MWSUG 2008

INDIANAPOLIS, INDIANA

CROSSROADS OF AMERICA

OCTOBER 12-14

Global Clinical Data Classification: A Discriminate Analysis

Amurthur Ramamurthy, Gordon Kapke and Jodi Yoder
Covance Central Laboratories, Indianapolis, IN

How different is Clinical Laboratory data sets across world geographies and gender ?

- Do global gender specific data sets meet the goals to combine data across geographical regions?
  - Bias Criteria
  - MDA
- Can Multiple Discriminant Analysis distinguish between genders?

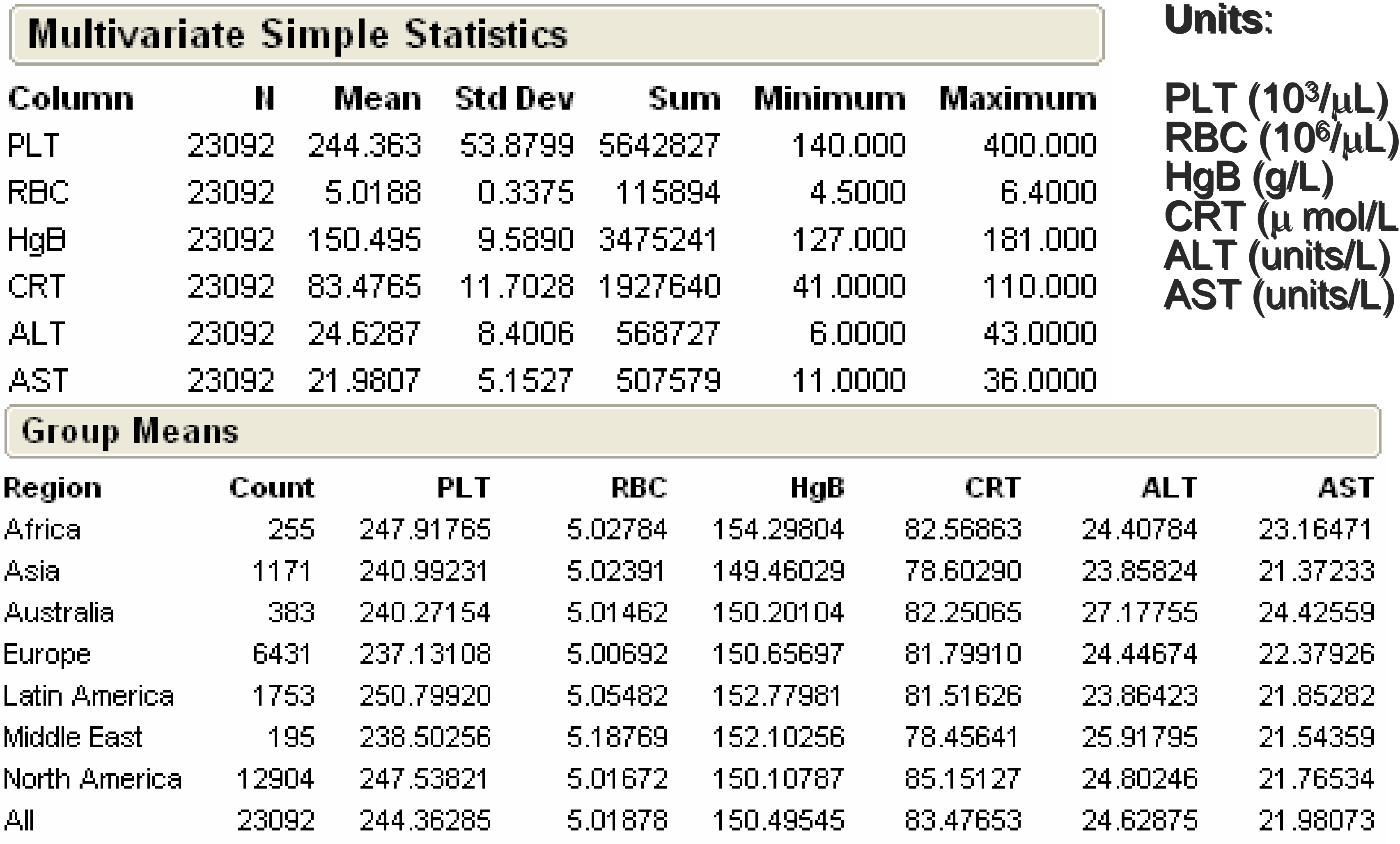
Analytical Data and Multivariate Methods used

- Analytical data
  - Generated with same method
  - Collected across all projects (> 1000)
  - Categorized by age and sex (Adult male and Adult female)
  - Sorted by geography
  - Data truncated using “reference intervals”
- Multivariate Methods
  - Exploratory
    - Principal Component Analysis (PCA)
  - Inferential
    - Multiple Discriminant Analysis (MDA)

Multivariate Hypothesis

- By Gender
  - MDA has good discriminatory power with respect to gender given six exploratory variables
- By Region
  - MDA has poor classification rate within gender across global regions
    - Can we show if there is equivalence among regions

