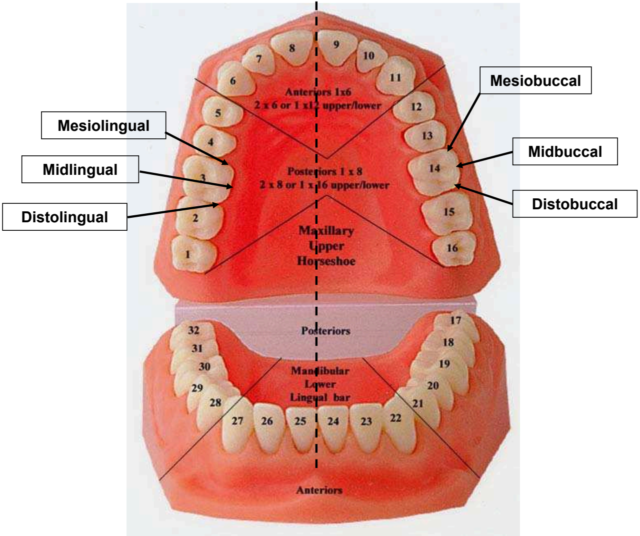
Elizabeth G. Hill and Elizabeth H. Slate

Division of Biostatistics and Epidemiology, MUSC

September 26, 2011

Dental anatomy

From http://dentalimplants-usa.com/generalinfo/toothnumbering.html



Calibrating examiners

- Larger clinical periodontal studies require multiple examiners
- Accurate measurement of PD requires training
- Calibration studies demonstrate degree of agreement among examiners and with a gold standard examiner
- Agreement varies with examiner, but may also depend on characteristics of the site

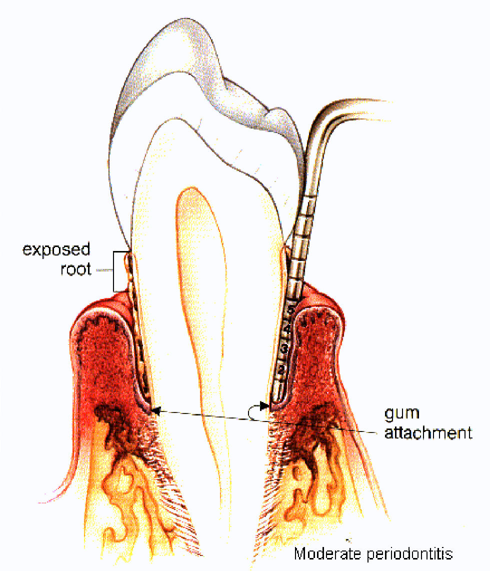
Calibrating examiners

- Larger clinical periodontal studies require multiple examiners
- Accurate measurement of PD requires training
- Calibration studies demonstrate degree of agreement among examiners and with a gold standard examiner
- Agreement varies with examiner, but may also depend on characteristics of the site

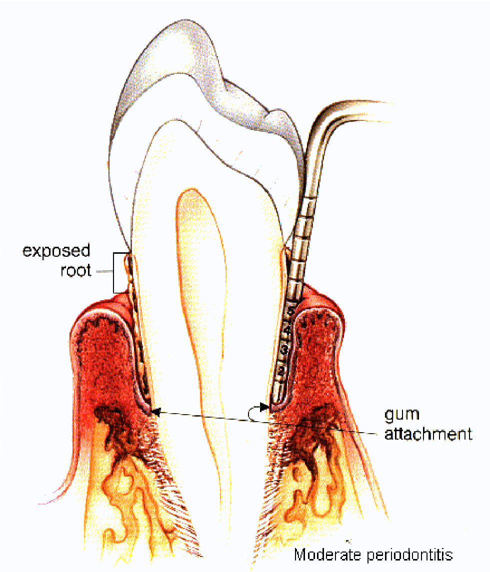
Goals:

