PROFICIENCY TESTING IN CLINICAL CHEMISTRY

by

JOSEPH RONALD KELLY

B.Sc., The University of British Columbia, 1994

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as confirming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April 1996

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date April 23, 1996

Abstract

In this thesis we will investigate proficiency testing in clinical chemistry. External quality control agencies utilize proficiency testing programs in order to assess and improve the analytical performance of clinical laboratories. In particular, we will consider the **CHEM^{PLUS}** proficiency testing program which has been developed by the Canadian Reference Laboratory (C.R.L.) The **CHEM^{PLUS}** program is designed to assess the validity of clinical tests performed by a laboratory, that is, the laboratory's analytical performance. Statistically valid and medically relevant evaluation criteria should be used to assess laboratory performance. However, there is debate whether this requirement is fulfilled by the use of current evaluation methodology. We will use data from the **CHEM^{PLUS}** proficiency testing program to compare several standard methods of evaluating laboratory performance. In addition, we will construct an overall measure of laboratory performance that can be used to rank clinical laboratories.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Dr. Harry Joe, for his encouragement and assistance while working on this thesis, and Dr. Michael Schulzer for his comments and suggestions for improving this manuscript. I would also like to thank the staff of the Canadian Reference Laboratory. In particular, thanks to Cathy McGuinness and Pam Cribbs for their limitless support and friendship. As well, thanks to Dr. David Seccombe for providing me with the opportunity to work with the Canadian Reference Laboratory. I would also like to thank the National Research Council of Canada for funding my work through the Industrial Research Assistance Program.

Finally, I would like to take this opportunity to especially thank my family and friends for believing in me and supporting me throughout my life. Without you I would have accomplished nothing. Thank you.

Chapter 1

Introduction

1.1 Proficiency Testing

The clinical laboratory community uses proficiency testing programs as a means of assessing and improving analytical performance [1, 2, 3]. Proficiency testing programs are developed and delivered by external quality control agencies. The Canadian Reference Laboratory (C.R.L.) is one such agency. The C.R.L. provides external quality control for laboratories across Canada. Moreover, all laboratories in British Columbia and Alberta are required by provincial law to participate in the C.R.L. program. In this regard, the C.R.L. also helps regulate clinical laboratories.

A typical clinical laboratory performs hundreds, perhaps thousands, of tests every day. In this context, a clinical test is a concentration measurement of an analyte from a serum sample. The serum sample is collected from a patient, or subject. The results of these tests can be used for diagnosing a disease, determining the severity of a disease, monitoring the progress of a disease, monitoring therapy, monitoring drug toxicity, predicting the response to a treatment, predicting a prognosis, or reassuring a patient [4]. Every time a laboratory performs a test, laboratory error can adversely affect the result. It is beneficial to clinical laboratories, and society, to minimize this laboratory error.

Laboratory error can be divided into three components. These components are preanalytical, analytical, and postanalytical error. The preanalytical component of laboratory error affects the test result before the concentration measurement is actually performed. This component can be divided into in vivo and in vitro factors. In vivo, or biological, factors act in the subject before and during serum sample collection. In vivo factors can be uncontrollable (for example, the age or race of the patient) or controllable (for example, the eating or exercise habits of the patient). In vitro, or environmental, factors can act on the subject or the serum sample. In vitro factors act on the subject before and during serum sample collection (for example, the preparation of the patient or the time and temperature during serum sample collection). In vitro factors act on the serum sample before the concentration measurement is actually performed (for example, the handling of the serum sample after collection or the preparation of the serum sample for the test). The analytical component of laboratory error affects the test result during the concentration measurement. Proficiency testing programs are developed to monitor analytical error. Clinical laboratories also use internal quality control to assess analytical error. The postanalytical component of laboratory error affects the test result after the concentration measurement is performed (for example, the recording of test results, the transmission of test results or the interpretation of test results).

The total laboratory error is mainly due to preanalytical in vivo factors (that is, biological factors) and analytical factors. Also, preanalytical in vitro factors partially contribute to the total laboratory error. However, postanalytical error seems to only result in a delay or restart of the testing process. In fact, the final observed value of a clinical test has basically no postanalytical

error [5]. Therefore, the validity of clinical tests can be significantly improved by reducing the analytical error, hence the importance of proficiency testing programs.

Proficiency testing programs are designed to assess the validity of clinical tests performed by a laboratory, that is, the laboratory's performance. This performance assessment should be comparable between laboratories and consistent over time and location. Although proficiency testing programs are all moderately different, the basic procedure underlying the programs is similar. External quality control agencies prepare common serum samples, also referred to as quality control samples, that are sent to laboratories participating in the agency's proficiency testing program. The participating laboratories measure concentrations of various analytes from the samples. The laboratories then report the results of the measurements back to the external quality control agency. The agency then compares and evaluates the laboratories' results. Laboratories participating in proficiency testing programs on a long term basis, while satisfying regulatory requirements, will attempt to improve the quality of their clinical test results.

1.2 The C.R.L. and CHEM^{PLUS}

The C.R.L. has developed several proficiency testing programs, each designed for different purposes. These programs are summarized in table 1.1. These proficiency testing programs share several ideal characteristics. For example, the quality control samples prepared by the C.R.L. are fresh human serum. As well, the quality control samples cover a clinically relevant range of analyte concentration levels, that is, the samples are representative of both clinically

normal and abnormal individuals. Also, all the programs report results quickly, allowing analytical problems to be identified promptly. Finally, the programs provide educational services which assist participating laboratories in improving their performance.

Table 1.1: The C.R.L. Proficiency Testing Programs	
CHEM^{PLUS}	A proficiency testing program for general chemistry
CHEM^{JR}	A full service proficiency testing program for small laboratories
LIPID^{PLUS}	A proficiency testing program for lipids (for example, cholesterol or triglycerides)
NeoBil	A proficiency testing program for neonatal bilirubin
Glycated Hemoglobin	A proficiency testing program for HbA_{1c}

Although all the C.R.L. programs are important, here, the **CHEM^{PLUS}** program will be the focus. The **CHEM^{PLUS}** program assesses laboratory performance for general chemistry. General chemistry can be divided into three segments; general analytes (for example, creatinine or glucose), electrolytes which are charged analytes (for example, calcium or sodium), and enzymes (for example, aspartate transaminase or amylase). An enzyme is a catalyst in a chemical reaction, while an analyte is a participant in a chemical reaction. Although differences exist, for simplicity, the term analyte will be used interchangeably to refer to either an enzyme or an analyte (general or electrolyte). That is, an analyte will be defined as the desired component of a serum sample whose concentration is being measured in a clinical test.

A clinical laboratory participating in the **CHEM**$^{\textbf{PLUS}}$ program receives six bimonthly shipments of five serum samples. Each sample, also referred to as a vial, consists of 3.0 mL of fresh human serum. Participating laboratories measure analyte concentrations from each vial. Although participating laboratories are aware that the vials are **CHEM**$^{\textbf{PLUS}}$ quality control samples, ideally the vials should be treated as actual patient samples. However, some laboratories treat the vials with extra care and attention [6], hence, concentration measurements from the vials do not represent concentration measurements from actual patient samples. Note that this problem could be solved by "blinding" the participating laboratories, that is, the laboratories would not be aware that the vials are **CHEM**$^{\textbf{PLUS}}$ quality control samples.

For each **CHEM**$^{\textbf{PLUS}}$ shipment, the five vials are adjusted so that the analyte concentrations vary over both a normal and an abnormal clinical range. The five vials are randomly color coded as blue, green, yellow, orange or red. The analytes available for measurement in the **CHEM**$^{\textbf{PLUS}}$ program are listed in table 1.2. Participating laboratories do not necessarily perform concentration measurements for all the analytes listed in table 1.2.

Table 1.2: CHEM^{PLUS} Analytes		
Albumin	Iron	Urea
Total CO_2	Total Iron Binding Capacity (TIBC)	Uric Acid
Total Bilirubin	Lithium	Alanine Transaminase (ALT)
Direct Bilirubin	Magnesium	Alkaline Phosphatase (ALP)
Calcium	Osmolality	Amylase
Ionized Calcium	Phosphate	Aspartate Transaminase (AST)
Chloride	Potassium	Creatine Kinase (CK)
Creatinine	Total Protein	Gamma Glutamyl Transferase (GGT)
Glucose	Sodium	Lactate Dehydrogenase (LDH)

The C.R.L. assesses the validity of analyte concentration measurements observed by participating laboratories. Observed concentration measurements should be compared to statistically valid and medically relevant evaluation criteria. However, there is debate whether this requirement is fulfilled by the use of the current evaluation methodology employed by the C.R.L. and other external quality control agencies.

1.2.1 The Acceptable Concentration Interval

Laboratories participating in the **CHEM^{PLUS}** program measure analyte concentrations from each of the five quality control samples prepared by the C.R.L. Each laboratory's observed

concentration measurements are compared to a statistically determined interval of acceptable concentrations. This interval can be expressed as follows:

$$Acceptable\ Concentration\ Interval = Target \pm Limit \quad (1.1)$$

If an observed concentration measurement lies within this interval, the observed concentration is "acceptable." Similarly, if an observed concentration measurement lies outside of this interval, the observed concentration is "unacceptable." Therefore, the C.R.L. uses the acceptable concentration interval to evaluate analyte concentration measurements observed by participating laboratories. However, different definitions of the acceptable concentration interval can be implemented, depending on the methodology used to define the target and the limit. In Chapter 2, the problems associated with defining the target are addressed and in Chapter 3, the limit is considered.

1.2.2 Analytical Performance

The C.R.L. uses the **CHEM**$^{\textbf{PLUS}}$ program to assess the analytical performance of participating laboratories. Accuracy, precision and linearity are the three main elements of laboratory performance. Accuracy is the ability of a laboratory to measure the true concentration of an analyte during a clinical test. Inaccuracy, or bias, is the lack of accuracy. Precision is the ability of a laboratory to consistently measure similar concentrations during repeated clinical tests. Imprecision is the lack of precision. Inaccuracy and imprecision are the two components of

analytical error. Inaccuracy is the component of analytical error due to systematic variation, and imprecision is the component of analytical error due to random variation.

Linearity is also an important element of laboratory performance. For each analyte, laboratories should aim to perform concentration measurements with low analytical error (i.e. inaccuracy and imprecision). If possible, low analytical error should be achieved at all possible concentration levels of the analyte. Linearity is the ability of a laboratory to perform concentration measurements, with low analytical error, over the entire concentration range of an analyte. For some analytes, however, linearity may be an impossibility.

An ideal proficiency testing program should incorporate accuracy, precision and linearity when evaluating participating laboratories. The **CHEM**$^{\textbf{PLUS}}$ program uses the acceptable concentration interval to help evaluate the analytical performance of participating laboratories. Although the current methodology used to define the target and the limit may not be adequate, a properly constructed acceptable concentration interval can be employed which accurately assesses a laboratory's bias and imprecision. As well, a properly defined target and limit can be used to construct a measure of overall laboratory performance that incorporates linearity as well as accuracy and precision. In Chapter 4, different measures of overall laboratory performance are proposed.

1.3 Reference Intervals

An important characteristic of the **CHEM**$^{\textbf{PLUS}}$ program is that, for each shipment, the five vials are adjusted so that the analyte concentrations vary over both normal and abnormal clinical levels. This implies that normal concentration values and abnormal concentration values are known for each analyte. Reference intervals are used to determine these normal and abnormal concentration values [7, 8].

Reference intervals are based on a reference population. A reference population consists of all valid reference individuals. A valid reference individual must satisfy certain criteria. These criteria can be categorized as either inclusion or exclusion criteria. Inclusion criteria are used to include individuals as valid reference individuals (for example, include females aged 30 to 40 years). Exclusion criteria are used to exclude individuals as valid reference individuals (for example, exclude pregnant females). Inclusion and exclusion criteria are used to create a reference population based on reference individuals who share many common characteristics. Note that reference individuals are not necessarily healthy individuals. In fact, results from unhealthy individuals are just as important as results from healthy individuals. A reference value is the value obtained by measuring the concentration of an analyte from a reference individual. Since it would be impractical to measure the concentration of an analyte from all reference individuals in a reference population, a reference sample group is randomly selected to represent the reference population. The reference distribution is the statistical distribution of the reference values from a specific reference population. The reference distribution is

determined from the reference sample group. The reference interval is a tolerance interval based on the reference distribution. Usually, a ninety five percent tolerance interval is used to construct the reference interval. Based on the reference distribution, either a parametric approach (usually normality assumptions are used, however, other distribution assumptions are possible) or a nonparametric approach is used to construct the reference interval.

A reference interval based on healthy individuals is used to determine the normal reference interval of an analyte. A reference interval based on unhealthy individuals is used to determine the abnormal reference interval of an analyte. Note that analyte concentration measurements from an unhealthy individual can be either abnormally low or abnormally high, depending on the reason for the individual's lack of health. Therefore, both an abnormally low reference interval and an abnormally high reference interval are determined. Laboratories use these reference intervals to assess the health of a patient. However, laboratories can define reference intervals differently since the analytical method used to generate a reference interval, the criteria used to determine a reference population, and the statistical procedure used to derive a reference interval can all vary between laboratories.

A laboratory participating in the **CHEM**$^{\textbf{PLUS}}$ program measures the concentration of an analyte from the five vials. Each concentration measurement is identified as either abnormally low, normal, or abnormally high, depending on the reference intervals set by the laboratory. The fact that the reference intervals are set by the participating laboratories, and not the C.R.L., creates a potential problem. Laboratories with poor performance can often justify their incompetence

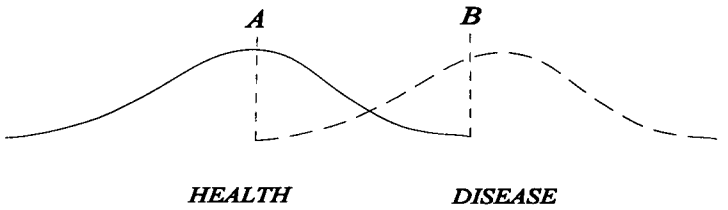
because different laboratories use significantly different reference intervals. This problem can be solved by issuing standardized reference intervals for each analyte. That is, the C.R.L. will identify the quality control samples as either abnormally low, normal, or abnormally high according to standardized reference intervals. Although standardized intervals have not yet been incorporated by the C.R.L., the literature defines reference intervals and decision levels for most analytes [9].

A decision level is the threshold value above which (or below which) a particular action is recommended. A decision level can either be a point of exclusion or a point of confirmation. Figure 1.1 shows an example of these two types of decision levels. The solid curve represents the reference distribution for a healthy reference population and the dotted curve represents the reference distribution for a diseased reference population. The decision level at point A is a point of exclusion. That is, if a clinical test for a patient produces a result less than point A, then the patient can be excluded from the diseased population. The decision level at point B is a point of confirmation. That is, if a clinical test for a patient produces a result greater than point B, then the patient can be included in the diseased population. In addition, if a clinical test for a patient produces a result greater than point A but less than point B, then the patient cannot be definitely included in the diseased population or definitely excluded from the diseased population.

Since a decision level represents the point at which possible medical action is taken, it is imperative that sound laboratory performance is achieved at concentration levels near the

decision levels. Therefore, decision levels should be incorporated into a proficiency testing program so that the importance of medically relevant concentration levels is stressed. Of course, laboratory performance should also be assessed at other possible concentration levels, hence the importance of normal and abnormal reference intervals.

Figure 1.1: Decision Levels



Chapter 2

Standardizing the Target

2.1 Defining the Acceptable Concentration Interval

The acceptable concentration interval defined in (1.1) is used by the C.R.L. to evaluate analyte concentration measurements observed by participating laboratories. The calculation of the acceptable concentration interval depends on the participating laboratory being evaluated. Laboratories in British Columbia and Alberta are required by provincial law to participate in the **CHEM**$^{\textbf{PLUS}}$ program. Results obtained by the C.R.L. are used by the accreditation bodies which regulate the laboratories in British Columbia and Alberta. The Diagnostic Accreditation Program of British Columbia (British Columbia DAP) and the Laboratory Proficiency Testing Program of Alberta (Alberta LPTP) are the respective accreditation bodies in British Columbia and Alberta. The British Columbia DAP and the Alberta LPTP require the use of a peer group approach (discussed in section 2.2) to calculate the acceptable concentration interval. Here, the target equals the mean concentration of a laboratory's peer group, and the limit equals two standard deviations of a laboratory's peer group. All other laboratories participating in the **CHEM**$^{\textbf{PLUS}}$ program are evaluated by a different acceptable concentration interval. Here, the target equals the all-group mean concentration (discussed in section 2.5), and the limit is based on biological variation (discussed in Chapter 3).

2.2 The Peer Group Approach

Peer groups are used in order to compare laboratories that use similar method principles, instrument manufacturers and instrument models. For example, a laboratory measuring the concentration of Albumin might use the Kodak Ektachem 700 which uses bromocresol green dye binding. Here, bromocresol green dye binding is the method principle, Kodak is the instrument manufacturer and Ektachem 700 is the model. For each analyte, a three level hierarchy can be developed which identifies a laboratory first by method, then by instrument and then by model. Within the method level of the hierarchy, a laboratory is placed in a method group, i.e., grouped with other laboratories using the same method. Similarly, within the instrument level of the hierarchy, a laboratory is placed in a method/instrument group and within the model level of the hierarchy, a laboratory is placed in a method/instrument/model group. As one proceeds down this hierarchy, the levels become more refined resulting in more homogeneous groups of laboratories, however, these groups contain fewer number of laboratories. Similarly, as one proceeds up this hierarchy, the levels become less refined resulting in larger groups of laboratories, however, these groups are less homogeneous. Therefore, within each level of the hierarchy there is a balance between the size of the laboratory groups and the homogeneity of the groups. The C.R.L. defines the peer group as the laboratory group with maximum homogeneity such that the size of the group is greater than twenty. For example, assume 100 laboratories measure Albumin using the bromocresol green dye binding method, 25 laboratories use the Kodak instrument and 15 laboratories use the Ektachem 700 model. Then, any laboratory using the Kodak Ektachem 700 instrument system will have a peer

group defined as the method/instrument group, i.e., the peer group will contain the 25 laboratories using the bromocresol green dye binding method and the Kodak instrument. Here, the method/instrument/model group will not be used as the peer group because less than twenty laboratories use the same model.

Therefore, for laboratories in British Columbia and Alberta, the C.R.L. compares an observed concentration of an analyte to the acceptable concentration interval which is determined by the laboratory's peer group. The target equals the peer group mean concentration and the limit equals two peer group standard deviations.

The advantages of using the peer group approach for determining the acceptable concentration interval are that comparisons are made between laboratories that use similar techniques and matrix effects (discussed in section 2.3) are neutralized.

However, there are several disadvantages associated with using the peer group approach for determining the acceptable concentration interval. For instance, the possibility exists that a concentration measurement that accurately reflects the true concentration will be deemed unacceptable, or a concentration measurement that significantly differs from the true concentration will be deemed acceptable. This problem follows from the fact that a concentration measurement is compared to peer concentration measurements and not to the true concentration. Also, the possibility exists that either a laboratory that uses the most popular (but not the best) method will be labeled superior, or a laboratory that uses the least popular (but

superior) method will be labeled poor. As well, a poor method is protected as soon as some peer group forms around the method. Furthermore, by using the peer group approach, medically relevant analytical goals are not considered.