Multivariate Data: For Regions (Adult Male and truncated using reference intervals)



Multivariate Simple Statistics

Column	N	Mean	Std Dev	Sum	Minimum	Maximum
PLT	23092	244.363	53.8799	5642827	140.000	400.000
RBC	23092	5.0188	0.3375	115894	4.5000	6.4000
HgB	23092	150.495	9.5890	3475241	127.000	181.000
CRT	23092	83.4765	11.7028	1927640	41.0000	110.000
ALT	23092	24.6287	8.4006	568727	6.0000	43.0000
AST	23092	21.9807	5.1527	507579	11.0000	36.0000

Group Means

Region	Count	PLT	RBC	HgB	CRT	ALT	AST
Africa	255	247.91765	5.02784	154.29804	82.56863	24.40784	23.16471
Asia	1171	240.99231	5.02391	149.46029	78.60290	23.85824	21.37233
Australia	383	240.27154	5.01462	150.20104	82.25065	27.17755	24.42559
Europe	6431	237.13108	5.00692	150.65697	81.79910	24.44674	22.37926
Latin America	1753	250.79920	5.05482	152.77981	81.51626	23.86423	21.85282
Middle East	195	238.50256	5.18769	152.10256	78.45641	25.91795	21.54359
North America	12904	247.53821	5.01672	150.10787	85.15127	24.80246	21.76534
All	23092	244.36285	5.01878	150.49545	83.47653	24.62875	21.98073

Units:

PLT ($10^3/\mu L$)
RBC ($10^6/\mu L$)
HgB (g/L)
CRT (μ mol/L)
ALT (units/L)
AST (units/L)

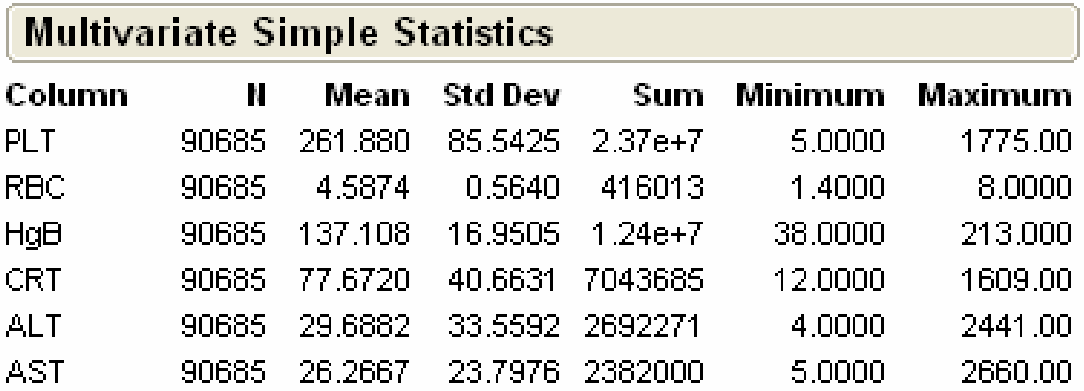
What is the minimal acceptable bias to combine data with a single reference interval?

$$\text{Bias} < 0.375\,(CV_i^{\,2} + CV_g^{\,2})^{1/2}$$

CV: Coefficient of Variation

Analyte	Minimal Acceptable Range	Regions within Range
PLT	± 8.9 %	All
RBC	± 2.6 %	All
HgB	± 2.79 %	All
CRT	± 5.10 %	All
ALT	± 18.0 %	All
AST	± 8.10 %	All

Multivariate Data: For Gender (Adult Male and Female data not truncated)

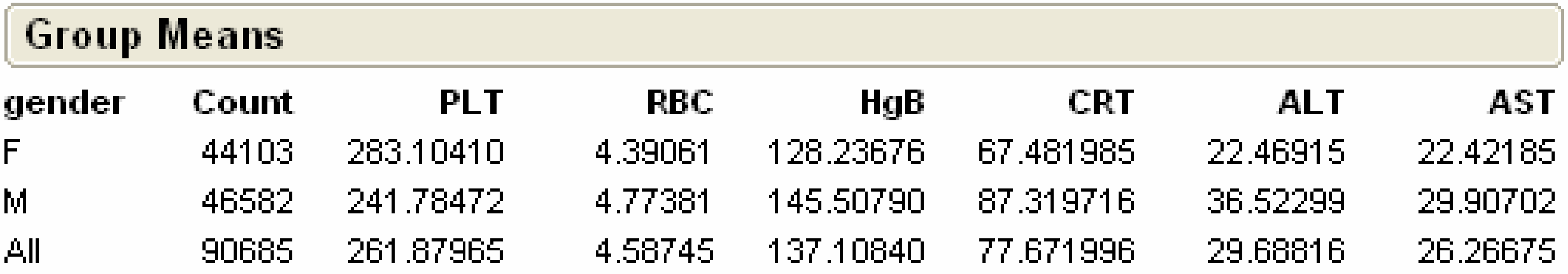


Multivariate Simple Statistics

Column	N	Mean	Std Dev	Sum	Minimum	Maximum
PLT	90685	261.880	85.5425	2.37e+7	5.0000	1775.00
RBC	90685	4.5874	0.5640	416013	1.4000	8.0000
HgB	90685	137.108	16.9505	1.24e+7	38.0000	213.000
CRT	90685	77.6720	40.6631	7043685	12.0000	1609.00
ALT	90685	29.6882	33.5592	2692271	4.0000	2441.00
AST	90685	26.2667	23.7976	2382000	5.0000	2660.00