- Quantify agreement
- Determine targets for enhanced training

Pocket and probing depth

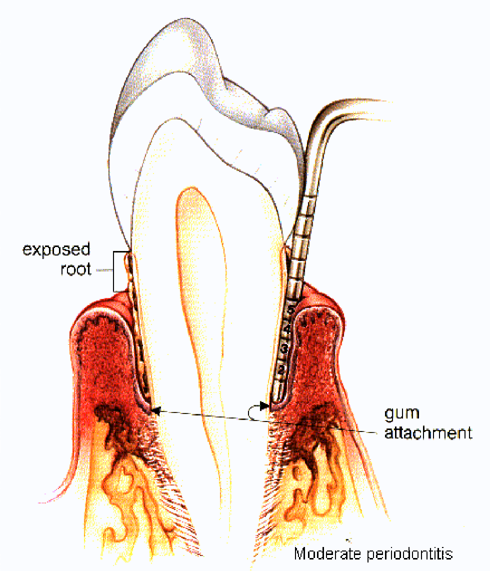


Pocket and probing depth



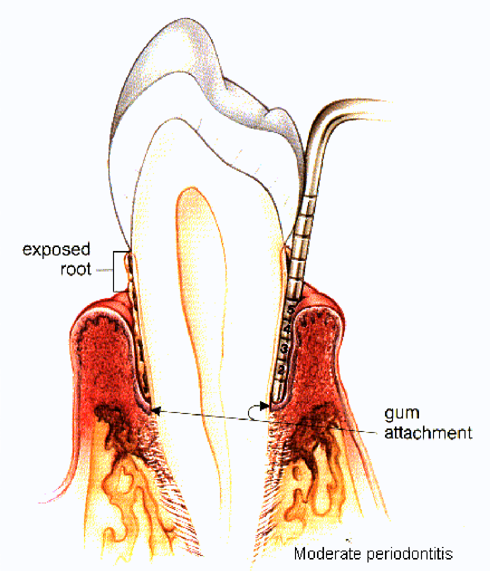
True pocket depth

Pocket and probing depth



True pocket depth
PD = 4.6 mm

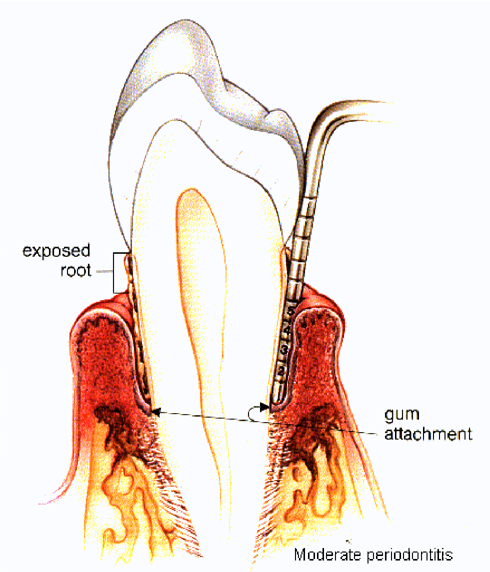
Pocket and probing depth



True pocket depth
PD = 4.6 mm

Observed probed pocket depth

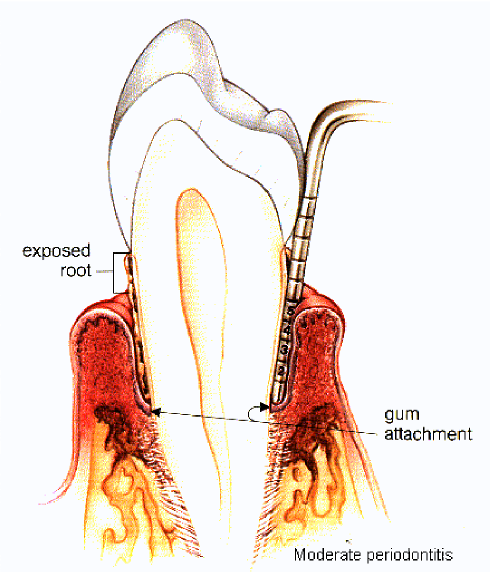
Pocket and probing depth



True pocket depth
PD = 4.6 mm

Observed probed pocket depth
5 mm < Obs PPD < 6 mm

Pocket and probing depth

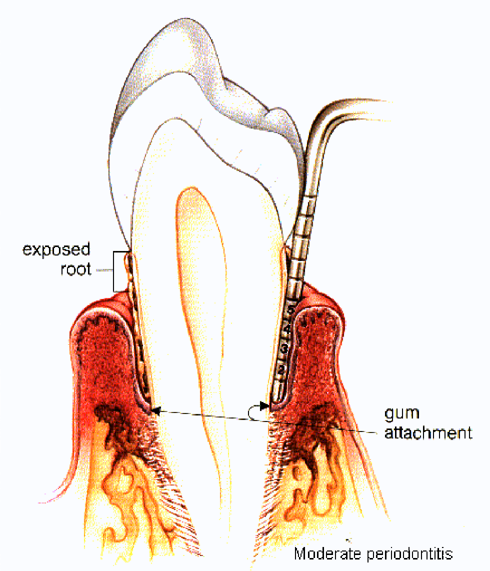


True pocket depth
PD = 4.6 mm

Observed probed pocket depth
5 mm < Obs PPD < 6 mm

Recorded PPD

Pocket and probing depth



True pocket depth
PD = 4.6 mm

Observed probed pocket depth
5 mm < Obs PPD < 6 mm

Recorded PPD
Rec PPD = 5 mm

Examiner calibration pilot study

- Three hygienists (A, B, C) and one standard examiner (S)
- Nine study subjects
- Subjects' quadrants randomized to examiner pairs
- Pocket depth measured at six sites for all available teeth by two examiners at each site
- As many as 168 sites (336 measurements) per subject
- 1080 duplicate observations (2160 measurements) total
- Inter- (AS, BS, CS, AB, AC, BC) and intra-examiner (AA, BB, CC, SS) measurements collected

Percent agreement

Examiner pair	k	n	Exact %	Exact 95% CI	±1 mm %	±1 mm 95% CI
AS	5	180	62	(36,88)	94	(83,100)
BS	5	156	49	(25,73)	88	(71,100)
CS	5	180	43	(34,51)	92	(82,100)
AB	3	108	45	(28,63)	81	(50,100)
AC	3	96	44	(0,89)	88	(64,100)
BC	3	120	47	(13,80)	81	(59,100)
AA	2	60	73	NA	98	(77,100)
BB	2	72	56	(44,67)	94	(36,100)
CC	2	78	79	(59,100)	97	(67,100)
SS	1	30	80	NA	100	NA

Our approach