The considerable disadvantages of the peer group approach signify a need to reevaluate the construction of the acceptable concentration interval. This could possibly result in a restructuring of the **CHEM**^{PLUS} program. A "perfect" proficiency testing program should assign target values based on chemical truth, set evaluation criteria based on medically relevant analytical goals, and reduce matrix effects.

2.3 Matrix Effects

Matrix effects are an important concern for proficiency testing programs. The matrix of a sample is defined as the environment in which an analyte and all components other than the analyte reside [10]. The matrix effect is defined as the error introduced by any other component in the sample besides the analyte [10]. Therefore, an important characteristic of a proficiency testing program is that the quality control samples sent to participating laboratories must have properties similar to those of actual specimens used in the laboratories. That is, the matrix of a quality control sample and the matrix of an actual laboratory sample must be similar so that matrix effects do not interfere with the assessment of a laboratory's ability to measure an analyte's concentration.

Many proficiency testing programs use quality control specimens that are derived from modified serum, often lyophilized and augmented with nonhuman materials [11]. Therefore, laboratories that perform unacceptable concentration measurements for the quality control samples may, in fact, be performing well for actual human samples. Similarly, laboratories that perform acceptable concentration measurements for the quality control samples may, in fact, be performing poorly for actual human samples. This discrepancy may be due to matrix effects. The C.R.L. has avoided the problem of matrix effects by using fresh human samples which accurately represent the samples used daily in a clinical laboratory.

2.4 Standardizing the Target via Reference Methods

Therefore, the C.R.L. has successfully reduced matrix effects; however, by using the peer group approach, target values are not based on chemical truth and evaluation criteria are not based on medically relevant analytical goals. Consequently, alternative approaches of constructing the acceptable concentration interval will be investigated. In particular, the acceptable concentration interval will be based on a fixed target and a clinically relevant limit (derivations of the limit are outlined in Chapter 3).

Defining a fixed target based on reference methods is a commonly proposed idea in the clinical chemistry community [12, 13, 14]. A reference method uses the best technology available to perform analyte concentration measurements. However, reference methods are usually time consuming and expensive. Therefore, clinical laboratories use routine methods to perform

analyte concentration measurements. For any given analyte, there may be several routine methods (i.e. different peer groups). In comparison to reference methods, routine methods are relatively quick and inexpensive, however, analytical performance is sacrificed. Using a reference method to measure the concentration of an analyte produces a more accurate estimate of the true concentration. In fact, for most clinical purposes, the concentration measured by a reference method can be viewed as the true concentration. Therefore, a fixed target set by an accepted reference method will be based on chemical truth.

The fixed target, clinically relevant limit approach has been employed in Germany since 1971 [11]. The German Society for Clinical Chemistry (DGKCh) uses a method independent target value determined by a reference method and a limit based on clinical needs. The implementation of this approach has already caused the elimination of many inferior methods.

There are many advantages of using an approach based on a fixed reference method target and a clinically relevant limit for determining the acceptable concentration interval. For instance, laboratories will be motivated to use the best methods and instrument manufacturers will be motivated to develop superior equipment. As well, laboratories that use different methodologies can be compared. Also, true accuracy is reflected since a concentration measurement is compared to a reference method concentration measurement. Finally, medically relevant analytical goals are considered.

However, there are also some disadvantages associated with using an approach based on a fixed reference method target and a clinically relevant limit for determining the acceptable concentration interval. One disadvantage is that differing reference methods may exist for an analyte. As well, reference methods may not exist for an analyte. Also, reference methods may be too expensive or time consuming to incorporate into a proficiency testing program.

2.5 Standardizing the Target via the All-group Mean

Although a fixed reference method target may be optimal, the use of reference methods by a proficiency testing program may be impractical and unrealistic. Another commonly proposed approach is the use of a fixed target based on the all-group mean. The C.R.L. evaluates participating laboratories outside of British Columbia and Alberta by an approach which sets the target equal to the all-group mean. The all-group mean is simply the mean of all the participating laboratories' concentration measurements for an analyte. The literature has shown that for several analytes, the all-group mean correlates closely with reference method values [15, 16, 17]. However, this is not true for every analyte and, in fact, several of the problems associated with the peer group mean are still present with the all-group mean.

Chapter 3

Defining the Limit

3.1 The Limit and its Relation to the Analytical Goal

An important characteristic of a proficiency testing program is that evaluation criteria should be based on medically relevant goals. For the **CHEM^{PLUS}** program, this means that the acceptable concentration interval defined in (1.1) should incorporate a clinically relevant limit. Although different approaches of defining the limit are available, not all are based on clinical relevance. In this chapter, the most common methods of defining the limit will be assessed.

Before proceeding, however, the concept of the analytical goal must be introduced. In the clinical chemistry community, the limit is usually expressed as a percentage of the target. That is, the limit is standardized in terms of the target. When standardized, the limit is referred to as the analytical goal. The analytical goal represents the allowable analytical error during a clinical test. The advantage of using the analytical goal, instead of the limit, is that the analytical goal does not vary depending on the concentration level of the target, whereas the limit does. Therefore, only one analytical goal needs to be defined for each analyte. Note that the acceptable concentration interval can easily be constructed using the analytical goal along with the target value.

3.1.1 The Coefficient of Variation

In the derivation of analytical goals, a commonly used measure of variability is the coefficient of variation (CV). The CV is used in place of the standard deviation. The CV is statistically defined as follows:

$$CV = \frac{standard\ deviation}{mean}$$

The CV standardizes the standard deviation in terms of the mean. For our purposes, however, the standard deviation will be standardized in terms of the target. Therefore, the CV will be defined as follows:

$$CV = \frac{standard\ deviation}{target}$$

The advantage of using the CV instead of the standard deviation is that CV's are comparable over any range of analyte concentration levels.

3.2 Analytical Goals

The analytical error of a clinical test is composed of the inaccuracy (or bias) and the imprecision of the test. Although analytical goals indirectly incorporate both inaccuracy and imprecision, usually no differentiation is made between these two sources of error. Unfortunately, important

information is lost by using analytical goals that do not explicitly account for inaccuracy and imprecision. Laboratories that run clinical tests with high analytical error are uninformed whether inaccuracy or imprecision is their major problem area. However, since all clinical laboratories do some form of internal quality control to assess imprecision, a laboratory should already know if imprecision is a problem. Therefore, the use of analytical goals that do not explicitly account for inaccuracy and imprecision is justified by proficiency testing programs, such as **CHEM**$^{\mathbf{PLUS}}$, which do not require replicate analyses of the same quality control sample (i.e. no way of estimating imprecision). The literature suggests several methods of defining such analytical goals [18, 19]. In this section, the four most common approaches of deriving analytical goals will be discussed.

3.2.1 Analytical Goals Based on Reference Intervals

The first widely used analytical goals were derived from reference intervals. Tonks (1963) originally defined analytical goals as follows:

$$Goal = \frac{0.25 \times Range\ of\ Normal\ Reference\ Interval}{Mean\ of\ Normal\ Reference\ Interval} \times 100\% \qquad (3.1)$$

Tonks' goal can be approximately interpreted as a CV×100%. This follows from the fact that the numerator (one quarter of a range) is approximately a standard deviation, and the denominator is a mean. This interpretation of Tonks' goal is important since it relates the form

of Tonks' goal to the form of the other analytical goals derived hereafter.

Clinical relevance is assured by using reference intervals to derive Tonks' goal. However, there are many disadvantages associated with this goal. The fraction (0.25) used in Tonks' goal is empirical. Also, the normal reference interval is not standardized. Therefore, different goals can be calculated, depending on the normal reference interval used. Factors such as the inaccuracy and imprecision of the analytical method used to generate the reference interval, the characteristics of the reference population, and the statistical procedure used to derive the reference interval can all contribute to producing different reference intervals. Therefore, despite simplicity and clinical relevance, Tonks' goal is too subjective and impractical for clinical use.

3.2.2 Analytical Goals Based on the Opinions of Clinicians

Opinions of clinicians have also been used to derived analytical goals. Barnett (1968) used a series of clinically significant coefficients of variation to define analytical goals. These CV's were obtained by asking clinicians what they viewed as a significant change between two consecutive clinical tests on an individual. Using Barnett's clinically significant CV's, analytical goals are defined as follows:

$$Goal = (2 \times Clinically\ Significant\ CV) \times 100\% \qquad \textbf{(3.2)}$$

Table 3.1 contains Barnett's clinically significant CV's and the respective analytical goals for several analytes.

Table 3.1: Analytical Goals Based on the Opinions of Clinicians		
Analyte	**Clinically Significant CV**	**Analytical Goal**
Albumin	0.070	14.0 %
Calcium	0.023	4.6 %
Chloride	0.018	3.6 %
Cholesterol	0.080	16.0 %
Glucose	0.042	8.4 %
Phosphate	0.056	11.2 %
Potassium	0.042	8.4 %
Sodium	0.013	2.6 %
Protein	0.043	8.6 %
Urea	0.074	14.8 %
Urate	0.083	16.6 %

The advantage of using analytical goals based on clinician opinions is that the goals are based on "expert" opinion and represent clinical relevance. However, it is possible that the opinions of different clinicians could vary significantly. As well, clinicians asked to identify a significant change between two consecutive clinical tests will most likely add preanalytical and biological variation to analytical variation when establishing their opinion. Therefore, the defined analytical goals will incorrectly include sources of error other than analytical variation.

3.2.3 Analytical Goals Based on State of the Art

Analytical goals have also been based on state of the art performance. For laboratories participating in the **CHEM**$^{\textbf{PLUS}}$ program which are located in British Columbia or Alberta, the analytical goals used are based on state of the art performance. This method of deriving analytical goals uses the peer group approach described previously (the limit is defined as two standard deviations of a laboratory's peer group).

In order to express the analytical goal as a percentage, the peer group CV is used in place of the peer group standard deviation. Therefore, state of the art analytical goals are defined as follows:

$$Goal = (2 \times Peer\ Group\ CV) \times 100\% \qquad \textbf{(3.3)}$$

The state of the art goal varies for every quality control sample used in a proficiency testing program. As well, this goal differs for every peer group. This problem of a constantly changing analytical goal is serious since laboratory performance should be comparable over time and locale. Furthermore, state of the art goals are designed to represent the standards currently achievable by clinical laboratories. Unfortunately, these standards may not always represent desirable performance. Hence, state of the art performance goals may not be clinically relevant.

3.2.4 Analytical Goals Based on Biological Variation

Analytical goals have also been derived from biological variation [20]. For laboratories participating in the **CHEM**^{PLUS} program which are not located in British Columbia or Alberta, the analytical goals used are based on biological variation.

Biological variation is the preanalytical component of laboratory error due to in vivo factors in an individual. Biological variation can be divided into two components, between and within biological variation. Between biological variation is the natural variation which occurs between individuals. Within biological variation is the natural variation which occurs within an individual.

Biological variation is only one component of laboratory error. Preanalytical factors (besides biological variation), analytical factors (inaccuracy and imprecision) and postanalytical factors all contribute to the total laboratory error. Laboratory error can adversely affect the result of any clinical test, including the tests run on quality control samples from proficiency testing programs. The following expression models a laboratory's replicate concentration measurements from quality control samples in terms of laboratory error:

$$X_{ij} = T_i + B_i + \alpha_{ij} + \beta_i + \omega_{ij} + \gamma_{ij} \qquad \textbf{(3.4)}$$

where X_{ij} is a laboratory's jth (j=1,...,J) replicate concentration measurement of an analyte from

the ith (I=1,...,I) quality control sample, T_i is the true concentration of the analyte from the ith quality control sample, B_i is the effect of bias (or inaccuracy) on the ith quality control sample, α_{ij} is the effect of imprecision on the jth replicate measurement from the ith quality control sample, β_i is the effect of between biological variation on the ith quality control sample, ω_j is the effect of within biological variation on the jth replicate measurement from the ith quality control sample, and γ_{ij} is the effect of preanalytical variation (other than biological) and postanalytical variation on the jth replicate measurement from the ith quality control sample. Note that bias (B_i) is a fixed effect, while imprecision (α_{ij}), between biological variation (β_I), within biological variation (ω_{ij}) and preanalytical variation (other than biological) and postanalytical variation (γ_{ij}) are all random effects.

We will assume that α_{ij}, β_i, ω_{ij} and γ_{ij} are independent gaussian distributed variables with means zero and standard deviations σ_α, σ_β, σ_ω and σ_γ, respectively. Therefore, the mean and standard deviation of X_{ij} are as follows:

$$E(X_{ij}) = T_i + B_i \qquad \textit{and} \qquad SD(X_{ij}) = \sqrt{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\omega^2 + \sigma_\gamma^2} \qquad \textbf{(3.5)}$$

Hence, the mean of X_{ij} incorporates the bias (fixed effect), whereas the standard deviation of X_{ij} incorporates the other factors of laboratory error (random effects).

Analytical goals based on biological variation are designed to control both inaccuracy and imprecision. First, imprecision will be considered. The clinical chemistry community has

concluded that for a single patient sample (i.e. a single quality control sample), imprecision should add no more than approximately ten percent variation to the total laboratory variation. Using this constraint, analytical goals for imprecision can be derived.

First, note that for a single quality control sample, total laboratory variation is represented by the standard deviation of X_j, where X_j can be expressed by simplifying (3.4). This simplification gives,

$$X_j = T + B + \alpha_j + \omega_j + \gamma_j \qquad \textbf{(3.6)}$$

where X_j is a laboratory's jth (j=1,...,J) replicate concentration measurement of an analyte from the quality control sample, T is the true concentration of the analyte, B is the effect of bias (or inaccuracy), α_j is the effect of imprecision on the jth replicate measurement, ω_j is the effect of within biological variation on the jth replicate measurement, and γ_j is the effect of preanalytical variation (other than biological) and postanalytical variation on the jth replicate measurement. Note that the random effect of between biological variation is not present in this model since only one quality control sample is considered (i.e. only one patient sample).

Therefore, the mean and standard deviation of X_j are as follows:

$$E(X_j) = T + B \qquad \textit{and} \qquad SD(X_j) = \sqrt{\sigma_\alpha^2 + \sigma_\omega^2 + \sigma_\gamma^2} \qquad \textbf{(3.7)}$$

Now, by using the constraint that imprecision should add no more than approximately ten percent variation to the total laboratory variation, it follows that,

$$\left(\frac{\sqrt{\sigma_\alpha^2 + \sigma_\omega^2 + \sigma_\gamma^2}}{\sqrt{\sigma_\omega^2 + \sigma_\gamma^2}} - 1 \right) \times 100\% < 10\%$$

Simplification gives,

$$\sigma_\alpha < 0.458 \sqrt{\sigma_\omega^2 + \sigma_\gamma^2} \qquad \textbf{(3.8)}$$

By using the assumption that preanalytical variation (other than biological) and postanalytical variation have a negligible effect on the final observed value of a clinical test, σ_γ can be set equal to zero. Also, since the allowable percentage of total laboratory variation due to imprecision is empirical, the constant in (3.8) will be defined to be one half. This modification corresponds to imprecision adding no more than 11.8% (instead of 10%) variation to the total laboratory variation. Consequently, (3.8) simplifies to:

$$\sigma_\alpha < \frac{1}{2} \sigma_\omega \qquad \textbf{(3.9)}$$

For consistency, we will use the CV (i.e. the standard deviation standardized in terms of the target) instead of the standard deviation. Thus, (3.9) can be expressed as,

$$CV_{\alpha} < \frac{1}{2} CV_{\omega} \tag{3.10}$$

Hence, (3.10) provides a limit for imprecision based on biological variation. A limit for inaccuracy based on biological variation is not as easily derived. There has been debate whether a limit for inaccuracy is even necessary. Some believe that since laboratory methodology should ideally have no bias, the limit for imprecision given in (3.10) should be applied to total analytical error (inaccuracy and imprecision). However, this ideal is unrealistic for clinical use. Recently, a limit for inaccuracy that has proven clinically useful has been proposed [21]. This limit is defined as,

$$B < \frac{1}{4} \sqrt{\sigma_{\omega}^2 + \sigma_{\beta}^2}$$

Again, we will use the CV instead of the standard deviation. Therefore, it follows that,

$$B' < \frac{1}{4} \sqrt{CV_{\omega}^2 + CV_{\beta}^2} \tag{3.11}$$

where B′ is the bias divided by the target (i.e. bias standardized in terms of the target), CV_{ω} is the within biological variation and CV_{β} is the between biological variation. Note that in (3.11), the bias is standardized in terms of the target so that comparisons to CV's can be made.

The constraints in (3.10) and (3.11) can be used to derive analytical goals based on biological variation. Analytical goals are used to help restrict the amount of laboratory error due to analytical factors (i.e. inaccuracy and imprecision). For a laboratory performing concentration measurements on a single quality control sample, the laboratory's analytical error is represented by $B+\alpha_j$, as in (3.6). The expected value of the analytical error is B and standard deviation of the analytical error is σ_α. Therefore, a laboratory's analytical error is significantly large if the analytical error exceeds $B+2\sigma_\alpha$. Hence, $B+2\sigma_\alpha$ can be used as a limit for analytical error. By standardizing in terms of the target, $B'+2CV_\alpha$ can be used to define analytical goals. However, both B' and CV_α vary depending on the laboratory. In order to define analytical goals that do not depend on the laboratory, maximum allowable B' as defined in (3.11) can be used in place of B' and maximum allowable CV_α as defined in (3.10) can be used in place of CV_α. Therefore, analytical goals based on biological variation can be expressed as,

$$Goal = (Allowable\ B' + 2 \times Allowable\ CV_\alpha) \times 100\%$$

That is,

$$Goal = \left(\frac{1}{4} \sqrt{CV_\omega^2 + CV_\beta^2} + CV_\omega \right) \times 100\% \qquad \textbf{(3.12)}$$

The calculation of analytical goals based on biological variation requires the knowledge of CV_ω and CV_β for every analyte. The data are contained in the literature [22].

The disadvantages of using analytical goals based on biological variation are that the allowable percentage of total laboratory error due to analytical variation is empirical, the data on biological variation may be inaccurate and the data on biological variation are often obtained only from healthy individuals over a short time period. However, these disadvantages may not be of major significance. Previous findings have indicated that biological variation data remain constant over time, between different sizes of subject groups and between subject groups from different countries [17, 18]. Hence, an accurate estimate of biological variation can be obtained by averaging the biological variation data from different studies. Also, for many analytes, the biological variation obtained from healthy individuals is smaller than the biological vartiation obtained from unhealthy individuals. Therefore, by using biological variation obtained only from healthy individuals, the most stringent analytical goals are usually adopted. Although these goals may be unrealistic for current methodology, they represent desirable laboratory performance.

3.3 Imprecision Goals

Usually, analytical goals do not explicitly account for inaccuracy and imprecision. However, goals for imprecision, that explicitly account for inaccuracy, have also been proposed [23, 24].

3.3.1 Imprecision Goals Based on Reference Intervals

Harris (1988) proposed the idea of using reference intervals to derive imprecision goals [23].