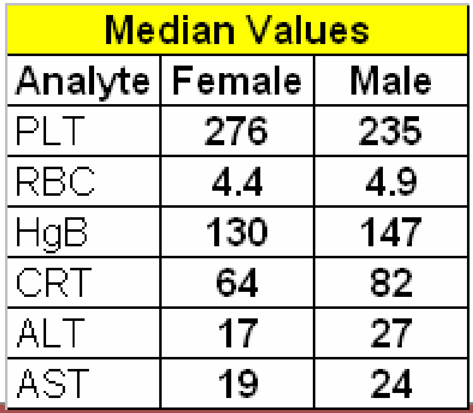
Units:

PLT (10^3/μL)
RBC (10^6/μL)
HgB (g/L)
CRT (μ mol/L)
ALT (units/L)
AST (units/L)



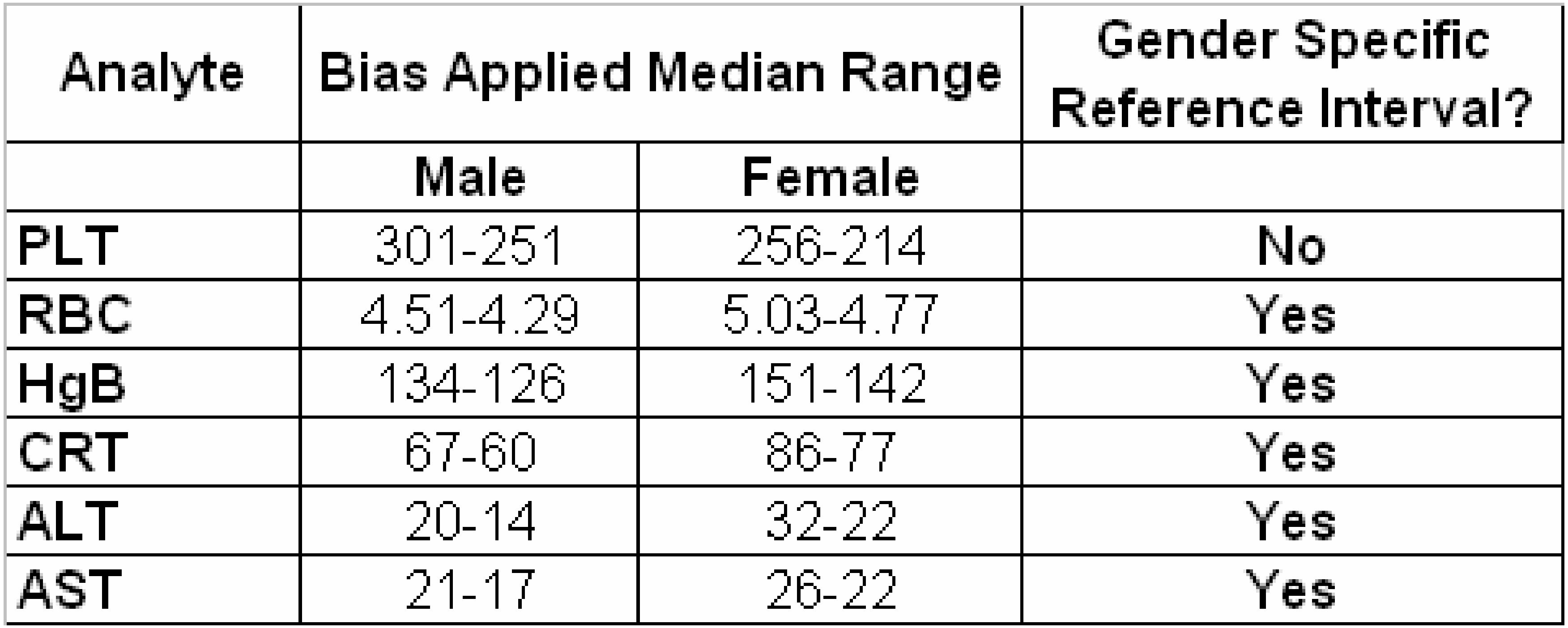
Group Means

gender	Count	PLT	RBC	HgB	CRT	ALT	AST
F	44103	283.10410	4.39061	128.23676	67.481985	22.46915	22.42185
M	46582	241.78472	4.77381	145.50790	87.319716	36.52299	29.90702
All	90685	261.87965	4.58745	137.10840	77.671996	29.68816	26.26675



Median Values		
Analyte	Female	Male
PLT	276	235
RBC	4.4	4.9
HgB	130	147
CRT	64	82
ALT	17	27
AST	19	24