- Use model-based approach that naturally incorporates dependence among observations
- Use *ALL* data to borrow strength
- Construct reliability measures (κ_w, % agreement) from realizations from posterior predictive distribution

1. *TRUE* pocket depth

$$\log(\theta_{ij}) = \mu + b_i + \varepsilon_{ij}$$

- θ_{ij} is pocket depth for jth site of ith subject
- $b_i|\sigma_b^2 \sim \text{Normal}(0, \sigma_b^2)$
- $\varepsilon_{ij}|\sigma_\varepsilon^2 \sim \text{Normal}(0, \sigma_\varepsilon^2)$
- b_i and ε_{ij} independent

2. *OBSERVED* probed pocket depth

$$\log(O_{ijk}) = \underbrace{\log(\theta_{ij})}_{truth} + \underbrace{\mathbf{X}'_{ijk}\boldsymbol{\beta}_{ij}}_{bias} + \underbrace{\gamma_{ijk}}_{noise}$$

- O_{ijk} is depth observed on probe for kth replicate of jth site of ith subject
- $\mathbf{X}_{ijk} = (X_{A,ijk}, X_{B,ijk}, X_{C,ijk})'$ - examiner indicators
- $\boldsymbol{\beta}_{ij} = (\beta_{A,ij}, \beta_{B,ij}, \beta_{C,ij})'$ - associated parameter vector
- $\gamma_{ijk} \sim \text{Normal}(0, \sigma^2_E)$
- σ^2_E subscripts are A, B, C or S

3. *RECORDED* probed pocket depth

$$T_{ijk} = \begin{cases} \lfloor O_{ijk} \rfloor & \text{if } 0 \leq O_{ijk} < 15 \\ 15 & \text{otherwise} \end{cases}$$

- T_{ijk} is recorded probed pocket depth for kth replicate of jth site for ith subject
- PPD recorded as floor of observed PPD
- Manual probe scored to a maximum depth of 15 mm

3. *RECORDED* probed pocket depth (cont.)

$$\begin{aligned}\pi_{t,ijk} &= \text{Prob}(T_{ijk} = t \,|\, \log(\theta_{ij}), \boldsymbol{\beta}_{ij}, \sigma_E) \\ &= \begin{cases} \zeta_{t+1} & \text{if } t = 0 \\ \zeta_{t+1} - \zeta_t & \text{if } t = 1, \ldots, 14 \\ 1 - \zeta_t & \text{if } t = 15 \end{cases}\end{aligned}$$