This approach will be outlined here. First, R is defined to be the range of the normal reference interval. Furthermore, R' is defined to be R divided by the target (i.e. the range of the normal reference interval standardized in terms of the target). Note that a reference method, with no bias, is used to generate the normal reference interval. Therefore, the results obtained by the reference method are subject to all sources of laboratory error, except inaccuracy. As well, the results obtained by the reference method will be assumed to follow a normal distribution. Hence, R' can be estimated by,

$$R' = 4\sqrt{CV_{\alpha}^2 + CV_{\omega}^2 + CV_{\beta}^2 + CV_{\gamma}^2} \qquad \textbf{(3.13)}$$

where CV_{α} is the random analytical variation (imprecision), CV_{ω} is the within biological variation, CV_{β} is the between biological variation and CV_{γ} is the preanalytical variation (other than biological) and postanalytical variation. Note that (3.13) follows from the fact that, for normally distributed data, a range is approximately equal to four standard deviations.

By using the assumption that CV_{γ} equals zero, (3.13) can be restated as follows:

$$CV_{\omega}^2 + CV_{\beta}^2 = \left(\frac{R'^2}{16}\right) - CV_{\alpha}^2 \qquad \textbf{(3.14)}$$

Now, the following constraint is implemented:

$$CV_{\alpha}^2 + B'^2 < \frac{1}{4}\left(CV_{\omega}^2 + CV_{\beta}^2\right) \tag{3.15}$$

where B′ is the bias divided by the target (i.e. bias standardized in terms of the target). This constraint guarantees that imprecision and bias will add no more than approximately ten percent variation to the total laboratory variation.

By substituting (3.14) into (3.15), the following inequality holds:

$$CV_{\alpha}^2 + B'^2 < \frac{1}{4}\left(\frac{R'^2}{16} - CV_{\alpha}^2\right)$$

Simplifying gives the following limit for imprecision, as proposed by Harris,

$$CV_{\alpha} < R'\sqrt{\frac{1}{80} - \left(\frac{4}{5}\right)\left(\frac{B'}{R'}\right)^2} \tag{3.16}$$

Assuming that the bias can be estimated, (3.16) can be used to derive imprecision goals for each analyte. These imprecision goals will vary for every laboratory, since inaccuracy differs between laboratories. Note that laboratories that run tests with small bias have more lenient imprecision goals than laboratories that run tests with large bias.

Imprecision goals are advantageous since these goals explicitly account for inaccuracy. However, imprecision goals are unrealistic for proficiency testing programs, such as **CHEM**$^{\textbf{PLUS}}$, which do not require replicate analyses of the same quality control sample. Without replicate analyses, an estimate of bias cannot be determined and, therefore, imprecision goals cannot be calculated. Even with limited replicate analyses, a reliable estimate of bias cannot be determined.

Besides the problem of estimating bias, imprecision goals have other disadvantages. Imprecision goals vary for every laboratory and, therefore, laboratory comparability is sacrificed. Also, allowable imprecision, as defined in (3.16), depends on the reference range. As previously mentioned, many problems are associated with using the reference range to derive goals.

Therefore, the use of imprecision goals is not motivated for proficiency testing programs, such as **CHEM**$^{\textbf{PLUS}}$. Analytical goals that do not explicitly account for inaccuracy and imprecision are more suitable for proficiency testing. However, imprecision goals can be very useful for internal quality control.

3.3.2 Imprecision Goals Based on a Clinically Relevant Change

Alternative derivations of imprecision goals have been outlined in the literature. One such derivation, proposed by Fraser, Petersen and Larsen (1990), uses a clinically relevant change to

define imprecision goals [24]. This approach, like other methods of defining imprecision goals, is more useful for internal quality control than for proficiency testing. Therefore, the derivation of imprecision goals based on a clinically relevant change will only be briefly outlined.

Successive tests run on one individual are commonly used for monitoring the health of the individual. If the individual's state of health remains constant from the first clinical test to the second, then there should be little change in the two test results. In contrast, a clinically relevant change in the two test results reflects a change in the individual's state of health.

First, X and Y are defined to be the results from two successive clinical tests performed on an individual. Note that the mean and standard deviation of both X and Y are given in (3.7). Also, D is defined to be the change in the two successive results. Assuming that the individual's state of health remains constant and the two test results are independent observations, the mean and standard deviation of D can be shown to be,

$$E(D) = 0 \qquad \textit{and} \qquad SD(D) = \sqrt{2(\sigma_\alpha^2 + \sigma_\omega^2 + \sigma_\gamma^2)}$$

where σ_α is the random analytical variation (imprecision), σ_ω is the within biological variation and σ_γ is the preanalytical variation (other than biological) and postanalytical variation.

A change in the individual's state of health is reflected by a clinically relevant change in the two test results (i.e. a significantly large value of D). For D to be a clinically relevant change, the following constraint should hold:

$$D > 2 \times \sqrt{2(\sigma_\alpha^2 + \sigma_\omega^2 + \sigma_\gamma^2)}$$

By using the CV instead of the standard deviation, this becomes,

$$D' > 2 \times \sqrt{2(CV_\alpha^2 + CV_\omega^2 + CV_\gamma^2)}$$

where D′ is the change in the two successive results divided by the target (i.e. change standardized in terms of the target).

Using the assumption that CV_γ equals zero gives the following limit for imprecision, as proposed by Fraser, Petersen and Larsen,

$$CV_\alpha < \sqrt{\frac{D'^2}{8} - CV_\omega^2} \qquad \textbf{(3.17)}$$

By setting D′ to be a clinically relevant change, (3.17) can be used to define imprecision goals. However, deciding on an appropriate value for D′ is a difficult and subjective task.

Imprecision goals based on a clinically relevant change are useful for monitoring the precision of successive clinical tests run on an individual. Although limiting the precision of successive clinical tests on one patient is important, monitoring the bias is not necessary since the bias remains constant for successive tests run on the same patient. However, monitoring the bias is essential for proficiency testing, hence, imprecision goals based on a clinically relevant change are not useful for the **CHEM**$^{\textbf{PLUS}}$ program.

Chapter 4

Assessing Overall Laboratory Performance

4.1 Evaluating Laboratories Across Analytes

Recently, the idea of using analytical goals to evaluate laboratories across analytes has been investigated [25, 26, 27, 28]. An overall measure of laboratory performance can be determined by evaluating laboratories across analytes. Furthermore, a properly constructed overall measure of laboratory performance can incorporate all three elements of laboratory performance (i.e. accuracy, precision and linearity). Remember that linearity is the ability of a laboratory to perform concentration measurements, with low analytical error, over the entire concentration range of an analyte.

First, let us introduce an error measure which is commonly defined as follows:

$$e_{hij} = \frac{|X_{hij} - t_{hi}|}{E_{hi}} \qquad \textit{or, equivalently} \qquad e_{hij} = \sqrt{\left(\frac{X_{hij} - t_{hi}}{E_{hi}}\right)^2} \tag{4.1}$$

where e_{hij} is the error measure for the jth (j=1,...,J) replicate concentration measurement from the ith (I=1,...,I) quality control sample of the hth analyte (h=1,...,H), X_{hij} is a laboratory's jth replicate concentration measurement from the ith quality control sample of the hth analyte, t_{hi} is the target value for the ith quality control sample of the hth analyte, and E_{hi} is the error limit

for the ith quality control sample of the hth analyte (i.e. the limit as defined in the acceptable concentration interval). Note that the definition of error measure in (4.1) can be simplified if only one concentration measurement is required for each quality control sample, as in the **CHEM**$^{\textbf{PLUS}}$ program.

Now, the following normality assumptions are made:

$$X_{hij} \sim N\left(T_{hi} + B_{hi},\ \sigma_{1h}^2\right) \qquad \textit{and} \qquad t_{hi} \sim N\left(T_{hi},\ \sigma_{2h}^2\right) \tag{4.2}$$

where T_{hi} is the true concentration from the ith quality control sample of the hth analyte, B_{hi} is the effect of bias (or inaccuracy) on the ith quality control sample of the hth analyte, σ_{1h} is the standard deviation associated with the laboratory method used for the hth analyte, and σ_{2h} is the standard deviation associated with the reference method used for the hth analyte. Note that in (4.2), the standard deviation of X_{hij} is defined simply as σ_{1h} instead of the more complex definition previously used in (3.5) that explicitly incorporates all the random components of laboratory error. Also, in (4.2), the assumption is made that a small amount of error is associated with the reference method used to determine the target value. This error is represented by σ_{2h}, which is much smaller than σ_{1h}, the error associated with the simpler laboratory method. Note that if σ_{2h} is not available, or if a reference method is not used to determine the target value, then σ_{2h} can be set to zero.

An overall measure of laboratory performance, that incorporates the error measure, will be derived. As well, by using the normality assumptions in (4.2), the distribution of the overall performance measure will be found. These derivations are based on the theory originally proposed by Tholen (1988) [25]. First, notice that,

$$\frac{X_{hij} - t_{hi}}{E_{hi}} \sim N\left(\frac{B_{hi}}{E_{hi}}, \frac{\sigma_{1h}^2 + \sigma_{2h}^2}{E_{hi}^2}\right)$$

Furthermore, by standardizing the variance to one, we see that,

$$\frac{X_{hij} - t_{hi}}{E_{hi}} \cdot \frac{E_{hi}}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}} \sim N\left(\frac{B_{hi}}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}}, 1\right)$$

Now, by squaring the above expression, the χ^2 distribution can be used as follows:

$$\left(\frac{X_{hij} - t_{hi}}{E_{hi}}\right)^2 \cdot \left(\frac{E_{hi}}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}}\right)^2 \sim \chi^2_{1,\lambda_{hi}}$$

where $\lambda_{hi} = \frac{B_{hi}^2}{\sigma_{1h}^2 + \sigma_{2h}^2}$ is the noncentrality parameter of the χ^2 distribution (with one degree of freedom).

Consequently,

$$\sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J} e_{hij}^2 \left(\frac{E_{hi}}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}} \right)^2 \sim \chi_{HIJ,\lambda}^2 \qquad \textbf{(4.3)}$$

where $\lambda = J\sum_{h=1}^{H}\sum_{i=1}^{I} \frac{B_{hi}^2}{\sigma_{1h}^2 + \sigma_{2h}^2}$ is the noncentrality parameter of the χ^2 distribution (with HIJ degrees of freedom).

Now, the overall performance measure is defined as follows:

$$\begin{aligned} P &= \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J} e_{hij}^2 \left(\frac{E_{hi}}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}} \right)^2 \\ &= \sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J} \left(\frac{X_{hij} - t_{hi}}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}} \right)^2 \end{aligned} \qquad \textbf{(4.4)}$$

Therefore, in order to evaluate a laboratory across analytes, a performance goal, P_g, can be defined that limits P. That is, a laboratory with overall performance measure, P, less than the performance goal, P_g, is considered to have acceptable overall performance. Similarly, a laboratory with overall performance measure greater than the performance goal is considered to have unacceptable overall performance.

The performance goal, P_g, is defined as follows:

$$P_g = \chi^2_{HIJ,\lambda}(1-\alpha) \qquad \textbf{(4.5)}$$

where λ is the noncentrality goal parameter for the χ^2 distribution (with HIJ degrees of freedom). As well, the value of α used in (4.5) determines the probability of giving a poor evaluation to a laboratory that actually meets the performance goal.

Traditionally in proficiency testing programs, the assessment of a laboratory's overall performance has been based on the percentage of concentration measurements deemed acceptable (via the acceptable concentration interval). For example, a common approach of assessing overall performance is that a laboratory must, for at least eighty percent of the analytes, perform at least four acceptable concentration measurements out of the five quality control samples. However, "acceptable" does not distinguish between measurements with little error and measurements with maximum allowable error. Similarly, "unacceptable" does not distinguish between measurements just outside the acceptable concentration interval and measurements far beyond the acceptable concentration interval boundaries. Therefore, much information is lost by using this traditional approach of assessing overall performance.

This loss of information does not occur, however, by using the overall performance measure, P, to assess a laboratory's overall performance. This is an important advantage of using the overall performance measure. However, disadvantages of using the overall performance measure also

exist. The normality assumptions made in (4.2) are based on the commonly used assumption that analytical error is normally distributed. However, this assumption may not be true for every analyte. Also, an important assumption used to obtain (4.3) is that the variables, X_{hij}-t_{hi} , are independent. However, this assumption may not always hold. For example, the concentrations of certain analytes are sometimes correlated within a single quality control sample. This occurs when the quality control samples are adjusted by the external quality control agency, such as the C.R.L. The quality control samples are adjusted so that analyte concentrations vary over both a normal and an abnormal clinical range. Often, when adjusting a quality control sample in order to obtain a specific concentration for an analyte (for example, abnormally high), other analyte concentrations are also affected in a similar manner. That is, the analyte concentrations are correlated. Therefore, the variables, X_{hij}-t_{hi} , will not be independent. The fact that this independence assumption may not always hold creates a potential problem for deriving the performance goal, P_g. The disadvantages of using the overall performance measure give reason to investigate an alternative measure of overall laboratory performance.

4.2 A Simple Measure of Overall Laboratory Performance

The derivation of a performance goal for overall laboratory performance usually requires underlying distribution assumptions. However, these assumptions may not always hold. Therefore, the use of a simple performance goal that is not based on distribution assumptions may be advantageous.

Again, let us define the error measure as follows:

$$e_{hij} = \sqrt{\left(\frac{X_{hij} - t_{hi}}{E_{hi}} \right)^2} \qquad \textbf{(4.6)}$$

where e_{hij} is the error measure for the jth (j=1,...,J) replicate concentration measurement from the ith (I=1,...,I) quality control sample of the hth analyte (h=1,...,H), X_{hij} is a laboratory's jth replicate concentration measurement from the ith quality control sample of the hth analyte, t_{hi} is the target value for the ith quality control sample of the hth analyte, and E_{hi} is the error limit for the ith quality control sample of the hth analyte.

The interpretation of the error measure is simple. If the error measure obtained from a given concentration measurement is less than or equal to one, then the concentration measurement is deemed acceptable. If the error measure from a given concentration measurement is greater than one, then the concentration measurement is deemed unacceptable. Also note that a lower error measure value is associated with a more credible concentration measurement. Similarly, a higher error measure value is associated with a less credible concentration measurement. For example, assume that two laboratories, say lab A and lab B, each perform a concentration measurement on the same quality control sample and obtain error measures of 0.9 and 0.5, respectively. Although both laboratories have performed acceptable concentration measurements, lab B has actually performed a more credible measurement than lab A.

Let us now propose an overall average error measure that can be used to evaluate a laboratory across analytes. The overall average error measure is defined as follows:

$$\bar{e} = \sqrt{\frac{1}{HIJ}\sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J} e_{hij}^{2}}$$

$$= \sqrt{\frac{1}{HIJ}\sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{X_{hij} - t_{hi}}{E_{hi}}\right)^{2}}$$

Therefore, by expanding the above expression, we see that,

$$\bar{e} = \sqrt{\frac{1}{HIJ}\sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(X_{hij} - \bar{X}_{hi})^2 + (\bar{X}_{hi} - t_{hi})^2}{E_{hi}^2}} \qquad \textbf{(4.7)}$$

In (4.7), the term, $(X_{hij} - \bar{X}_{hi})^2$, can be interpreted as the imprecision and the term, $(\bar{X}_{hi} - t_{hi})^2$, can be interpreted as the bias. Although the overall average error measure as defined in (4.7) provides important information about imprecision and bias, replicate concentration measurements from each quality control sample are required to obtain this information. If replicate concentration measurements are not made, no distinction between imprecision and bias can be made.

The interpretation of the overall average error measure is similar to the interpretation of the error measure defined in (4.6). That is, an overall average error measure less than or equal to one

indicates acceptable laboratory performance across analytes. Similarly, an overall average error measure greater than one indicates unacceptable laboratory performance across analytes. Note that the overall average error measure uses a performance goal simply equal to one and, hence, the performance goal is not based on any distribution assumptions. Also note that a lower overall average error measure is associated with better laboratory performance and a higher overall average error measure value is associated with worse laboratory performance.

4.3 Incorporating Decision Levels

A decision level represents the point at which possible medical action is taken. For each analyte, it is imperative that sound laboratory performance is achieved at concentration levels near the decision levels. Therefore, by incorporating decision levels, an overall weighted measure of laboratory performance can be determined.

Let us now propose an overall weighted error measure that incorporates a weight, w_{hij}, into (4.7). The overall weighted error measure is defined as follows:

$$\bar{e} = \sqrt{\frac{1}{\sum_h \sum_i \sum_j w_{hij}} \sum_{h=1}^{H} \sum_{i=1}^{I} \sum_{j=1}^{J} w_{hij} \cdot \frac{(X_{hij} - \bar{X}_{hi})^2 + (\bar{X}_{hi} - t_{hi})^2}{E_{hi}^2}} \qquad \textbf{(4.8)}$$

As with the overall average error measure defined in (4.7), the overall weighted error measure can be used to assess laboratory performance across analytes. That is, an overall weighted error

measure less than or equal to one indicates acceptable laboratory performance across analytes. Similarly, an overall weighted error measure greater than one indicates unacceptable laboratory performance across analytes.

In (4.8), the weight, w_{hij}, can be defined appropriately so that laboratory performance is assessed more strictly at concentration levels near decision levels and assessed less strictly at concentration levels far from decision levels. Let us assume that the weight, w_{hij}, can be assigned any value from zero to one. In particular, for a given laboratory performing the jth replicate concentration measurement from the ith quality control sample of the hth analyte, the weight, w_{hij}, will be assigned a value close to one in one of two cases. In case one, the target value, t_{hi}, is near a decision level, regardless of the laboratory's observed concentration measurement, X_{hij}. In this case, the target value, which approximates the true concentration level, is near a decision level, hence, sound laboratory performance should be achieved here. In case two, the laboratory's observed concentration measurement, X_{hij}, is near a decision level, regardless of the target value, t_{hi}. In this case, the laboratory's observed concentration measurement, which the laboratory believes is the true concentration level, is near a decision level, hence, sound laboratory performance should be achieved here. Furthermore, if either case one or case two do not hold then the weight, w_{hij}, will not be assigned a value close to one. Rather, the weight will begin to approach zero.

The above requirements are satisfied by defining the weight, w_{hij}, as follows:

$$w_{hij} = \exp\left[-\min\left(\frac{|D^t - t_{hi}|}{E_{hi}}, \frac{|D^X - X_{hij}|}{E_{hi}}\right)\right] \qquad \textbf{(4.9)}$$

where X_{hij} is a laboratory's jth replicate concentration measurement from the ith quality control sample of the hth analyte, t_{hi} is the target value for the ith quality control sample of the hth analyte, E_{hi} is the error limit for the ith quality control sample of the hth analyte, D^t is the decision level nearest to t_{hi}, and D^X is the decision level nearest to X_{hij}.

A disadvantage of using (4.9) to define the weights for a given laboratory is that the possibility exists that some weights can be assigned a value close to zero, while the other weights can be assigned a value close to one. This means that no importance is given to laboratory performance at concentration levels far from decision levels, while all the importance is given to laboratory performance at concentration levels near decision levels. Although laboratory performance should be assessed more strictly at concentration levels near decision levels and assessed less strictly at concentration levels far from decision levels, some importance should still be given to laboratory performance at all concentration levels. In order to correct this problem, a minimum constraint can be required for the weights. For example, all weights can be required to exceed a minimum constraint of 0.05. Of course, other appropriate definitions of the weight can be used in place of the definition given in (4.9).