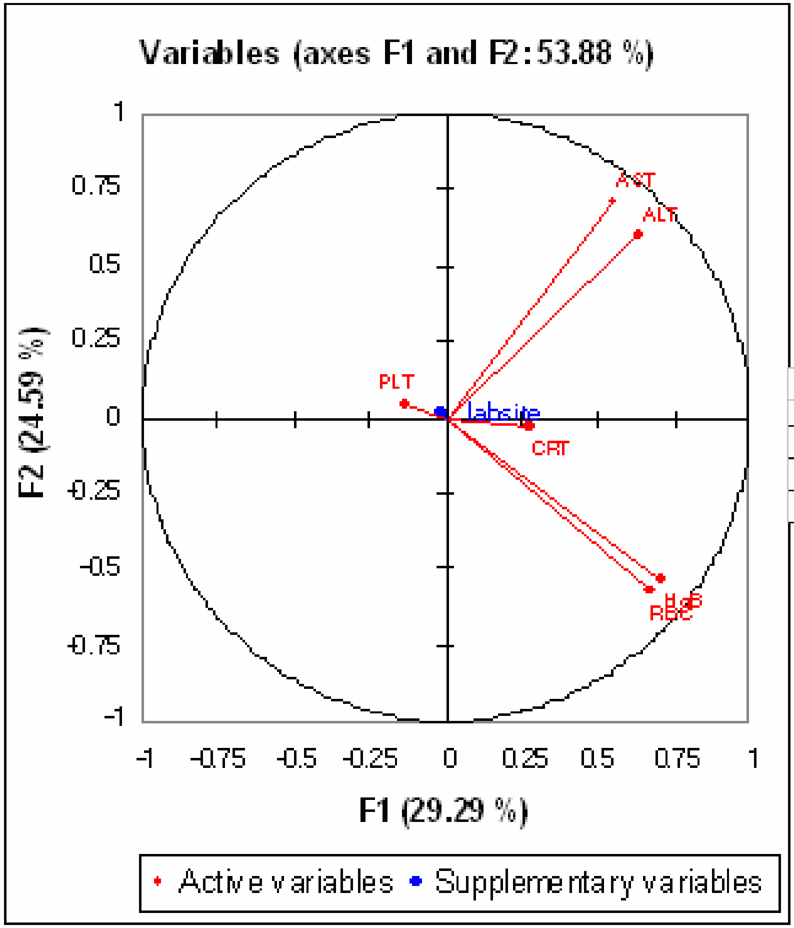
Are Gender Specific Reference Intervals Justified on Bias Criteria Applied to Median Values?

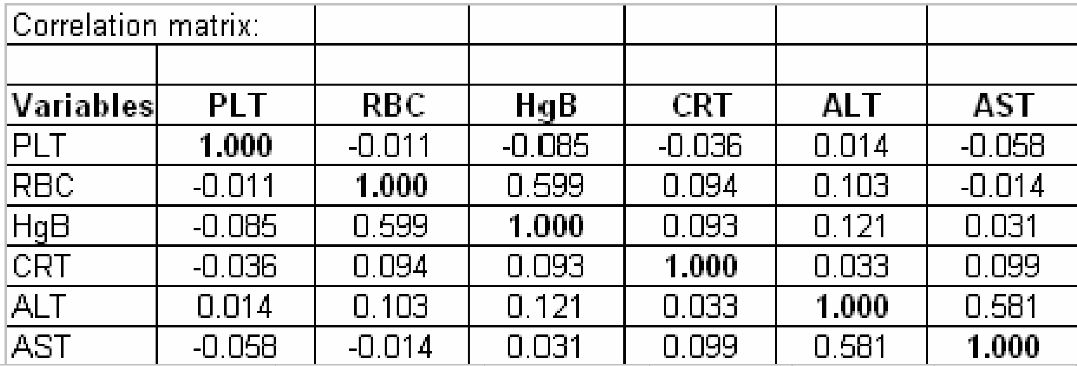


Analyte	Bias Applied Median Range		Gender Specific Reference Interval?
	Male	Female	
PLT	301-251	256-214	No
RBC	4.51-4.29	5.03-4.77	Yes
HgB	134-126	151-142	Yes
CRT	67-60	86-77	Yes
ALT	20-14	32-22	Yes
AST	21-17	26-22	Yes

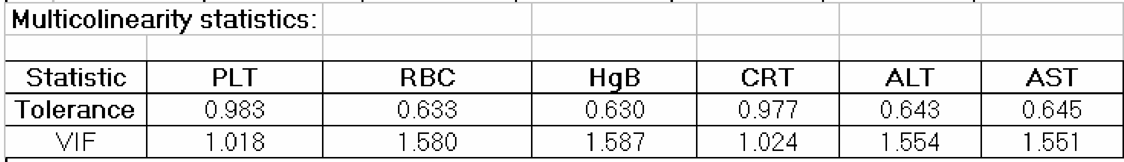
Gender Specific Reference Intervals are not Justified when ranges overlap