- $\zeta_t = \Phi\left(\frac{\log(t) - \log(\theta_{ij}) - \mathbf{X}'_{ijk}\boldsymbol{\beta}_{ij}}{\sigma_E}\right)$
- $\Phi(\cdot)$ is standard normal CDF

Likelihood

Conditional likelihood proportional to

$$\prod_{i=1}^{n} \prod_{j=1}^{m_i} \prod_{k=1}^{2} \prod_{t=0}^{15} \pi_{t,ijk}^{V_{t,ijk}}$$

- $\mathbf{V}_{ijk} = (V_{0,ijk}, V_{1,ijk}, \ldots, V_{15,ijk})'$ is vector of 15 zeros and a single one
- $T_{ijk} = t \Rightarrow V_{t,ijk} = 1$
- $\mathbf{V}_{ijk} \,|\, \log(\theta_{ij}), \boldsymbol{\beta}_{ij}, \sigma_E^2 \sim \text{Multinomial}(1; \pi_{0,ijk}, \pi_{1,ijk}, \ldots, \pi_{15,ijk})$

Prior and hyperprior specifications

- Mean of true distribution (μ) - vague mean zero Normal
- Error terms (five) and random effect - Normal$(0, 1/\sigma_\nu^2)$
- Standard deviations (σ_ν) - Uniform$(0, a)$ (Gelman, Bayesian Analysis 2005)

"Non"parametric Component:

- Rater bias parameters ($\beta_{E,ij}$s) - Dirichlet process prior (DPP)
  - DPP naturally gives rise to clusters
  - Reasonable to assume latent class structure
  - Preliminary analysis indicates site-level characteristics cause examiners to be more (less) prone to bias

Dirichlet process prior overview

- $y_i \sim f(y_i|\phi_i)$, $i = 1, \ldots, n$, with ϕ_i unknown
- ϕ_i is centered around base distribution G_0
- Candidate values for ϕ_i are drawn from G_0 according to concentration parameter α
- Traditional Bayesian approach places prior on ϕ_i
- DPP places prior on ϕ_i's *distribution*

$$\begin{aligned} \phi_i|\Gamma &\sim \Gamma \\ \Gamma &\sim DP(\alpha G_0) \end{aligned}$$

- Assignment of a given candidate value from G_0 to multiple ϕ_i may be expected so that ϕ_is cluster based on similarities among the y_is

Antoniak 1974, Escobar 1995, and Escobar and West 1998

Dirichlet process prior overview (cont.)

Practical approach based on finite implementation (Sethuraman 1994)

- Draw $M \leq n$ candidate values, denoted ϕ_m^* with $m = 1, \ldots, M$, from G_0
- $M^* \leq M$ of these are allocated to one or more of the ϕ_i
- Assignment of ϕ_i to ϕ_m^* determined by a multinomial distribution with probability vector $\mathbf{P} \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_M))$
- Assignment of multiple ϕ_i to the same candidate value ϕ_m^* results in the formation of a cluster
- Empty clusters result when none of the ϕ_i are assigned to one or more of the M candidate values.

Dirichlet process prior overview (cont.)

- ϕ_i's density is discrete and the fineness of the discretization increases with α
    - α large $\Rightarrow$ density of ϕ_i resembles G_0
    - α small $\Rightarrow$ density of ϕ_i similar to a finite mixture model
- Dirichlet Process Mixture Model (DPMM)

DPMM for examiner biases

- For our application

$$\begin{aligned} \beta_{E,ij}|\Gamma_E &\sim \Gamma_E \\ \Gamma_E &\sim DP(\alpha_E G_{E,0}) \end{aligned}$$

- $G_{E,0} \equiv G_0$ vague mean zero Normal
- DPMM identifies latent classes of periodontal sites that "cluster" based on examiner-specific biased rating behavior
- $\alpha_E \equiv 8$ (tried $\alpha = 0.5, 1, 2, \ldots, 10, 20$)
- Total number of classes, $M_E \equiv 6$ (tried $2, 3, 4, 5, 6$)
- M limited by number of distinct data values (Congdon 2001)
  - In our data, recorded PPD ranged from 0mm to 8mm
  - Differences in duplicate recorded PPD ranged from -4mm to 4mm
  - $M_E \leq 9$