There are many advantages of using the overall average error measure or the overall weighted error measure to assess laboratory performance. For example, distribution assumptions are not required by using a performance goal simply equal to one. As well, the loss of information that occurs by using traditional methods of assessing overall laboratory performance, based on the percentage of concentration measurements deemed acceptable, does not occur by using the overall average error measure or the overall weighted error measure. Also, the interpretation of the overall average error measure and the overall weighted error measure is simple. Finally, the two performance measures incorporate all three elements of laboratory performance, that is, accuracy, precision and linearity.

Therefore, the use of either the overall average error measure or the overall weighted error measure (or a combination of the two measures) provides a credible indication of overall laboratory performance. In addition, these performance measures can be used to easily rank the clinical laboratories participating in a proficiency testing program.

Chapter 5

Analysis of CHEM^{PLUS} Data

5.1 The CHEM^{PLUS} Data

The C.R.L. has developed the **CHEM^{PLUS}** proficiency testing program in order to assess laboratory performance for general chemistry. Recall that laboratories participating in the **CHEM^{PLUS}** program measure analyte concentrations from each of the five vials received during a shipment. The five vials are randomly color coded as blue, green, yellow, orange or red.

Our **CHEM^{PLUS}** data analysis will be based on data from the January 1996 **CHEM ^{PLUS}**shipment. Furthermore, the analysis will be limited to the data from only five analytes. These analytes are calcium, chloride, magnesium, potassium and sodium.

5.1.1 Removing Outliers

The mean and standard deviation of a population can be heavily influenced by outliers. Therefore, outliers are eliminated before any shipment statistics are generated by the C.R.L. For each vial, the following statistical procedure is used to designate outliers. First, the first quantile (Q1) and the third quantile (Q3) are calculated from all laboratory concentration measurements obtained for a vial. Next, a scaling factor (SF) is defined as 1.5(Q3 - Q1). Then, the outer fences of the population are set as Q1 - 2(SF) and Q3 + 2(SF). Now, any laboratory

measurement which falls outside these outer fences is designated as an outlier. For our **CHEM**$^{\textbf{PLUS}}$ data analysis, outliers will also be removed according to this procedure.

5.1.2 Removing Other CHEM^{PLUS} Data

Besides eliminating outliers, other observations will also be removed before commencing our **CHEM**$^{\textbf{PLUS}}$ data analysis. For each vial, certain "problem" laboratory measurements will be removed. Examples of "problem" measurements include late measurements or measurements performed on an insufficient serum sample. Besides "problem" measurements, measurements performed by any laboratory located outside of British Columbia or Alberta will also be removed. The C.R.L. has historically evaluated these laboratories by using an acceptable concentration interval different from the one used for laboratories located within British Columbia or Alberta. Therefore, for consistency, our **CHEM**$^{\textbf{PLUS}}$ data analysis will only be based on measurements performed by laboratories located within British Columbia or Alberta. The C.R.L. evaluates these laboratories by using an acceptable concentration interval derived by the peer group approach.

Table A.1, located in Appendix A, summarizes the January 1996 **CHEM**$^{\textbf{PLUS}}$ shipment data used for our analysis of calcium, chloride, magnesium, potassium and sodium.

5.1.3 Initial Data Analysis

After removing unwanted observations from the January 1996 **CHEM**^{PLUS} shipment data, initial data analysis can be performed. First, information regarding peer groups will be provided. Tables 5.1 to 5.5 list the peer groups for calcium, chloride, magnesium, potassium and sodium, respectively. These tables show the number of laboratories that use each peer group (i.e. the number of laboratories that use the statistics derived from the peer group). As well, the number of laboratories contained in each peer group (i.e. the number of laboratories used to derive the peer group statistics) is shown in parentheses. The number of laboratories contained in each peer group is always greater than or equal to the number of laboratories that use the peer group.

Table 5.1: Calcium Peer Groups

Method	Instrument	Model	Blue Vial	Green Vial	Yellow Vial	Orange Vial	Red Vial
All Group			1 (147)	1 (148)	1 (147)	1 (149)	1 (147)
ARSDYE			12 (82)	13 (84)	12 (82)	13 (84)	12 (83)
ARSDYE	BEKMAN		31 (31)	31 (31)	31 (31)	32 (32)	32 (32)
ARSDYE	KODAK		1 (39)	1 (39)	1 (39)	1 (39)	1 (39)
ARSDYE	KODAK	EKTSYS	38 (38)	38 (38)	38 (38)	38 (38)	38 (38)
CRESO			64 (64)	64 (64)	64 (64)	64 (64)	63 (63)

Table 5.2: Chloride Peer Groups							
Method	**Instrument**	**Model**	**Blue Vial**	**Green Vial**	**Yellow Vial**	**Orange Vial**	**Red Vial**
All Group			18 (178)	20 (181)	19 (180)	20 (181)	17 (177)
ISE			59 (160)	60 (161)	60 (161)	60 (161)	59 (160)
ISE	BEKMAN		19 (50)	19 (50)	19 (50)	19 (51)	19 (51)
ISE	BEKMAN	SYNSYS	31 (31)	31 (31)	31 (31)	32 (32)	32 (32)
ISE	KODAK		15 (51)	15 (51)	15 (51)	14 (50)	14 (50)
ISE	KODAK	EKTSYS	36 (36)	36 (36)	36 (36)	36 (36)	36 (36)

Table 5.3: Magnesium Peer Groups							
Method	**Instrument**	**Model**	**Blue Vial**	**Green Vial**	**Yellow Vial**	**Orange Vial**	**Red Vial**
All Group			2 (79)	2 (79)	2 (79)	2 (80)	1 (76)
CALG			3 (24)	3 (24)	3 (24)	3 (24)	3 (24)
CALG	BEKMAN	SYNSYS	21 (21)	21 (21)	21 (21)	21 (21)	21 (21)
COLOR			26 (53)	26 (53)	26 (53)	27 (54)	24 (51)
COLOR	KODAK	EKTSYS	27 (27)	27 (27)	27 (27)	27 (27)	27 (27)

Table 5.4: Potassium Peer Groups							
Method	**Instrument**	**Model**	**Blue Vial**	**Green Vial**	**Yellow Vial**	**Orange Vial**	**Red Vial**
All Group			17 (219)	19 (218)	18 (219)	19 (216)	17 (214)
ISE			72 (202)	71 (199)	72 (201)	71 (197)	72 (197)
ISE	BEKMAN		20 (53)	20 (52)	20 (52)	20 (53)	20 (53)
ISE	BEKMAN	SYNSYS	33 (33)	32 (32)	32 (32)	33 (33)	33 (33)
ISE	KODAK	EKTSYS	40 (77)	40 (76)	40 (77)	40 (73)	40 (72)
ISE	KODAK	KDT60	37 (37)	36 (36)	37 (37)	33 (33)	32 (32)

Table 5.5: Sodium Peer Groups							
Method	**Instrument**	**Model**	**Blue Vial**	**Green Vial**	**Yellow Vial**	**Orange Vial**	**Red Vial**
All Group			18 (219)	19 (219)	18 (220)	19 (217)	17 (214)
ISE			72 (201)	72 (200)	72 (202)	72 (198)	72 (197)
ISE	BEKMAN		19 (52)	19 (52)	19 (53)	19 (53)	19 (53)
ISE	BEKMAN	SYNSYS	33 (33)	33 (33)	34 (34)	34 (34)	34 (34)
ISE	KODAK	EKTSYS	40 (77)	40 (76)	40 (77)	40 (73)	40 (72)
ISE	KODAK	KDT60	37 (37)	36 (36)	37 (37)	33 (33)	32 (32)

The use of box plots provides an informative graphical display of data. For our **CHEM**$^{\textbf{PLUS}}$ data analysis, box plots of laboratory concentration measurements (by peer group) will be constructed. These box plots will be created for each vial of each analyte. Figures A.1 to A.5, located in Appendix A, display these box plots for calcium, chloride, magnesium, potassium and sodium, respectively. Note that for each peer group, the box plot is constructed from concentration measurements performed by laboratories contained in the peer group (i.e. the laboratories used to derive the peer group statistics). These box plots allow us to easily compare the concentration measurement distributions of the different peer groups. For example, the box plots for calcium show that both the center and the spread of the concentration measurement data vary depending on the peer group. This characteristic is also evident in the box plots of the other analytes, except for perhaps magnesium. The fact that the distribution of the concentration measurement data varies depending on the peer group provides initial evidence that the peer group approach for defining the acceptable concentration interval may not be adequate. One problem is that using differing target values (based on the peer group mean) for each vial contradicts the fact that there is only one true concentration level for the vial. Another problem is that using differing limits (based on two peer group standard deviations) for each vial results in a strict evaluation for some laboratories and a lenient evaluation for others. Further analysis is needed to strengthen the argument against the peer group approach and to investigate the alternative approaches of defining the acceptable concentration interval.

5.2 The Limit

The acceptable concentration interval defined in (1.1) is composed of the target and the limit. In Chapter 2, three methods of defining the target are outlined (i.e. the peer group mean approach, the all-group mean approach and the reference method approach). In Chapter 3, four methods of deriving the limit, or equivalently, the analytical goal are considered (i.e. the reference interval approach, the opinion of clinicians approach, the state of the art approach and the biological variation approach). In this section, the different methods of deriving the limit, or equivalently, the analytical goal are compared. In section 5.3, the different approaches of defining the target are compared.

The use of properly defined analytical goals insures that acceptable analytical performance is maintained in clinical laboratories. With this in mind, the four methods of deriving analytical goals will be analyzed by using the January 1996 **CHEM**$^{\textbf{PLUS}}$ shipment data for calcium, chloride, magnesium, potassium and sodium.

5.2.1 The Reference Interval Approach

Analytical goals based on reference intervals are defined in (3.1). Table 5.6 contains the normal reference interval, the range of the normal reference interval, the midpoint of the normal reference interval and the analytical goal for each of the five analytes of interest. Note that the normal reference interval data have been obtained from the literature [9].

Table 5.6: Analytical Goals Based on Reference Intervals				
Analyte	**Normal Reference Interval**	**Range of Reference Interval**	**Midpoint of Reference Interval**	**Analytical Goal**
Calcium	2.25 - 2.64 mmol/L	0.39	2.44	3.99 %
Chloride	98 - 109 mmol/L	11.0	103.5	2.66 %
Magnesium	0.6 - 1.2 mmol/L	0.6	0.9	16.67 %
Potassium	3.7 - 5.1 mmol/L	1.4	4.4	7.95 %
Sodium	138 - 146 mmol/L	8.0	142.0	1.41 %

5.2.2 The Opinion of Clinicians Approach

Analytical goals based on the opinions of clinicians are defined in (3.2). Table 5.7 contains the clinically significant CV and the analytical goal for each of the five analytes of interest.

Table 5.7: Analytical Goals Based on the Opinions of Clinicians		
Analyte	**Clinically Significant CV**	**Analytical Goal**
Calcium	0.023	4.6 %
Chloride	0.018	3.6 %
Magnesium	N/A	N/A
Potassium	0.042	8.4 %
Sodium	0.013	2.6 %

5.2.3 The State of the Art Approach

Analytical goals based on state of the art are defined in (3.3). The C.R.L. currently uses the state of the art approach to derive the limit for laboratories located in British Columbia and Alberta. One major problem with the state of the art approach is that the analytical goal varies for every vial and for every peer group. Therefore, for each analyte, the state of the art analytical goal cannot be expressed as a fixed value.

5.2.4 The Biological Variation Approach

Analytical goals based on biological variation are defined in (3.12). The C.R.L. currently uses the biological variation approach to derive the limit for laboratories located outside of British Columbia and Alberta. Table 5.8 contains the within biological variation (CV_{ω}), the between biological variation (CV_{β}), the allowable imprecision (CV_{α}), the allowable bias (B') and the analytical goal for each of the five analytes of interest. Note that the within biological variation data and the between biological variation data have been obtained from the literature [22]. Also, allowable imprecision is defined in (3.10) and allowable bias is defined in (3.11). The analytical goal, which is defined in (3.12), combines allowable imprecision and allowable bias.

Table 5.8: Analytical Goals Based on Biological Variation					
Analyte	**CV_ω**	**CV_β**	**Allowable CV_α**	**Allowable B′**	**Analytical Goal**
Calcium	0.0267	0.0375	0.0133	0.0115	3.82 %
Chloride	0.0121	0.0128	0.00605	0.00440	1.65 %
Magnesium	0.0469	0.0690	0.0235	0.0208	6.78 %
Potassium	0.0121	0.0128	0.00605	0.00440	1.65 %
Sodium	0.00972	0.00693	0.00486	0.00298	1.27 %

5.2.5 Comparison of the Analytical Goal Derivation Methods

For each vial of an analyte, an acceptable concentration interval can be constructed by using the analytical goal in conjunction with the target value. Of course, the acceptable concentration interval will depend on the methodology used to define the analytical goal and the target value. Figures A.6 to A.30, located in Appendix A, graphically display various acceptable concentration intervals (by peer group) for the twenty five analyte/vial combinations. Each of these figures contains four plots. The four plots differ by the analytical goal derivation method used to construct the acceptable concentration intervals (i.e. plot one uses the reference interval approach, plot two uses the opinion of clinicians approach, plot three uses the state of the art approach and plot four uses the biological variation approach). Furthermore, each plot displays two acceptable concentration intervals for each peer group (one that that uses the all-group mean target and one that uses the peer group mean target). As well, each plot shows the reference method target (the horizontal dotted line) and the concentration measurements

performed by laboratories that use each peer group. Note that the reference method targets used have been set by the Wisconsin State Laboratory of Hygiene (W.S.L.H.), an established reference laboratory. The W.S.L.H. uses reference methods validated by the National Committee for Clinical Laboratory Standards.

Figures A.6 to A.30 provide valuable information pertaining to both the target and the limit. Here, the limit will be considered. By comparing the four plots in these figures, important differences between the four methods of deriving the limit become evident. First, notice that for magnesium and potassium, the limits based on reference intervals are very lenient. Almost all of the laboratory concentration measurements are deemed acceptable. This characteristic is also apparent with limits based on opinions of clinicians. The problem with using a lenient limit to define the acceptable concentration interval is that laboratories with poor analytical performance will be given a favourable evaluation. In fact, no distinction will be made between laboratories with poor analytical performance and laboratories with superior analytical performance.

Other problems with the reference interval approach and the opinion of clinicians approach also exist. As indicated in Chapter 3, these approaches are too subjective and impractical for clinical use. Due to the problems associated with the reference interval approach and the opinion of clinicians approach, neither approach will be used to construct the acceptable concentration interval.

Now, notice that for all five analytes, the limits based on state of the art vary depending on the peer group. As noted previously, this results in a strict evaluation for some laboratories and a lenient evaluation for others. Hence, the laboratories that are strictly evaluated (i.e. use a peer group with a tight limit) will be forced to achieve superior analytical performance; however, the laboratories that are leniently evaluated (i.e. use a peer group with a loose limit) will not be motivated to improve their analytical performance.

As indicated in Chapter 3, other problems with the state of the art approach also exist. The state of the art limit reflects the standards currently achievable by clinical laboratories. Unfortunately, these standards may not always represent desirable performance. Hence, the limit may not be clinically relevant. Due to the problems associated with the state of the art approach, this approach will not be used to construct the acceptable concentration interval.

Finally, let us consider the biological variation approach. A limit based on biological variation exhibits several appealing characteristics. For instance, the biological variation limit is clinically relevant, since well researched biological variation data is used to define the limit. Also, the biological variation limit does not vary with the peer group, as is the case with the state of the art approach. As well, limits based on biological variation are consistently tight, which corresponds with a strict evaluation of analytical performance. Although, for some analytes (for example, potassium) the biological variation limits may be considered too tight. In these cases, the limits should be viewed as goals for desirable analytical performance. For our analysis, the biological variation approach will be used to construct the acceptable concentration interval.

5.3 The Target

Ideally, when constructing the acceptable concentration interval for a vial, the target value should represent the true concentration level of the vial. Using this criterion, the three approaches of defining the limit will be compared by using the January 1996 **CHEM**^{PLUS} shipment data for calcium, chloride, magnesium, potassium and sodium.

5.3.1 The Peer Group Mean Approach

The C.R.L. currently uses the peer group mean to define the target for laboratories located in British Columbia and Alberta. One major problem with using the peer group mean approach is that the possibility exists that different target values will be defined for different peer groups. Figures A.6 to A.30 graphically display the severity of this problem. For every vial of each analyte, with the exception of magnesium, the peer group mean target varies depending on the peer group. As noted previously, this characteristic contradicts the fact that there is only one true concentration level for each vial. For our analysis, we will assume that this true concentration level is best approximated by the reference method target, set by the W.S.L.H. Many of the plots in figures A.6 to A.30 illustrate the case where a reference method target is contained in the acceptable concentration intervals of some peer groups, but not in others. Therefore, the possibility exists that a laboratory, using a particular peer group, can perform an accurate concentration measurement that is deemed "unacceptable". Similarly, a laboratory, using a different peer group, can perform an inaccurate concentration measurement that is

deemed "acceptable". This provides strong evidence against using the peer group mean approach to construct the acceptable concentration interval.

To further illustrate the problem of using a target value that varies depending on the peer group, let us consider the blue vial of calcium. Plot four of figure A.6 shows the acceptable concentration intervals, constructed with a limit based on biological variation, for the blue vial of calcium. In particular, notice that by using the peer group mean target, the acceptable concentration intervals vary significantly depending on the peer group. For example, the ARSDYE BEKMAN peer group and the CRESO peer group use completely distinct acceptable concentration intervals. This is a serious problem since one of the decision levels for calcium is 1.75 mmol/L [9]. Notice that this decision level lies approximately between the two acceptable concentration intervals. Therefore, the possibility exists that "acceptable" laboratories using the ARSDYE BEKMAN peer group will not take a particular medical action, while "acceptable" laboratories using the CRESO peer group will take the medical action. This is another major disadvantage of the peer group mean approach. Due to the many problems associated with the peer group mean approach (other problems are discussed in Chapter 2), this approach will not be used to construct the acceptable concentration interval.

5.3.2 The All-group Mean Approach

Defining a fixed target for each vial is the logical alternative to using the peer group mean approach. The all-group mean approach is one method of defining a fixed target for each vial.

The C.R.L. currently uses the all-group mean approach to define the target for laboratories located outside of British Columbia and Alberta. The literature has shown that for several analytes, the all-group mean correlates closely with reference method values [15, 16, 17]. This is an appealing characteristic since reference method values represent chemical truth.

In order to assess the all-group mean approach, the all-group mean will be compared to the reference method target set by the W.S.L.H. For each vial of an analyte, an acceptable concentration interval will be constructed that uses a limit based on biological variation and an all-group mean target. Also, a second acceptable concentration interval will be constructed that uses a limit based on biological variation and a reference method target. Then, a two by two contingency table will be used to compare the number of laboratory concentration measurements that pass (or fail) based on the two acceptable concentration intervals. Note that only one contingency table, that sums the number of passing (or failing) measurements for the five vials, will be constructed for each analyte. Similar contingency tables will also be constructed that compare each peer group mean, instead of the all-group mean, to the reference method target. That is, for each peer group of an analyte, a contingency table will be used to compare the number of passing (or failing) measurements based on two acceptable concentration intervals (one that uses a fixed target defined as the mean of the peer group and the other that uses a reference method target).