Principal Component Analysis: Regions





Correlation matrix:						
Variables	**PLT**	**RBC**	**HgB**	**CRT**	**ALT**	**AST**
PLT	**1.000**	-0.011	-0.085	-0.036	0.014	-0.058
RBC	-0.011	**1.000**	0.599	0.094	0.103	-0.014
HgB	-0.085	0.599	**1.000**	0.093	0.121	0.031
CRT	-0.036	0.094	0.093	**1.000**	0.033	0.099
ALT	0.014	0.103	0.121	0.033	**1.000**	0.581
AST	-0.058	-0.014	0.031	0.099	0.581	**1.000**



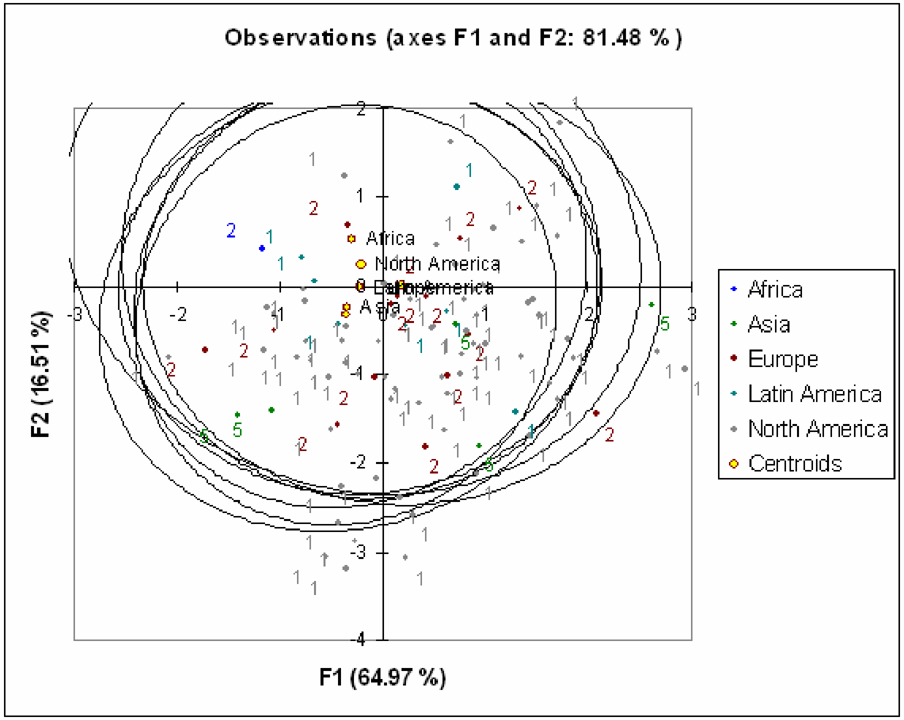
Multicolinearity statistics:						
Statistic	PLT	RBC	HgB	CRT	ALT	AST
Tolerance	0.983	0.633	0.630	0.977	0.643	0.645
VIF	1.018	1.580	1.587	1.024	1.554	1.551

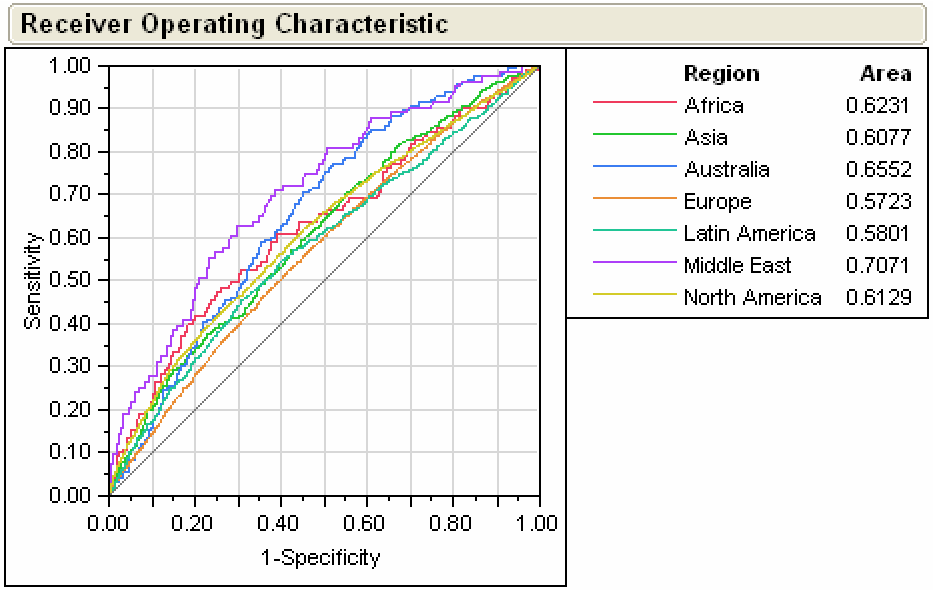
- First two PCA's contribute over 50 % of the observed variation.
- Correlations are seen between RGB/HgB and ALT/AST