Posterior clustering inference

- Least-squares clustering identifies examiner-specific classes of biased ratings (Dahl 2006)
- $\mathbf{c}_{E,1}, \ldots, \mathbf{c}_{E,D}$ are D draws from posterior clustering distribution of $\beta_{E,ij}$s
- For each $\mathbf{c}_E$ in $\mathbf{c}_{E,1}, \ldots, \mathbf{c}_{E,D}$, construct $\mathcal{L} \times \mathcal{L}$ association matrix, $\delta(\mathbf{c}_E)$ ($\mathcal{L} =$ total number of sites)
- $\delta(\mathbf{c}_E)_{\ell\ell'} = 1$ when sites ℓ and ℓ' jointly classified, 0 otherwise
- $\Delta_E =$ element-wise average of collection of association matrices

Examiner E's least-squares cluster

$$\mathbf{c}_E^{\mathrm{LS}} = \underset{\mathbf{c}_E \in \{\mathbf{c}_{E,1},\ldots,\mathbf{c}_{E,D}\}}{\operatorname{argmin}} \sum_{\ell=1}^{\mathcal{L}} \sum_{\ell'=1}^{\mathcal{L}} \left(\delta(\mathbf{c}_E)_{\ell\ell'} - \Delta_{E,\ell\ell'}\right)^2$$

Simulation study

- Comparable in size to real calibration data
  9 subjects, 3 examiners, 1 standard
- Examiners S, A no bias
- Examiner B biased on deep pockets; C biased, severely on DLMM

True pocket depth model

$$\begin{aligned}\log(T_{ijk}) &= \log(\theta_{ij}) + \beta_{B,ij} \cdot I(\theta_{ij} \geq 4\text{mm}) + \beta_{C_1,ij} + \\ &\quad \beta_{C_2,ij} \cdot I(\text{site } j \text{ is distolingual mandibular molar}) + \gamma_{ijk}\,,\end{aligned}$$

- $\beta_{B,ij} = -0.5$
- $\beta_{C_1,ij} = 0.25$
- $\beta_{C_2,ij} = -1$

Simulation results

10K iterations (50,500 burn-in)

Description	Parameter	Truth	Median	95% CI
Mean true PD	μ	1	1.03	$(0.80, 1.18)$
SD random effect	σ_b	0.2	0.19	$(0.11, 0.40)$
SD true PD	σ_ε	0.3	0.29	$(0.28, 0.30)$
SD Ex A obs PPD	σ_A	0.1	0.11	$(0.09, 0.13)$
SD Ex B obs PPD	σ_B	0.25	0.24	$(0.22, 0.28)$
SD Ex C obs PPD	σ_C	0.15	0.15	$(0.12, 0.17)$
SD Ex S obs PPD	σ_S	0.07	0.08	$(0.07, 0.10)$

Estimated $\hat{\kappa}_w$

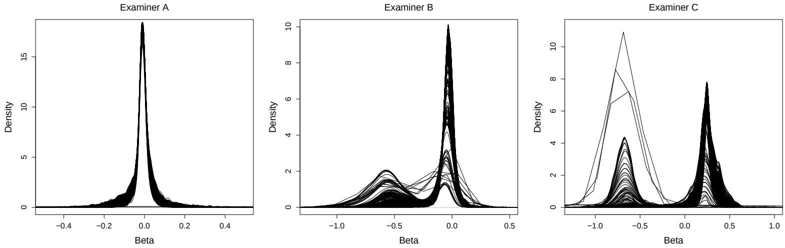
Examiner Pair	Truth	Median	95% CI
AS	0.89	0.85	$(0.77, 0.93)$
BS	0.69	0.67	$(0.55, 0.82)$
CS	0.66	0.61	$(0.50, 0.77)$
AB	0.68	0.63	$(0.47, 0.80)$
AC	0.66	0.59	$(0.45, 0.76)$
BC	0.55	0.50	$(0.33, 0.69)$
AA	0.87	0.84	$(0.73, 0.92)$
BB	0.56	0.62	$(0.44, 0.79)$
CC	0.72	0.82	$(0.71, 0.91)$
SS	0.91	0.87	$(0.78, 0.95)$

Estimated $\hat{\kappa}_W$ - Corrects for bias!