Figures A.31 to A.35, located in Appendix A, display the two by two contingency tables for calcium, chloride, magnesium, potassium and sodium, respectively. As well, an overall

proportion of agreement (with approximate standard deviation) is given for each table. This is the proportion of measurements that either pass according to both acceptable concentration intervals, or fail according to both acceptable concentration intervals. Notice that for calcium and chloride, the overall proportion of agreement is greatest for the contingency table that compares the all-group mean target to the reference method target. Therefore, for calcium and chloride, a fixed target based on the all-group mean correlates closest with the reference method target. Similarly, for sodium, a fixed target based on the mean of the ISE KODAK EKTSYS peer group is optimal, followed by the all-group mean target. For magnesium, the overall proportion of agreement does not significantly differ between contingency tables. This result follows from the fact that, for magnesium, the peer group mean target does not significantly vary depending on the peer group. Finally, for potassium, a fixed target based on the mean of the ISE BEKMAN SYNSYS peer group is optimal. However, since the limit based on biological variation is extremely tight for potassium, differences in the overall proportion of agreement may be exaggerated.

Overall, a fixed target based on the all-group mean seems to correlate closest with the reference method target. As well, by using the all-group mean, the definition of the target does not change from analyte to analyte. This will occur, however, if the mean of a specific peer group is used as the target for some analytes. Therefore, using a fixed target based on the all-group mean is an acceptable substitute for using a target based on reference methods.

Problems with the all-group mean approach do exist, however. One major problem is that the all-group mean is adversely affected by concentration measurements performed by laboratories that use inaccurate methods or instruments. To illustrate this problem, let us consider the blue vial of calcium. Plot four of figure A.6 shows the acceptable concentration intervals, constructed with a limit based on biological variation, for the blue vial of calcium. In particular, notice that the all-group mean target (1.67 mmol/L) is lower than the reference method target (1.71 mmol/L). This difference is due to the laboratories that use the CRESO method. These laboratories generally perform concentration measurements lower than the other laboratories. In fact, by removing the laboratories that use the CRESO method, the "modified" all-group mean (1.71 mmol/L) comes into exact agreement with the reference method target. Furthermore, the difference between the actual all-group mean and the "modified" all-group mean is statistically significant (P-value less than 0.01). Therefore, for the blue vial of calcium, the all-group mean is adversely altered by measurements performed by laboratories that use the CRESO method. This problem occurs with other analytes as well, indicating that the all-group mean approach is not perfect.

5.3.3 The Reference Method Approach

Although defining the target as the all-group mean is an acceptable approach, using reference method targets, if available, is the preferred method. Chemical truth is established by using a fixed target set by an accepted reference method. For our analysis, the reference method targets set by the W.S.L.H. will be used to construct the acceptable concentration interval.

5.4 Assessing Laboratory Performance

The acceptable concentration interval, constructed with a reference method target and a limit based on biological variation, will now be used to assess laboratory performance. Three levels of laboratory performance will be considered. These levels are vial performance, analyte performance and overall performance.

5.4.1 Vial Performance

Assessing laboratory performance for each vial is easily accomplished by using the acceptable concentration interval. For each vial, the concentration measurement performed by a laboratory passes evaluation if the measurement lies within the acceptable concentration interval.

Table 5.9 summarizes vial performance for the January 1996 **CHEM**$^{\textbf{PLUS}}$ shipment data. For each analyte, table 5.9 shows the number of laboratories evaluated and the percentage of laboratories that perform a passing concentration measurement for each vial. Note that the number of laboratories evaluated only includes those laboratories that, for all five vials, have performed concentration measurements that have not been removed from the analysis. Removed measurements include outliers, "problem" measurements or measurements performed by laboratories located outside of British Columbia and Alberta.

Table 5.9: Percentage of Laboratories that Perform a Passing Concentration Measurement						
Analyte	**Number of Labs**	**Blue Vial**	**Green Vial**	**Yellow Vial**	**Orange Vial**	**Red Vial**
Calcium	145	36.6 %	72.4 %	66.2 %	72.4 %	78.6 %
Chloride	174	44.3 %	68.4 %	58.1 %	56.3 %	48.9 %
Magnesium	74	58.1 %	91.9 %	82.4 %	20.3 %	75.7%
Potassium	210	0.0 %	40.0 %	43.8%	68.6 %	37.6 %
Sodium	211	46.9 %	49.3 %	12.8 %	68.7 %	57.4 %

Many of the percentages contained in table 5.9 are low, indicating that few laboratories are performing passing concentration measurements. However, a failing concentration measurement is not necessarily equivalent with a truly unacceptable concentration measurement. In fact, the low percentages observed in table 5.9 are mainly due to the fact that a tight limit (based on biological variation) is used to construct the acceptable concentration interval.

5.4.2 Analyte Performance

Assessing laboratory performance for each analyte is accomplished by using the acceptable concentration interval in conjunction with the theory contained in Chapter 4. Three measures of overall laboratory performance are derived in Chapter 4. The overall performance measure is defined in (4.4), the overall average error measure is defined in (4.7) and the overall weighted error measure is defined in (4.8). Although these performance measures are designed for evaluating overall laboratory performance, they can be easily simplified in order to assess

analyte performance. By setting H (the number of analytes) to one, the three overall performance measures can be used to assess laboratory performance for each analyte. Note that these three simplified overall performance measures will now be referred to as simply the performance measure, the average error measure and the weighted error measure.

Table 5.10 summarizes analyte performance for the January 1996 **CHEM**$^{\mathbf{PLUS}}$ shipment data. For each analyte, table 5.10 shows the number of laboratories evaluated, the percentage of laboratories that perform at least four passing concentration measurements out of the five vials, the percentage of laboratories that achieve a passing performance measure, the percentage of laboratories that achieve a passing average error measure and the percentage of laboratories that achieve a passing weighted error measure. Note that a laboratory's performance measure passes evaluation if it is less than the performance goal defined in (4.5), with alpha set to 0.05. As well, a laboratory's average error measure passes evaluation if it is less than one. Similarly, a laboratory's weighted error measure passes evaluation if it is less than one. Note that the weight used in the calculation of the weighted error measure is defined in (4.9). Also, all weights are required to exceed a minimum constraint of 0.05. In addition, the decision levels used in the calculation of the weight are contained in the literature [9].

Table 5.10: Percentage of Laboratories that Achieve Passing Analyte Performance					
Analyte	**Number of Labs**	**4 out of 5 Passing Measurements**	**Performance Measure**	**Average Error Measure**	**Weighted Error Measure**
Calcium	145	49.0 %	86.2 %	55.2 %	33.0 %
Chloride	174	37.4 %	85.6 %	37.9 %	48.3 %
Magnesium	74	55.4 %	86.5 %	56.8 %	54.1 %
Potassium	210	6.7 %	94.3 %	7.6 %	7.14 %
Sodium	211	25.6 %	91.0 %	10.9 %	5.2 %

The traditional approach of assessing analyte performance is represented by the criterion that a laboratory must perform at least four passing concentration measurements out of the five vials. Here, a passing concentration measurement is a measurement that lies within the acceptable concentration interval. As discussed in Chapter 4, much information is lost by using this traditional approach. This loss of information does not occur, however, by using the other three measures to assess analyte performance.

Many of the percentages contained in table 5.10 are low, indicating that few laboratories are achieving a passing evaluation for analyte performance. However, these low percentages are mainly due to the fact that the limit (based on biological variation) is tight. Notice, however, that low percentages are not apparent when the performance measure is used to assess analyte performance. This is due to the fact that the performance goal defined in (4.5) incorporates estimates of the bias and the imprecision of a "typical" laboratory. The bias estimate is obtained

by taking a weighted average of the absolute bias of each peer group, and the imprecision estimate is obtained by taking a weighted average of the imprecision of each peer group. For our **CHEM**$^{\mathbf{PLUS}}$ data analysis, the estimates obtained for the bias and the imprecision of a "typical" laboratory produce a lenient performance goal and, hence, many laboratories achieve a passing evaluation for analyte performance.

Also, note that the value of the performance goal varies depending on the value of alpha used to define the goal. That is, a performance goal defined with a small value of alpha results in a lenient goal, and a performance goal defined with a large value of alpha results in a strict goal. Hence, the value of alpha can be adjusted so that the performance goal is neither too lenient nor too strict.

Therefore, the problem of too few laboratories achieving a passing evaluation seems to be solved by using the performance measure. However, disadvantages with the performance measure exist. As indicated in Chapter 4, the distribution assumptions used to define the performance goal may not always hold. Another problem with the performance goal is that the bias estimate and the imprecision estimate used to calculate the goal are, in theory, laboratory dependent. However, an important characteristic of a performance goal is that the goal is not laboratory dependent. For this reason, the performance goal used in our analysis incorporates estimates of the bias and the imprecision of a "typical" laboratory. Unfortunately, the idea of a standardized bias estimate and imprecision estimate for every laboratory is unrealistic. In fact, by using standardized estimates to derive the performance goal, only a "rough" goal is obtained

for every laboratory. This is an acceptable compromise, however, since the role of proficiency testing agencies is to only monitor participating laboratories. Therefore, using the "rough" performance goal, based on standardized estimates of bias and imprecision, is acceptable; however, the computation of the bias and imprecision estimates is time consuming, considering a bias estimate is needed for every vial and an imprecision estimate is needed for every analyte.

The average error measure and the weighted error measure are feasible alternatives to the performance measure. As indicated in Chapter 4, these two error measures are easy to compute and interpret. As well, the goals used for these error measures are not based on distribution assumptions. In fact, a goal equal to one can be used for both of these measures. However, for our analysis, this goal proves to be too tight. Table 5.10 shows that too few laboratories are achieving a passing evaluation for analyte performance. As previously discussed, this is due to the fact that the limit (based on biological variation) is tight. A simple approach of solving this problem is to increase the goal so that more laboratories achieve a passing evaluation. Over time, the goal can be slowly tightened towards the desired value of one.

5.4.3 Overall Performance

Assessing overall laboratory performance is accomplished by using the three overall laboratory performance measures derived in Chapter 4. The overall performance measure is defined in (4.4), the overall average error measure is defined in (4.7) and the overall weighted error measure is defined in (4.8).

Table 5.11 summarizes overall performance for the January 1996 **CHEM**^{PLUS} shipment data. Table 5.11 shows the number of laboratories evaluated, the percentage of laboratories that, for at least eighty percent of the analytes, perform at least four passing concentration measurements out of the five vials, the percentage of laboratories that achieve a passing overall performance measure, the percentage of laboratories that achieve a passing overall average error measure and the percentage of laboratories that achieve a passing overall weighted error measure.

Table 5.11: Percentage of Laboratories that Achieve Passing Overall Performance

Number of Labs	4 out of 5 Passing Measurements for 80% of the Analytes	Overall Performance Measure	Overall Average Error Measure	Overall Weighted Error Measure
212	1.9 %	85.9 %	5.7 %	3.3 %

The traditional approach of assessing overall performance is represented by the criterion that a laboratory must, for at least eighty percent of the analytes, perform at least four passing concentration measurements out of the five vials. As previously indicated, this approach results in the loss of important information and, hence, should not be used.

Table 5.11 shows that for the overall average error measure and the overall weighted error measure, the percentage of laboratories that achieve passing overall performance is extremely low. This is not a problem, however, with the overall performance measure. Although, the

overall performance measure suffers from the same drawbacks as the performance measure used to assess analyte performance. For the overall average error measure and the overall weighted error measure, the percentage of laboratories that achieve passing overall performance can be increased by simply increasing the goal from one. For example, using a goal equal to 2.5 results in 94.8% of the laboratories achieving a passing overall average error measure and 87.7% of the laboratories achieving a passing overall weighted error measure. In this case, the proportion of overall agreement between the overall average error measure and the overall weighted error measure is 0.920 (with an approximate standard deviation of 0.019). This is the proportion of laboratories that either pass according to both measures or fail according to both measures. Therefore, the overall average error measure and the overall weighted error measure are comparable, although slightly more laboratories fail according to the overall weighted error measure.

5.5 Ranking Laboratory Performance

The overall average error measure and the overall weighted error measure can be used to easily rank either analyte performance or overall laboratory performance. A better performance ranking is simply associated with a lower error measure. By ranking laboratory performance, distinction can be made between two laboratories that both pass evaluation (or two laboratories that both fail evaluation).

Although the overall performance measure, P, can be used to rank analyte performance, problems arise when P is used to rank overall laboratory performance. This is due to the fact that P is not an average (or a weighted average). Therefore, laboratories that perform concentration measurements on a different number of analytes cannot be compared. Hence, ranking these laboratories is impossible. This problem can be solved by averaging the overall performance measure. That is, an overall average performance measure can be defined as $\frac{1}{HIJ}P$ (where H is the number of analytes, I is the number of vials and J is the number of replicate measurements). A performance goal that limits the overall average performance measure can be defined as $\frac{1}{HIJ}P_g$ (where P_g, defined in (4.5), is the performance goal for P). A laboratory with overall average performance measure less than the performance goal is considered to have acceptable overall performance. Similarly, a laboratory with overall average performance measure greater than the performance goal is considered to have unacceptable overall performance. Now, the overall average performance measure can be used to rank either analyte performance or overall laboratory performance.

5.6 Conclusions and Recommendations

In this thesis we have investigated several methods of constructing the acceptable concentration interval. We have considered three standard methods of defining the target (i.e. the peer group mean approach, the all-group mean approach and the reference method approach) and four common methods of deriving the limit (i.e. the reference interval approach, the opinion of clinicians approach, the state of the art approach and the biological variation approach).

Ideally, the target should represent chemical truth. Unfortunately, chemical truth is not reflected by using the peer group mean to define the target. One major problem with the peer group mean approach is that the target varies depending on the peer group. This problem is solved by using a fixed target set by an accepted reference method. Chemical truth is established by using a reference method target. If the use of reference methods is impractical, the target can be defined by the all-group mean. For many analytes, the all-group mean target correlates closely with the reference method target.

A clinically relevant limit should be used to construct the acceptable concentration interval. Unfortunately, a limit based on either reference intervals or opinions of clinicians does not always reflect clinical relevance. These two approaches often produce very lenient limits for many analytes and, hence, no distinction is made between laboratories with poor analytical performance and laboratories with superior analytical performance. The reference interval approach and the opinion of clinicians approach are also too subjective and impractical for clinical use. The state of the art approach and the biological variation approach are the most commonly used methods of deriving the limit. However, the state of the art approach does not always produce a clinically relevant limit. For many analytes, the state of the art limit significantly varies depending on the peer group. This results in a strict evaluation for some laboratories and a lenient evaluation for others. In contrast, the biological variation approach does produce a clinically relevant limit, since well researched biological variation data is used to define the limit. The biological variation approach tends to produce strict limits for most analytes. For some analytes, these limits may be considered too strict. In these cases, the limits

should be viewed as goals for desirable analytical performance.

We have also proposed three measures of overall laboratory performance (i.e. the overall performance measure, the overall average error measure and the overall weighted error measure). These measures can be used to assess laboratory performance across vials (i.e. analyte performance) or laboratory performance across analytes (i.e. overall performance). The derivation of the overall performance measure is based on distribution assumptions. Unfortunately, these assumptions may not always hold. A practical alternative to using the overall performance measure is either using the overall average error measure or using the overall weighted error measure. These two simple measures are not based on distribution assumptions.

The C.R.L. currently evaluates only the vial performance of laboratories participating in the **CHEM**$^{\textbf{PLUS}}$ program. Laboratory concentration measurements are compared to the acceptable concentration interval. For laboratories located within British Columbia or Alberta, the acceptable concentration interval is defined by a peer group mean target and a limit based on state of the art. For all other laboratories, the acceptable concentration interval is defined by an all-group mean target and a limit based on biological variation. Currently, the C.R.L. does not assess analyte performance or overall laboratory performance.

The C.R.L. should continue to use the acceptable concentration interval to evaluate vial performance. However, a standardized acceptable concentration interval should be used for all

laboratories participating in the **CHEM**$^{\textbf{PLUS}}$ program. This acceptable concentration interval is best constructed by using a reference method target and a limit based on biological variation. If the use of reference methods is impractical, the target should be defined by the all-group mean. The C.R.L. should also incorporate either the overall average error measure or the overall weighted error measure (or a combination of the two measures) to assess analyte performance and overall laboratory performance. These two measures are easy to compute, simple to interpret and provide valuable information about laboratory performance. In addition, these measures can be used to rank the laboratories participating in the **CHEM**$^{\textbf{PLUS}}$ program.

Bibliography

1. Rippey JH, Williamson WE. The overall role of a proficiency testing program. Arch Pathol Lab Med 1988;112:340-342.

2. Sunderman FW. The history of proficiency testing / quality control. Clin Chem 1992;38:1205-1209.

3. Ladenson JH. Is anyone regulating the regulators? Clin Chem 1993;39:375-376.

4. Young DS. Determination and validation of reference intervals. Arch Pathol Lab Med 1992;116:704-709.

5. Kazmierczak SC, Catrou PG. Laboratory error undetectable by customary quality control/quality assurance monitors. Arch Pathol Lab Med 1993;117:714-718.

6. Ehrmeyer SS, Laessig RH, Cembrowski GS. Performance of external quality control systems. Laboratory Medicine 1989;20:428-431.

7. Gräsbech R. Reference values, why and how. Scand J Clin Lab Invest 1990;50,Suppl 201:45-53.

8. Sasse EA. Determination of reference intervals in the clinical laboratory using the proposed guideline national committee for clinical laboratory standards C28-P. Arch Pathol Lab Med 1992;116:710-713.

9. Statland BE. Clinical decision levels for lab tests, 2nd ed. Oradell: Medical Economics Books, 1987:224pp.

10. Rej R. Accurate enzyme activity measurements. Arch Pathol Lab Med 1993;117:352-364.

11. Tietz NW, Rodgerson DO, Laessig RH. Are clinical laboratory proficiency tests as good as they can be? Clin Chem 1992;38:473-475.

12. Franzini C. The reference method value. Ann Ist Super Sanità 1991;27:359-364.

13. Stamm D. A new concept for quality control of clinical laboratory investigations in the light of clinical requirements and based on reference method values. J Clin Chem Clin Biochem 1982;20:817-824.

14. Bowers GN. Clinical chemistry analyte reference systems based on true value. Clin Chem 1991;37:1665-1666.

15. Hartmann AE, Naito HK, Burnett RW, Welch MJ. Accuracy of participant results utilized as target values in the CAP chemistry survey program. Arch Pathol Lab Med 1985;109:894-903.

16. Rej R, Jenny RW. How good are clinical laboratories? An assessment of current performance. Clin Chem 1992;38:1210-1225.

17. Jansen RTP, Bullock DG, Vassault A, Baadenhuijsen H, De Leenheer A, Dumont G, De Verdier CH, Zender R. Between-country comparability of clinical chemistry results: an international quality assessment survey of 17 analytes in six European countries through existing national schemes. Ann Clin Biochem 1993;30:304-314.

18. Fraser CG. The application of theoretical goals based on biological variation data in proficiency testing. Arch Pathol Lab Med 1988;112:404-415.

19. Fraser CG. Desirable performance standards for clinical chemistry tests. Advances in Clinical Chemistry 1983;23:299-339.