 Levels of Correlations/VIF's are not a concern for MDA

Adult Male Data: Truncated with reference intervals

Multiple Discriminant Analysis: For Regions





Region	Area
Africa	0.6231
Asia	0.6077
Australia	0.6552
Europe	0.5723
Latin America	0.5801
Middle East	0.7071
North America	0.6129

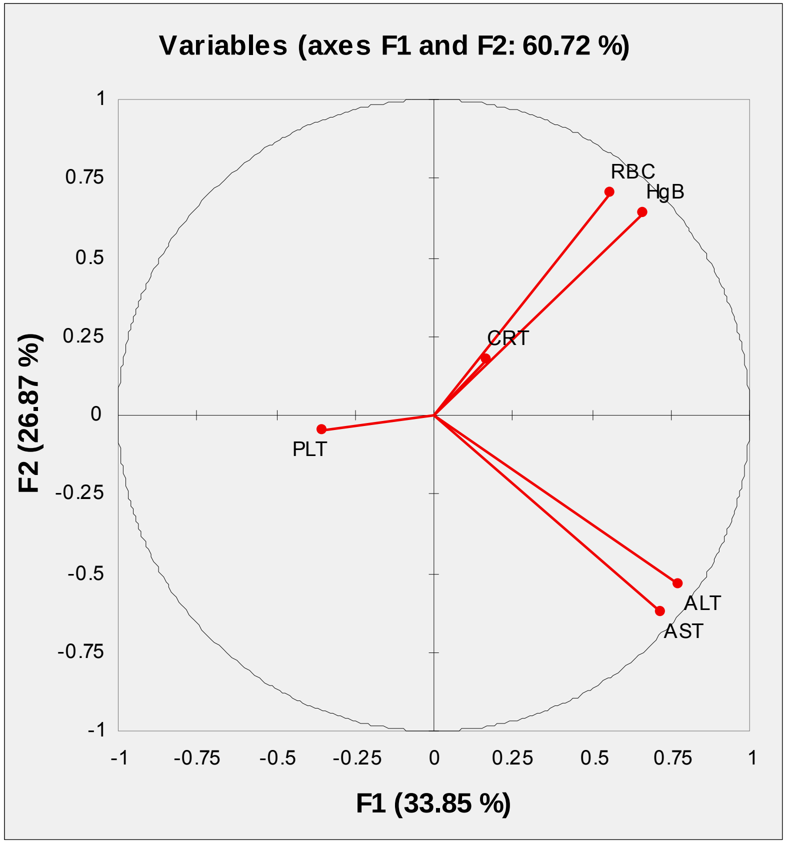
Adult Male Data: Truncated with reference intervals