Examiner Pair	Observed	Truth	Median	95% CI
AS	0.90	0.89	0.85	$(0.77, 0.93)$
BS	0.45	0.69	0.67	$(0.55, 0.82)$
CS	0.59	0.66	0.61	$(0.50, 0.77)$
AB	0.43	0.68	0.63	$(0.47, 0.80)$
AC	0.71	0.66	0.59	$(0.45, 0.76)$
BC	0.45	0.55	0.50	$(0.33, 0.69)$
AA	0.83	0.87	0.84	$(0.73, 0.92)$
BB	0.34	0.56	0.62	$(0.44, 0.79)$
CC	0.85	0.72	0.82	$(0.71, 0.91)$
SS	0.88	0.91	0.87	$(0.78, 0.95)$

Guggenmoos-Holzmann and Vonk (SIM 1998) noted this bias in κ

Posterior distributions of examiner effects



- Examiner A - 1 class
- Examiners B and C
  - 1 dominant and 1 subordinate class
  - B's class membership significantly associated with deep pockets ($p < 0.0001$)
  - C's class membership significantly associated with DLMM sites ($p < 0.0001$)

Application to real calibration data

Compared model fits for 4 variants:

- Model 0: No examiner biases, common examiner variances
- Model 1: No examiner biases, unequal variances
- Model 2: Fixed bias for each examiner ($\beta_{E,ij} = \beta_E$), unequal variances
- Model 3: Site-level biases, unequal variances

Application model selection

- Examined posterior predicted distributions for recorded PPD
- Examined DIC_3 values (Celeux et al. 2006)

Model		DIC_3
0:	No bias, common variance	4560.11
1:	No bias, unequal variance	4402.13
2:	Fixed bias, unequal variance	4129.07
3:	Site-level bias, unequal variance	3381.83

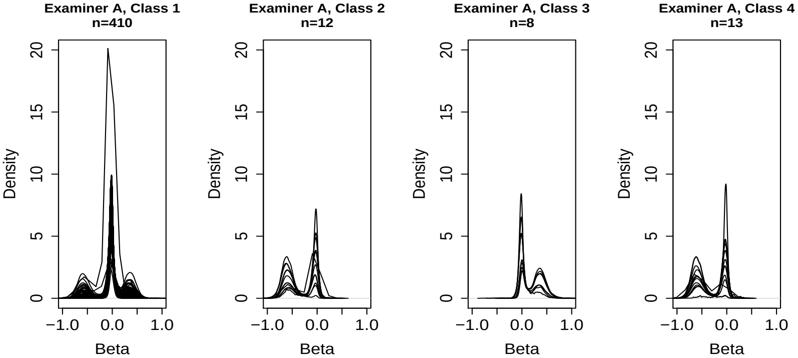
Agreement measures

Examiner	% ±1mm		κ_w	
Pair	Observed	Model	Observed	Model
AS	94 (83, 100)	95 (89, 99)	.71	.80 (.69, .90)
BS	88 (71, 100)	89 (78, 95)	.67	.64 (.49, .80)
CS	92 (82, 100)	92 (83, 97)	.69	.71 (.59,.84)
AB	82 (50, 100)	85 (73, 94)	.63	.60 (.42, .77)
AC	88 (64, 100)	88 (76, 96)	.59	.62 (.44, .79)
BC	81 (59, 100)	88 (77, 96)	.62	.60 (.43, .77)

Agreement measures

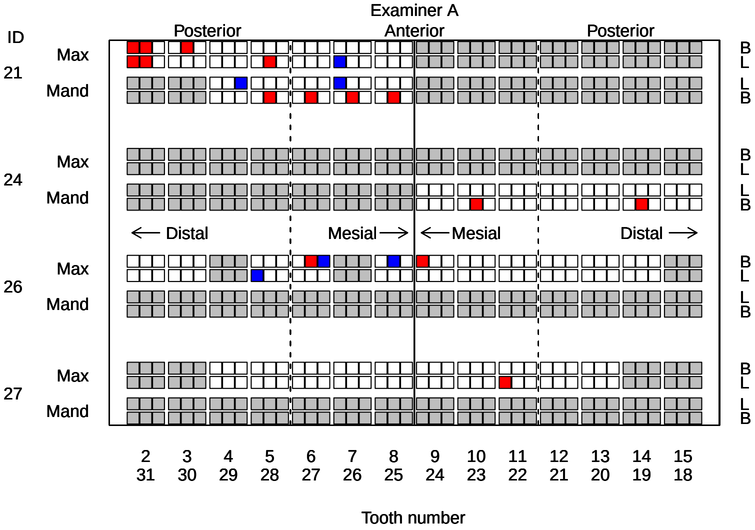
Examiner Pair	% ±1mm Observed	% ±1mm Model	κ_w Observed	κ_w Model
AS	94 (83, 100)	95 (89, 99)	.71	.80 (.69, .90)
BS	88 (71, 100)	89 (78, 95)	.67	.64 (.49, .80)
CS	92 (82, 100)	92 (83, 97)	.69	.71 (.59,.84)
AB	82 (50, 100)	85 (73, 94)	.63	.60 (.42, .77)
AC	88 (64, 100)	88 (76, 96)	.59	.62 (.44, .79)
BC	81 (59, 100)	88 (77, 96)	.62	.60 (.43, .77)
A·Truth		96 (91, 98)		.81 (.73, .90)
B·Truth		91 (83, 96)		.69 (.59, .81)
C·Truth		94 (87, 98)		.74 (.65, .84)
S·Truth		100 (99,100)		.93 (.87, .97)