20. Fraser CG, Petersen PH. Quality goals in external quality assessment are best based on biology. Scand J Clin Lab Invest 1993;53:8-9.

21. Gowans EMS, Petersen PH, Blabjerg O, Hørder M. Analytical goals for the acceptance of common reference intervals for laboratories throughout a geographical area. Scand J Clin Lab Invest 1988;48:757-764.

22. Fraser CG. Biological variation in clinical chemistry. Arch Pathol Lab Med 1992;116:916-923.

23. Harris EK. Proposed goals for analytical precision and accuracy in single-point diagnostic testing. Arch Pathol Lab Med 1988;112:416-420.

24. Fraser CG, Petersen PH, Larsen ML. Setting analytical goals for random analytical error in specific clinical monitoring situations. Clin Chem 1990;36:1625-1628.

25. Tholen DW. A statistical procedure for measuring and evaluating performance in interlaboratory comparison programs. Arch Pathol Lab Med 1988;112:462-470.

26. Chinchilli VM, Miller WG. Evaluating test methods by estimating total error. Clin Chem 1994;40:464-471.

27. Ehrmeyer SS, Laessig RH. Alternative statistical approach to evaluating interlaboratory performance. Clin Chem 1985;31:106-108.

28. Jansen AP, Van Kampen EJ, Leijnse B, Meijers CAM, Van Munster PJJ. Experience in the Netherlands with an external quality control and scoring system for clinical chemistry laboratories. Clinica Chimica Acta 1977;74:191-201.

Appendix A

Table A.1 summarizes the January 1996 **CHEM**^{PLUS} shipment data used for our analysis of calcium, chloride, magnesium, potassium and sodium. Figures A.1 to A.5 display box plots of laboratory concentration measurements for calcium, chloride, magnesium, potassium and sodium, respectively. Figures A.6 to A.30 display various acceptable concentration intervals for the twenty five analyte/vial combinations. Note that, in figures A.6 to A.30, the horizontal dotted line represents the reference method target. Figures A.31 to A.35 show the overall agreement between the all-group mean target (as well as the peer group mean targets) and the reference method target for calcium, chloride, magnesium, potassium and sodium, respectively.

Table A.1: Summary of the January 1996 CHEM^{PLUS} Shipment Data						
Analyte	**Vial**	**Total Number of Labs**	**Outliers**	**Problem Measurements**	**Labs Outside of B.C. or Alberta**	**Number of Labs used in Analysis**
Calcium	Blue	153	2	0	4	147
	Green	153	1	0	4	148
	Yellow	153	2	0	4	147
	Orange	153	0	0	4	149
	Red	153	2	0	4	147
Chloride	Blue	189	2	5	4	178
	Green	189	1	3	4	181
	Yellow	189	2	3	4	180
	Orange	189	0	4	4	181
	Red	189	0	8	4	177
Magnesium	Blue	83	1	0	3	79
	Green	83	1	0	3	79
	Yellow	83	1	0	3	79
	Orange	83	0	0	3	80
	Red	83	3	2	2	76
Potassium	Blue	227	1	3	4	219
	Green	227	3	2	4	218
	Yellow	227	2	2	4	219
	Orange	227	1	6	4	216
	Red	227	0	9	4	214
Sodium	Blue	227	2	2	4	219
	Green	227	1	3	4	219
	Yellow	227	1	2	4	220
	Orange	227	0	6	4	217
	Red	227	0	9	4	214