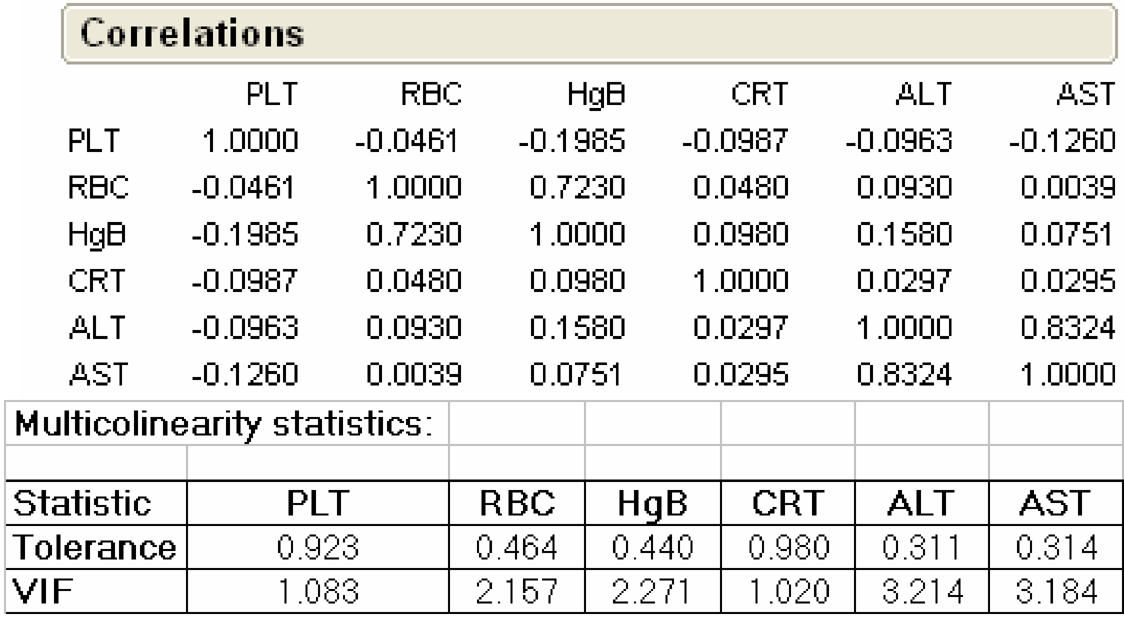
Misclassification Rate : 80 %

Note all six Analytes are Included in this Analysis

Note all Centroids Are with the 50 % Confidence circles

Principal Component Analysis: Gender





Correlations

	PLT	RBC	HgB	CRT	ALT	AST
PLT	1.0000	-0.0461	-0.1985	-0.0987	-0.0963	-0.1260
RBC	-0.0461	1.0000	0.7230	0.0480	0.0930	0.0039
HgB	-0.1985	0.7230	1.0000	0.0980	0.1580	0.0751
CRT	-0.0987	0.0480	0.0980	1.0000	0.0297	0.0295
ALT	-0.0963	0.0930	0.1580	0.0297	1.0000	0.8324
AST	-0.1260	0.0039	0.0751	0.0295	0.8324	1.0000

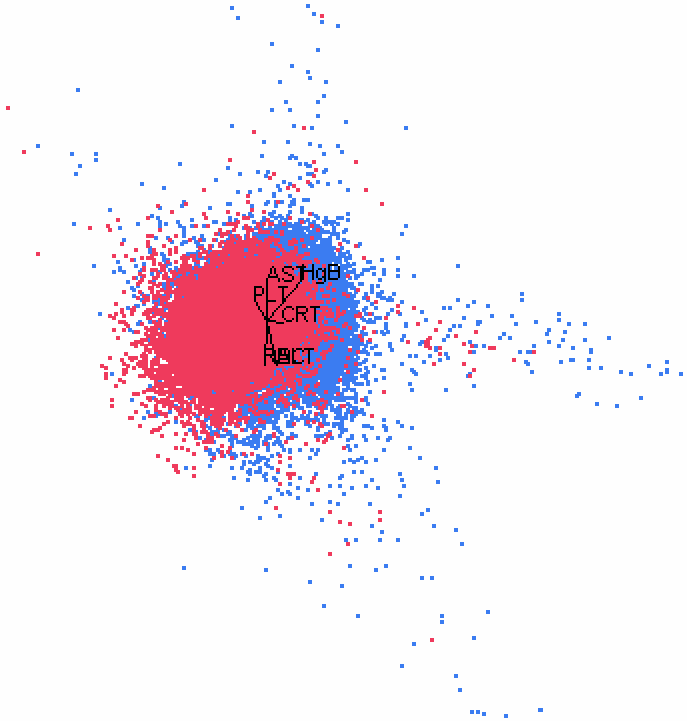
Multicolinearity statistics:

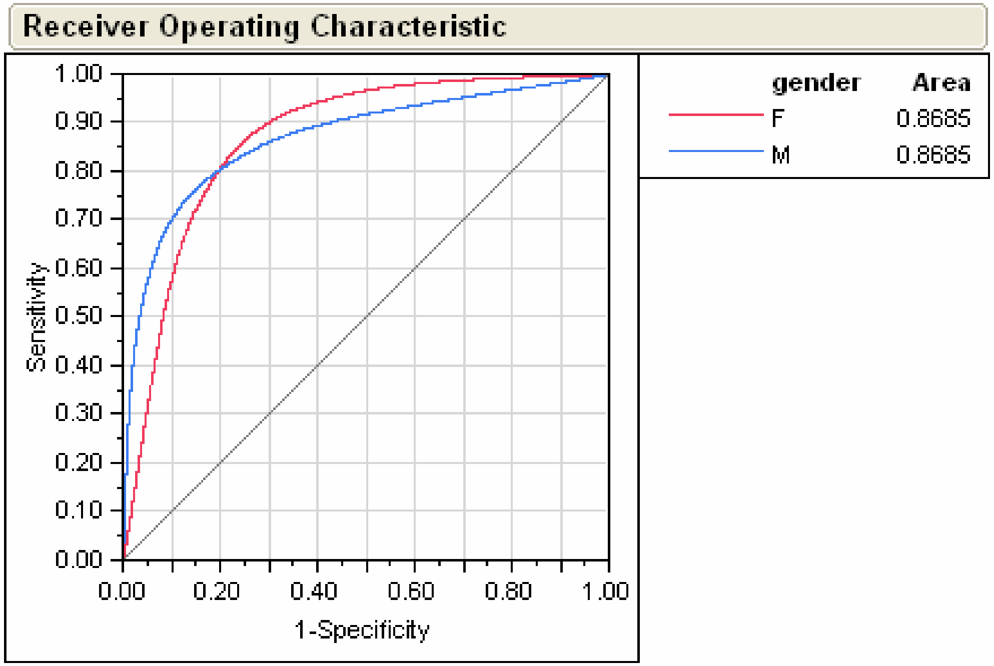
Statistic	PLT	RBC	HgB	CRT	ALT	AST
Tolerance	0.923	0.464	0.440	0.980	0.311	0.314
VIF	1.083	2.157	2.271	1.020	3.214	3.184

- First two PCA’s contribute over 60 % of the observed variation.
- Correlations are seen between RGB/HgB and ALT/AST