Examiner A posterior clustering inference



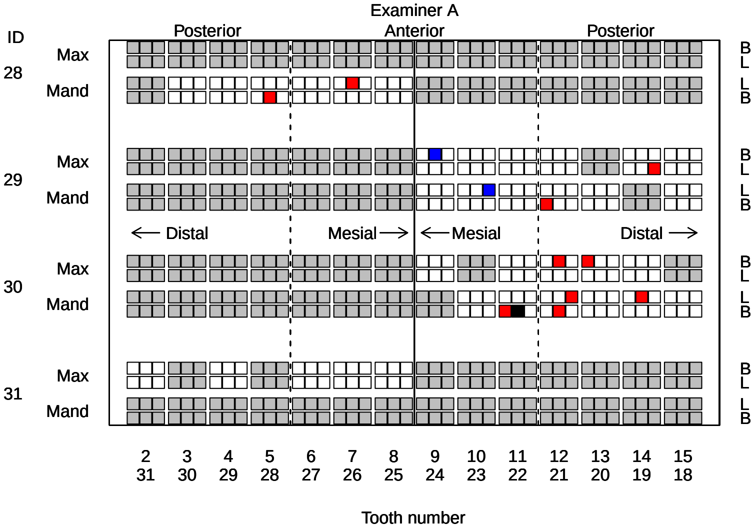
- Class 1 - predominantly unbiased
- Classes 2 and 4 - one class of significantly associated with
  - Negative bias
  - Significantly associated with mid-tooth ($p = 0.03$) and buccal ($p = 0.016$) sites
- Class 3 - positively biased on anterior teeth ($p = 0.028$)

Examiner A



Examiner A
Posterior
Anterior
Posterior
ID
21
24
26
27
Max
Mand
B
L
Distal
Mesial
Mesial
Distal
2 3 4 5 6 7 8 9 10 11 12 13 14 15
31 30 29 28 27 26 25 24 23 22 21 20 19 18
Tooth number

Examiner A


Examiner A
Posterior
Anterior
Posterior
ID
28
29
30
31
Max
Mand
B
L
Distal
Mesial
Mesial
Distal
2 3 4 5 6 7 8 9 10 11 12 13 14 15
31 30 29 28 27 26 25 24 23 22 21 20 19 18
Tooth number

Examiners B and C posterior clustering inference

- Examiner B - 2 classes
  - Dominant class - mildly negatively biased
  - In subordinate class, all recorded PPD = 0mm on mid-tooth sites
  - Possible association with mandibular sites ($p = 0.052$)
  - Significant association with shallow sites ($p = 0.012$)
- Examiner C - 2 classes
  - Dominant class - mildly negative biased
  - Subordinate class - mild positive bias
  - Possible association with anterior sites ($p = 0.052$)
  - Significant association with deeper sites ($p = 0.0007$)

Summary

- Model-based approach naturally incorporates multiple levels of dependence
- Borrowing strength
  - Corrects bias
  - Improves precision
- Accommodates estimation of agreement with truth
- Interpretation of DPMM classes for βs
  - Examiner specific
  - Target follow-up training

Acknowledgements

- Carlos Salinas, DMD and COBRE clinical core
- Sara Grossi, DDS, MS
- Keith Kirkwood, DDS, PhD and Steve London, DDS, PhD
- NCRR P20 RR17696, NIDCR U24 DE16508,
  NIDCR K25 DE016863