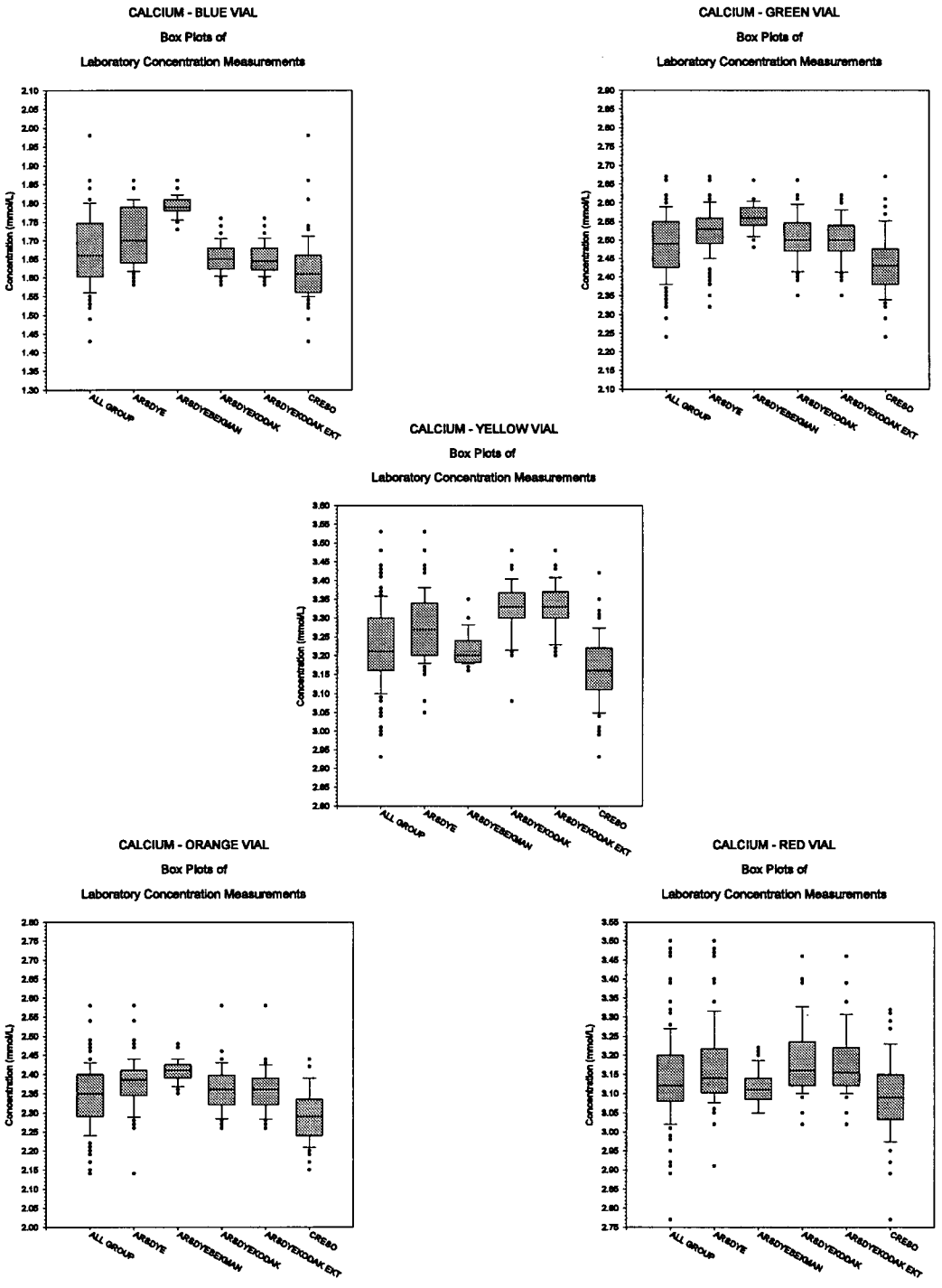


Figure A.1: Box Plots of Laboratory Concentration Measurements for Calcium

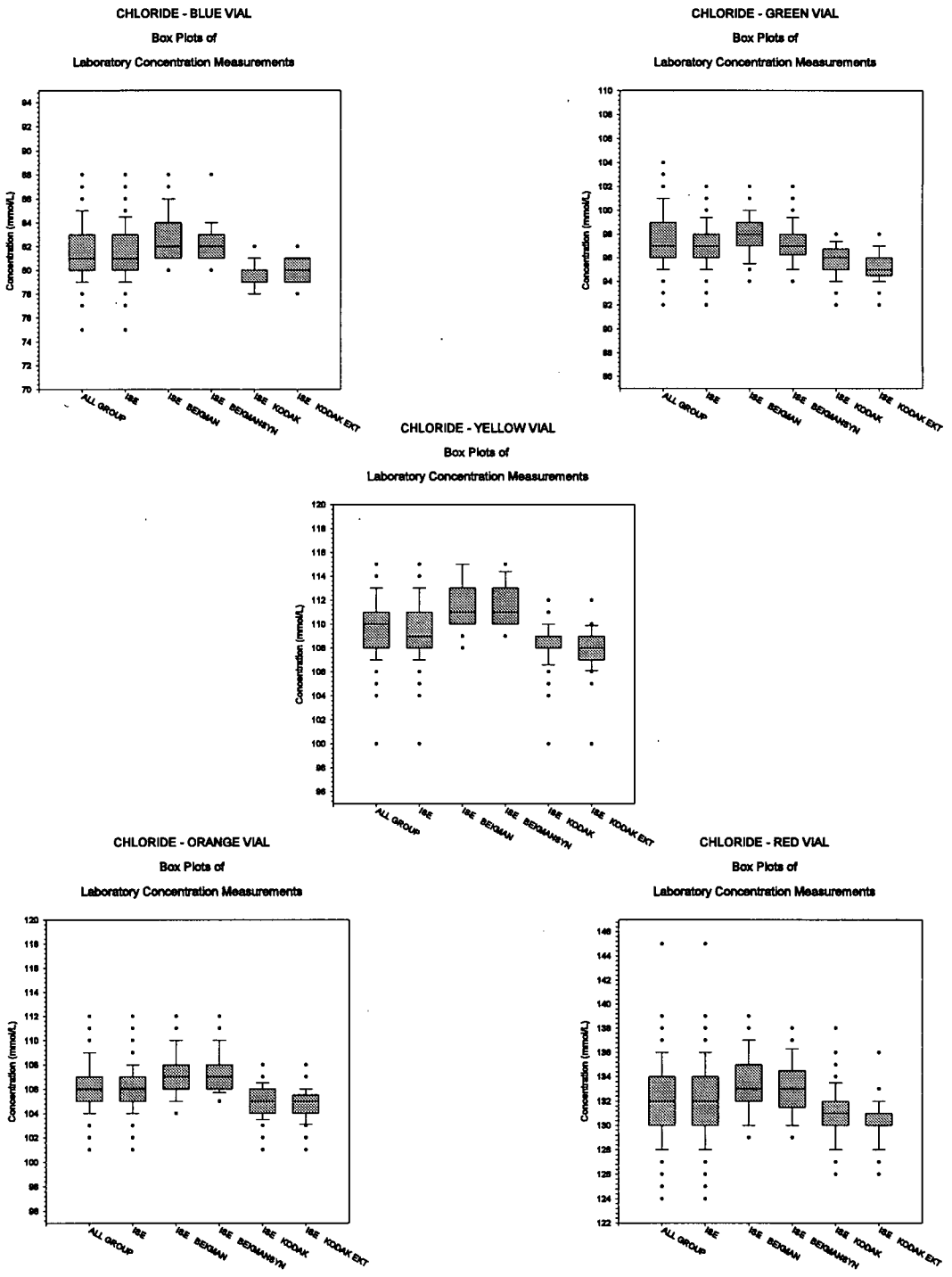


Figure A.2: Box Plots of Laboratory Concentration Measurements for Chloride

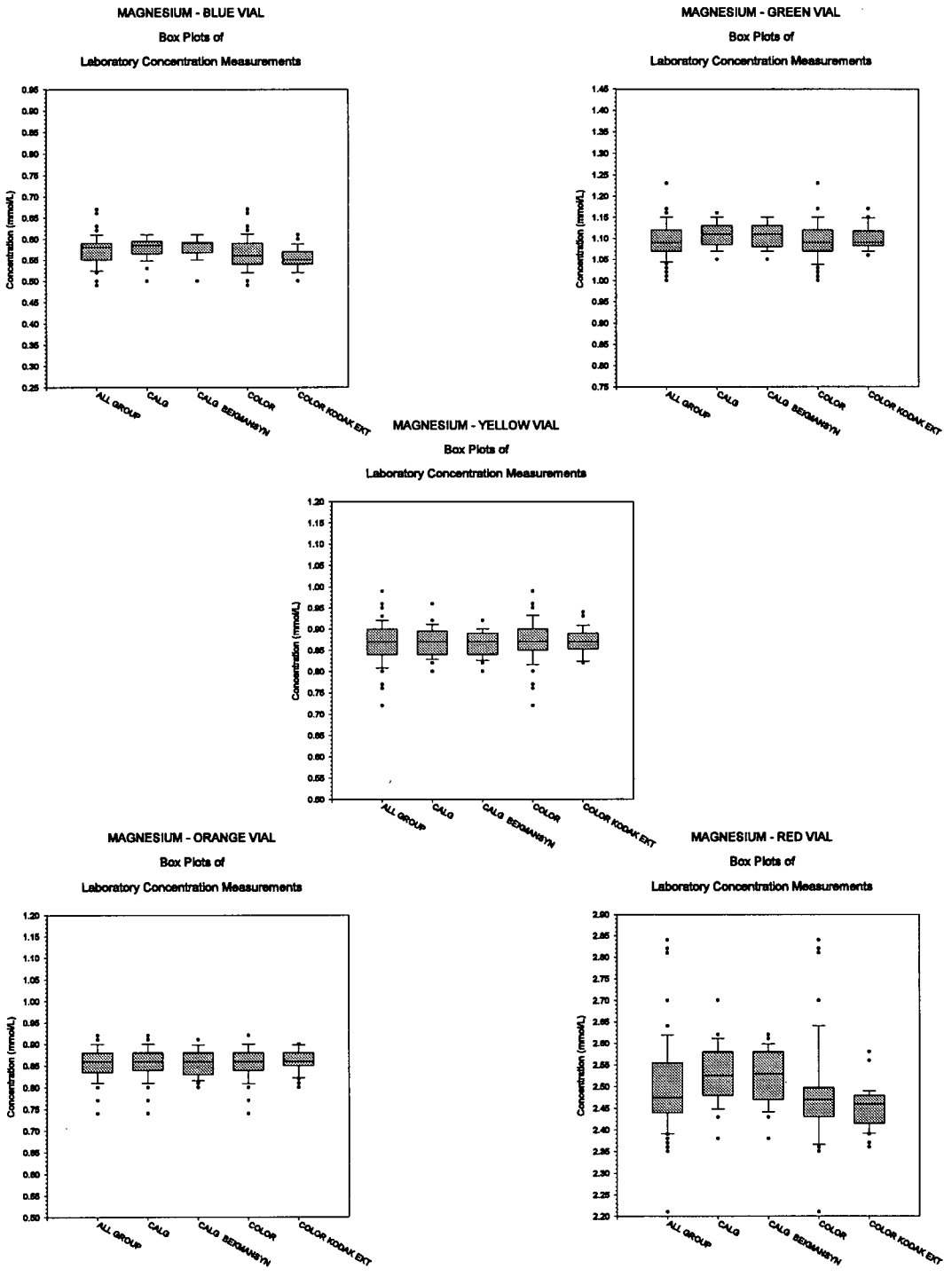


Figure A.3: Box Plots of Laboratory Concentration Measurements for Magnesium

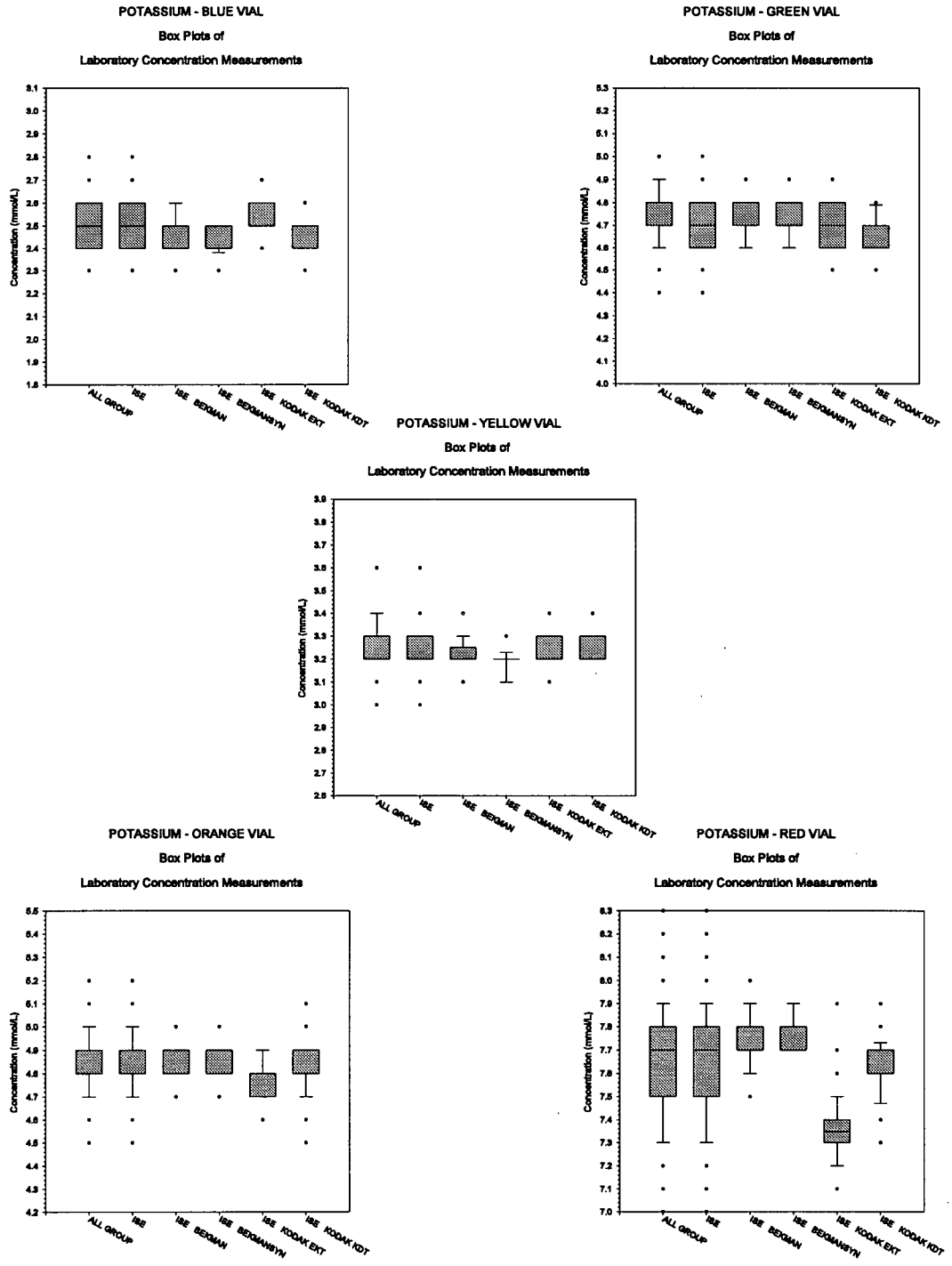


Figure A.4: Box Plots of Laboratory Concentration Measurements for Potassium

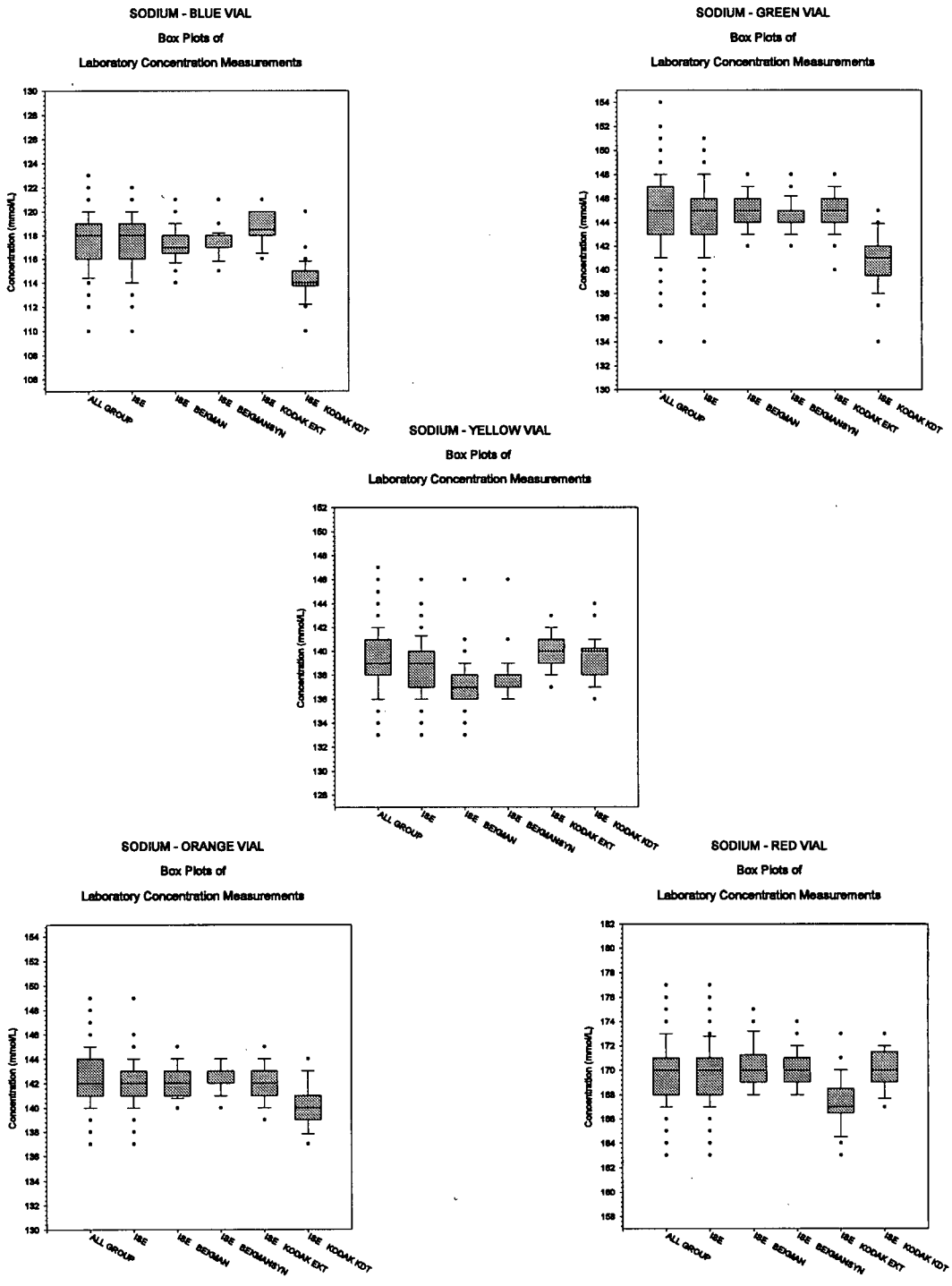


Figure A.5: Box Plots of Laboratory Concentration Measurements for Sodium

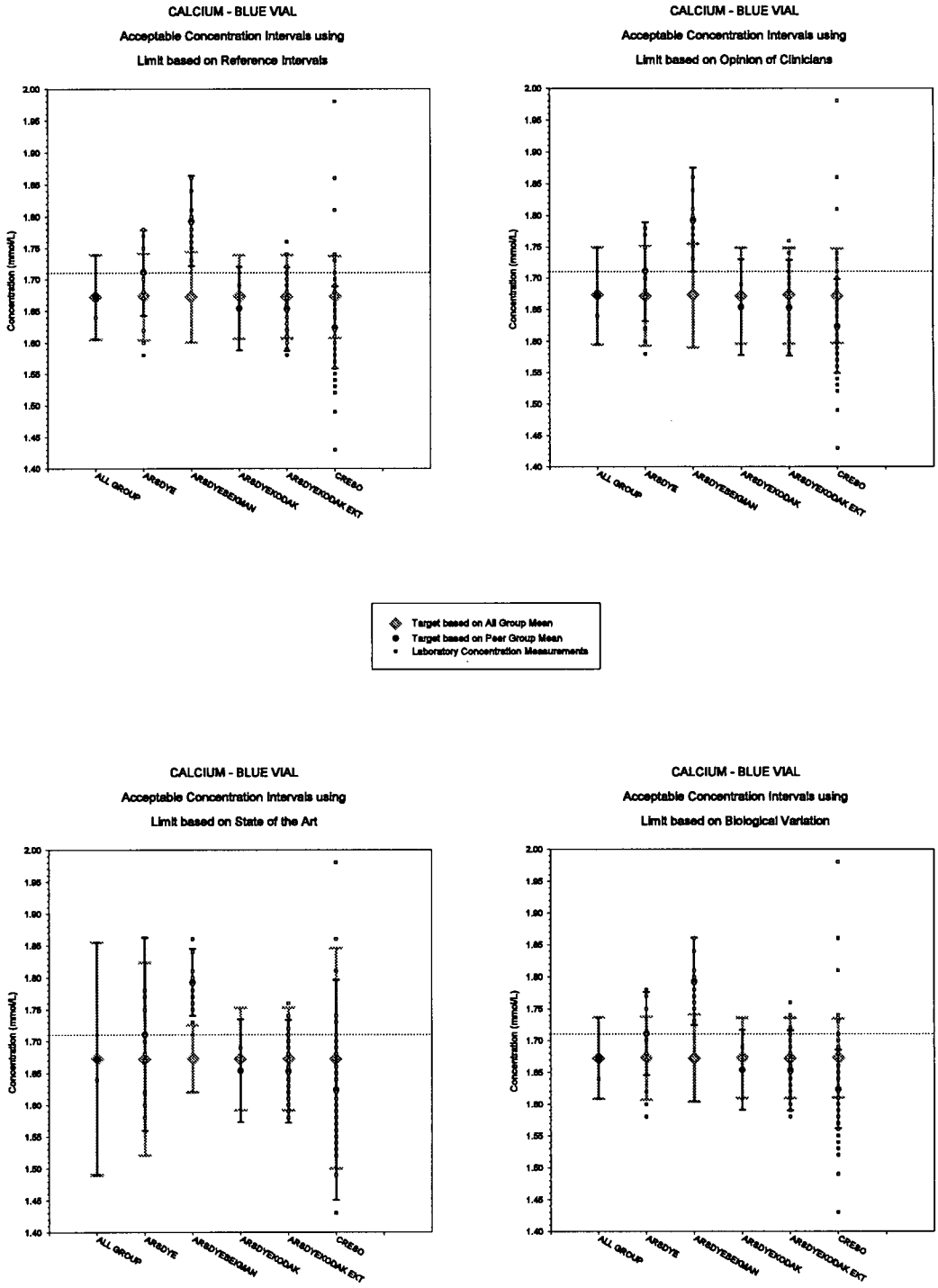


Figure A.6: Acceptable Concentration Intervals for the Blue Vial of Calcium

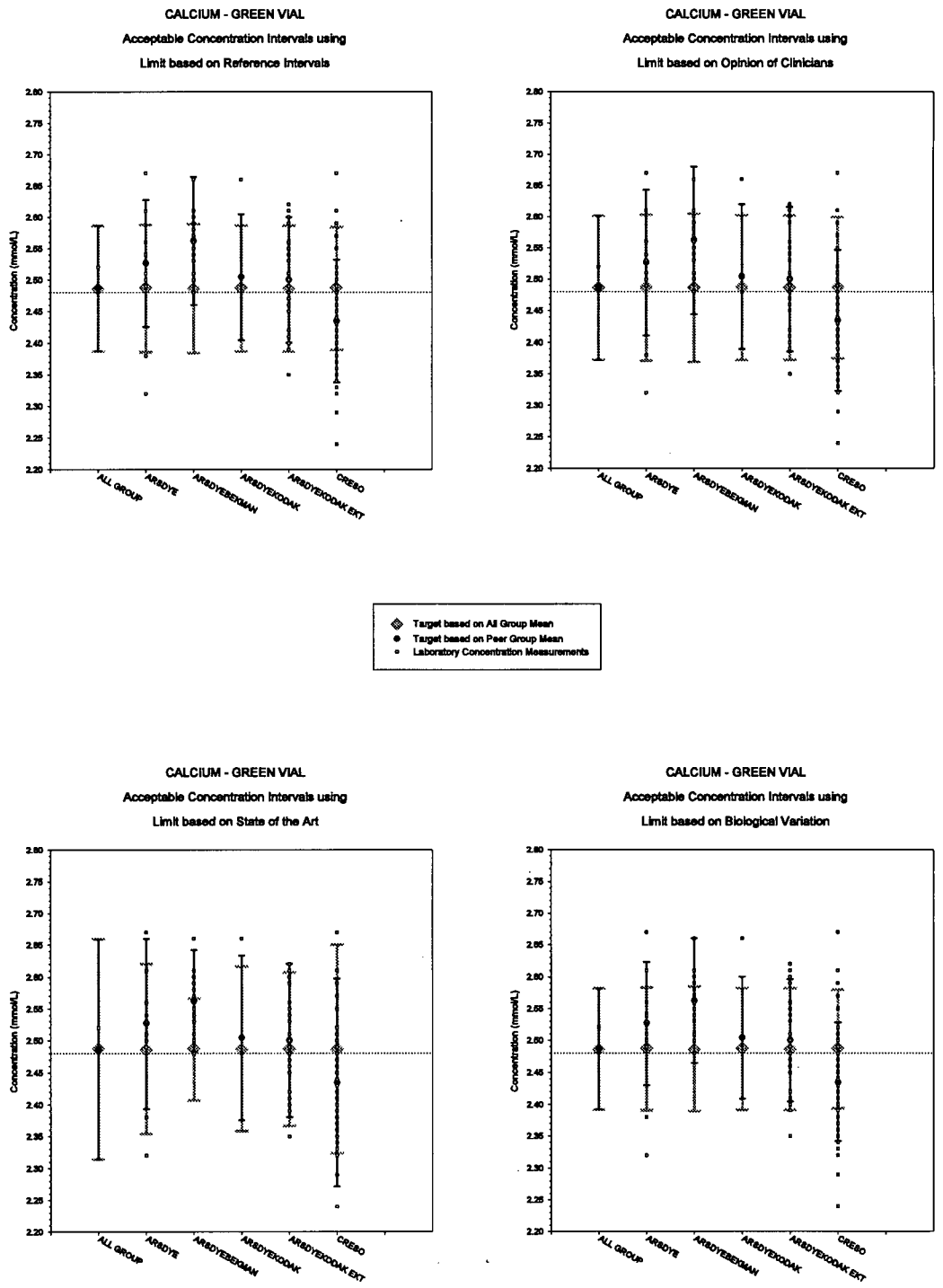


Figure A.7: Acceptable Concentration Intervals for the Green Vial of Calcium