Levels of Correlations/VIF’s are not a concern for MDA

Multiple Discriminant Analysis: For Gender




Receiver Operating Characteristic

gender	Area
F	0.8685
M	0.8685

Misclassification rate: 20 %

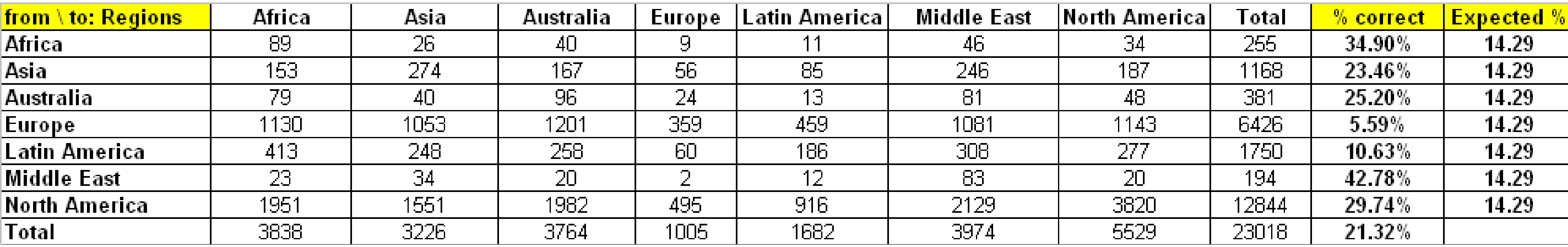
Note all six
Analytes are
Included in this
Analysis

Cross Validation

- **Leave-one-out cross-validation**
  - involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

Cross Validation Results: Adult male data truncated by reference intervals

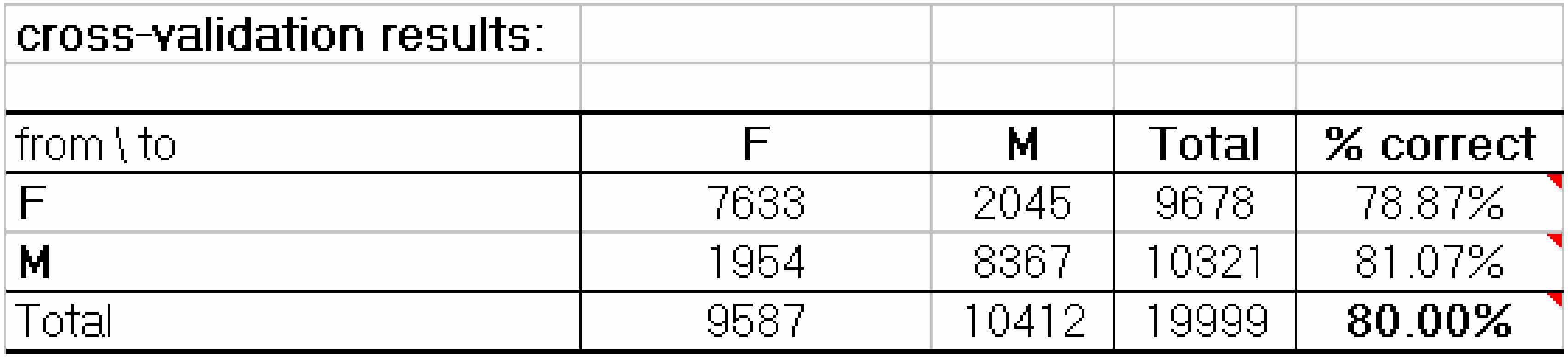
- Prior probabilities were not assumed in this analysis
- Each region has a 14.3 % chance
- Average misclassification is 79 %
- This data set showcases the limitation of MDA (unequal distributions) – Journal of Finance, XXXII(1977)875



from \ to: Regions	Africa	Asia	Australia	Europe	Latin America	Middle East	North America	Total	% correct	Expected %
Africa	89	26	40	9	11	46	34	255	34.90%	14.29
Asia	153	274	167	56	85	246	187	1168	23.46%	14.29
Australia	79	40	96	24	13	81	48	381	25.20%	14.29
Europe	1130	1053	1201	359	459	1081	1143	6426	5.59%	14.29
Latin America	413	248	258	60	186	308	277	1750	10.63%	14.29
Middle East	23	34	20	2	12	83	20	194	42.78%	14.29
North America	1951	1551	1982	495	916	2129	3820	12844	29.74%	14.29
Total	3838	3226	3764	1005	1682	3974	5529	23018	21.32%	

Cross Validation Results: Gender data not truncated by reference intervals

- Prior probabilities were not assumed in this analysis
- Each gender assumed to have 50 % chance
- Average misclassification is 20 %



cross-validation results:				
from \ to	F	M	Total	% correct
F	7633	2045	9678	78.87%
M	1954	8367	10321	81.07%
Total	9587	10412	19999	**80.00%**

Summary

- Global data bias (measured by AON or medians) suggests global data can be combined with a single reference interval
- MDA of global data shows poor classification rates within adult males consistent with combining data globally
- MDA has a very good discrimination among adult Male and Female populations.