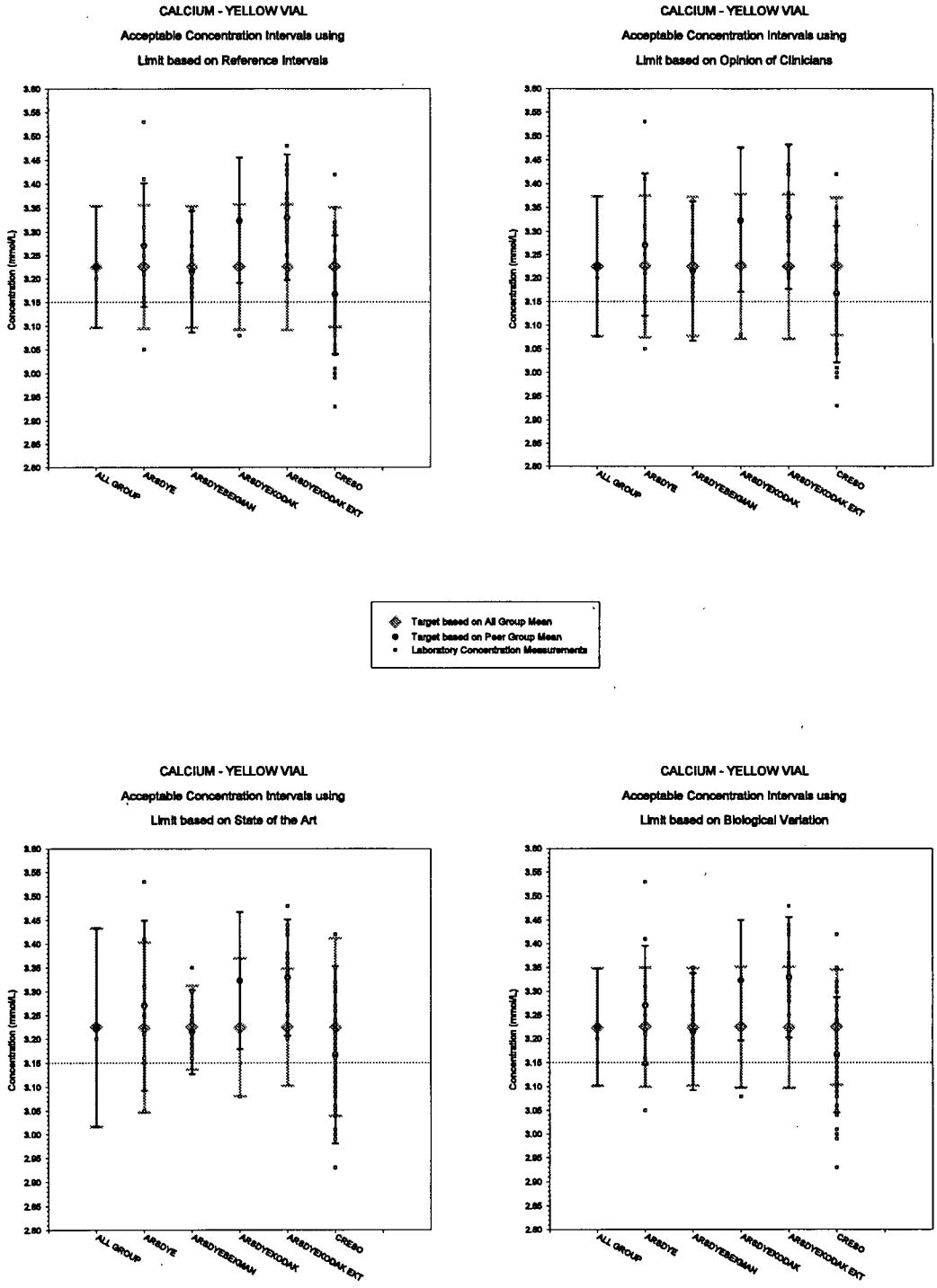


Figure A.8: Acceptable Concentration Intervals for the Yellow Vial of Calcium

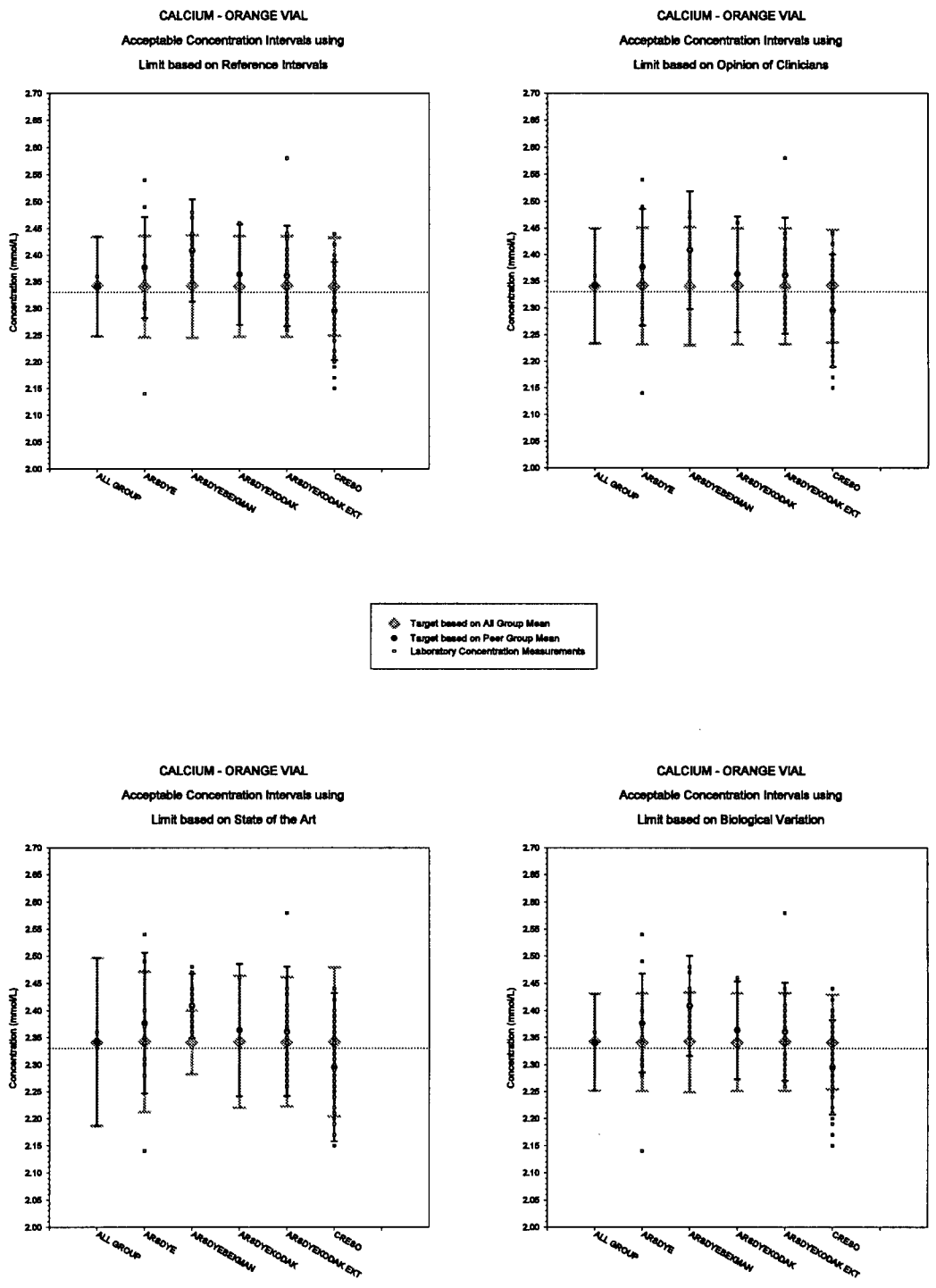


Figure A.9: Acceptable Concentration Intervals for the Orange Vial of Calcium

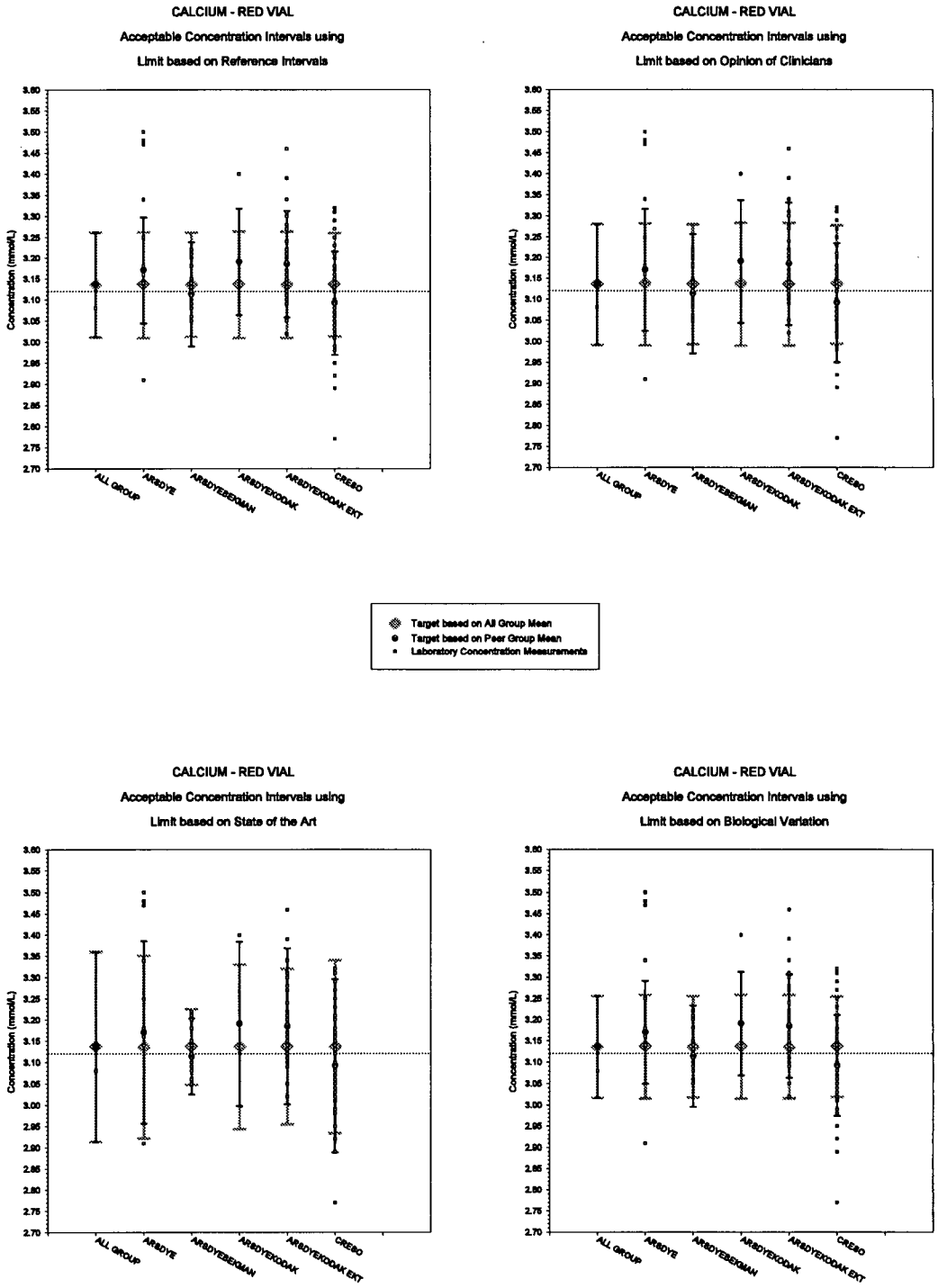


Figure A.10: Acceptable Concentration Intervals for the Red Vial of Calcium

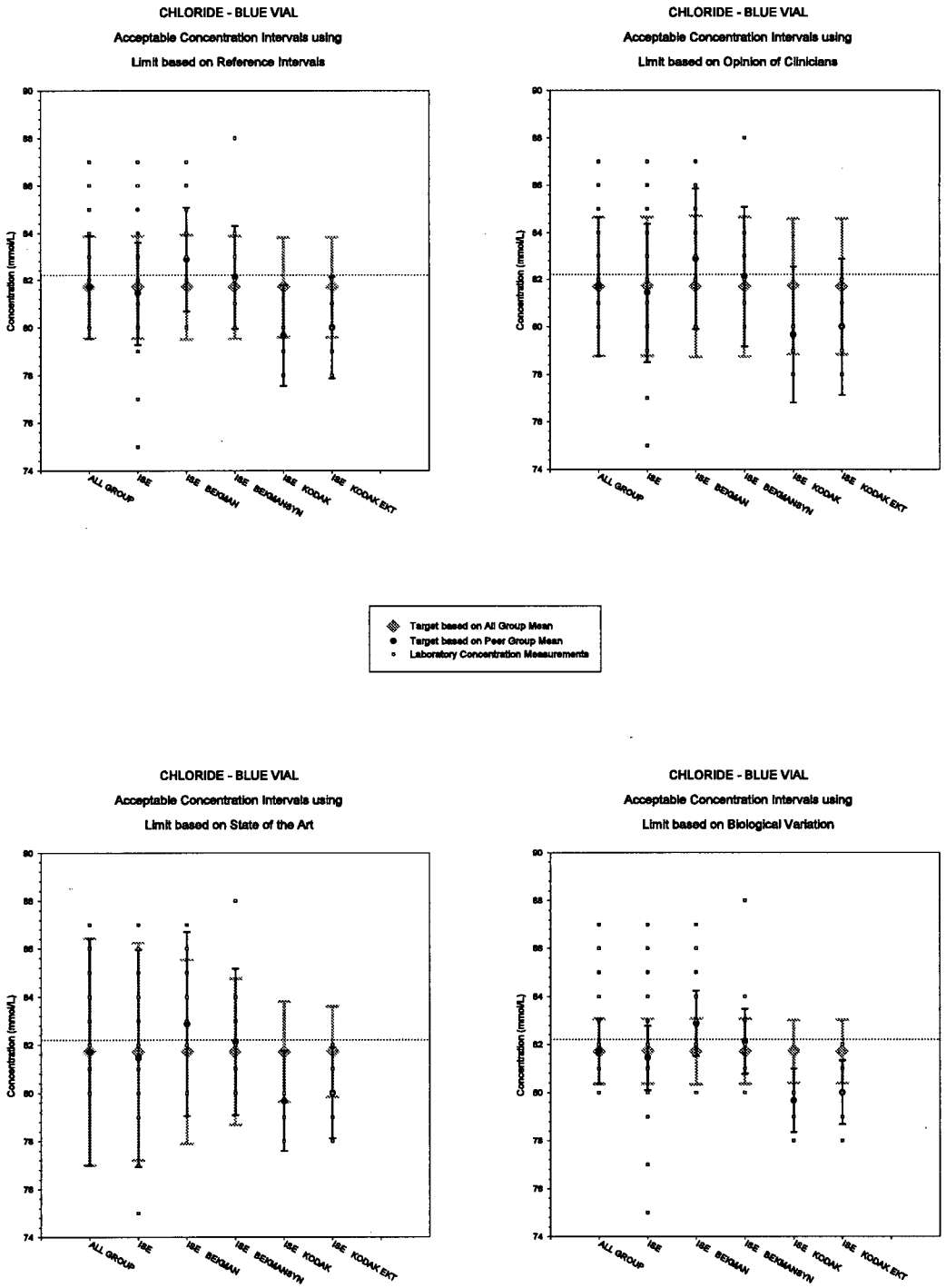


Figure A.11: Acceptable Concentration Intervals for the Blue Vial of Chloride

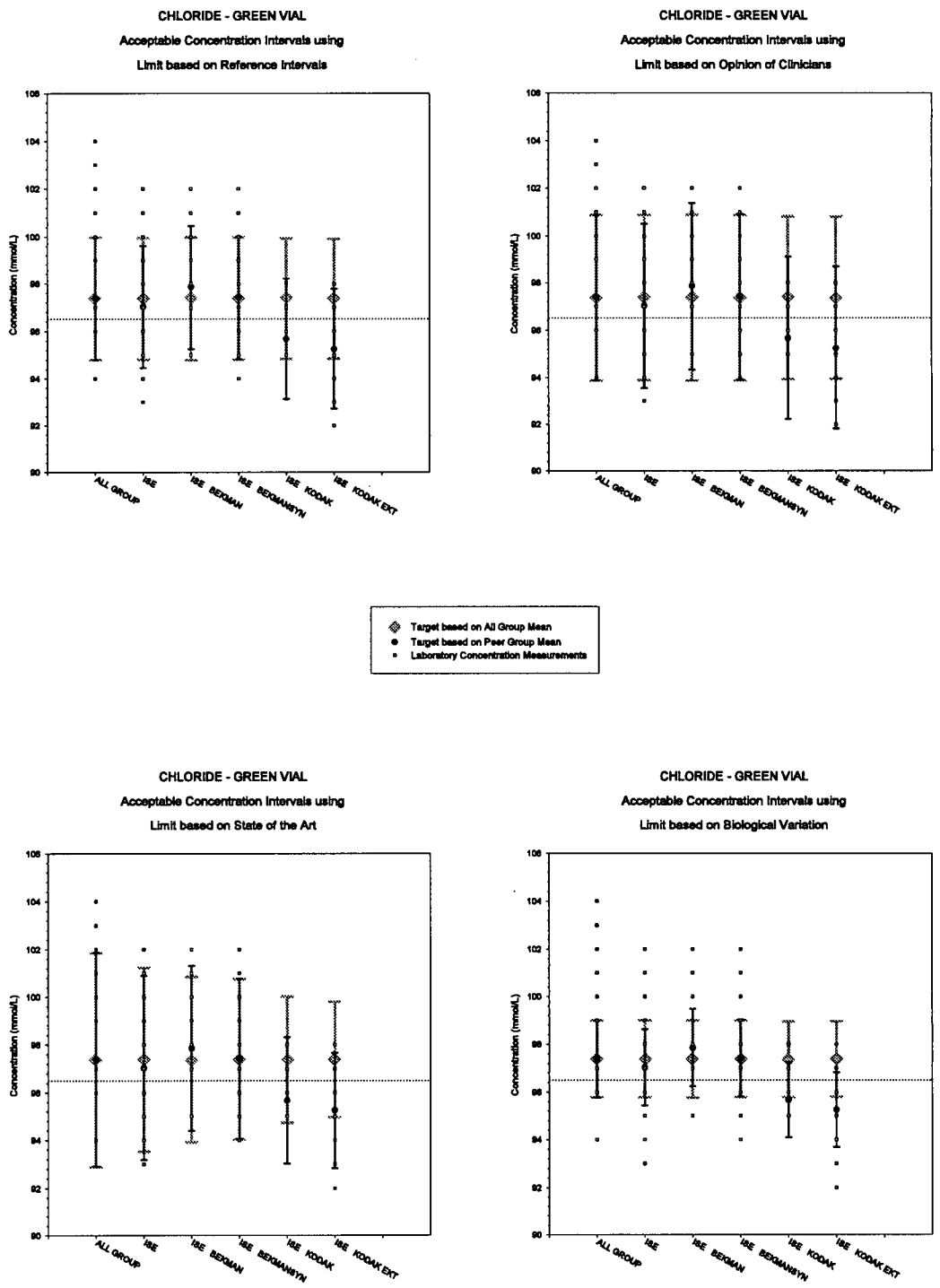


Figure A.12: Acceptable Concentration Intervals for the Green Vial of Chloride

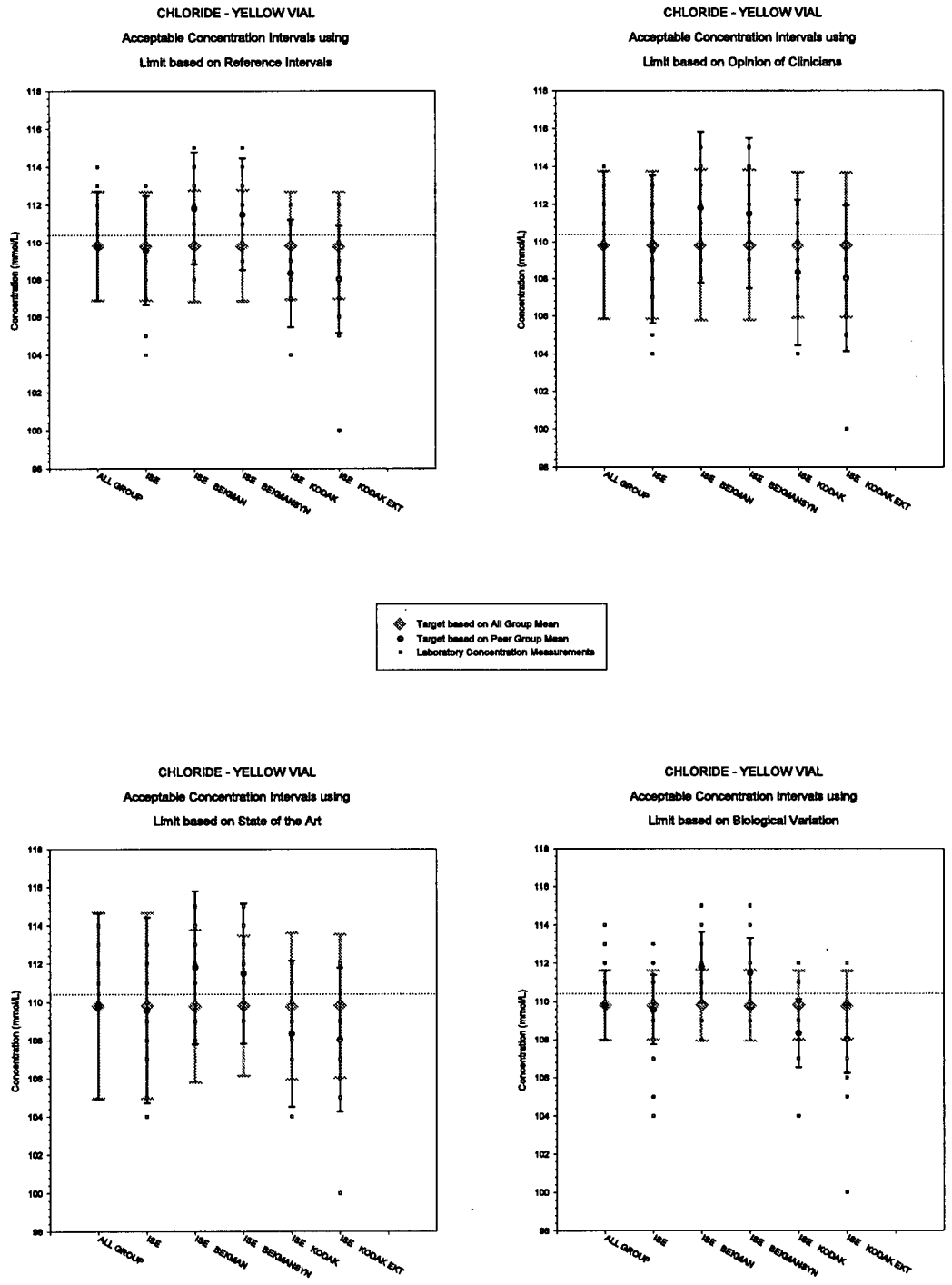


Figure A.13: Acceptable Concentration Intervals for the Yellow Vial of Chloride

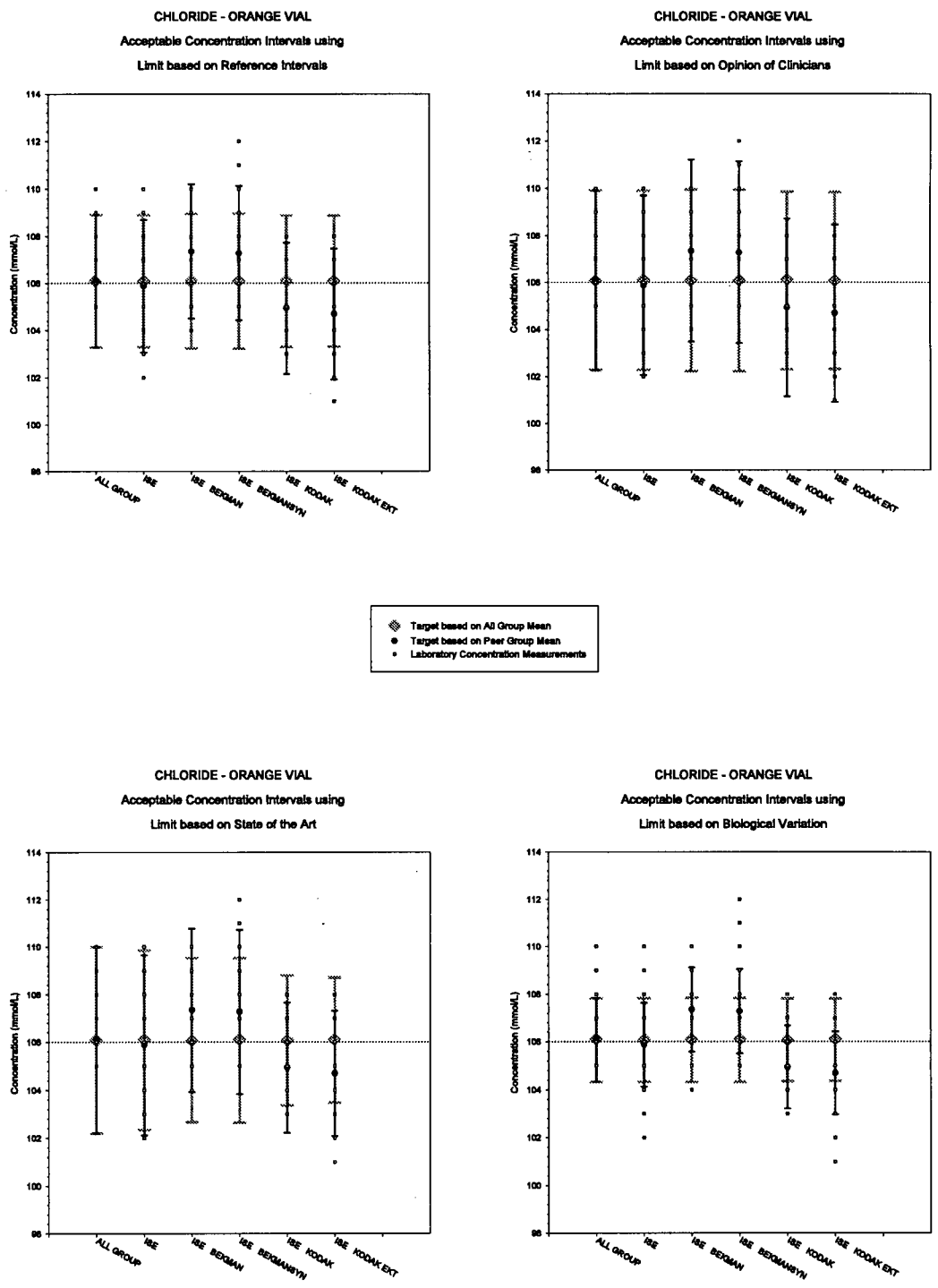


Figure A.14: Acceptable Concentration Intervals for the Orange Vial of Chloride

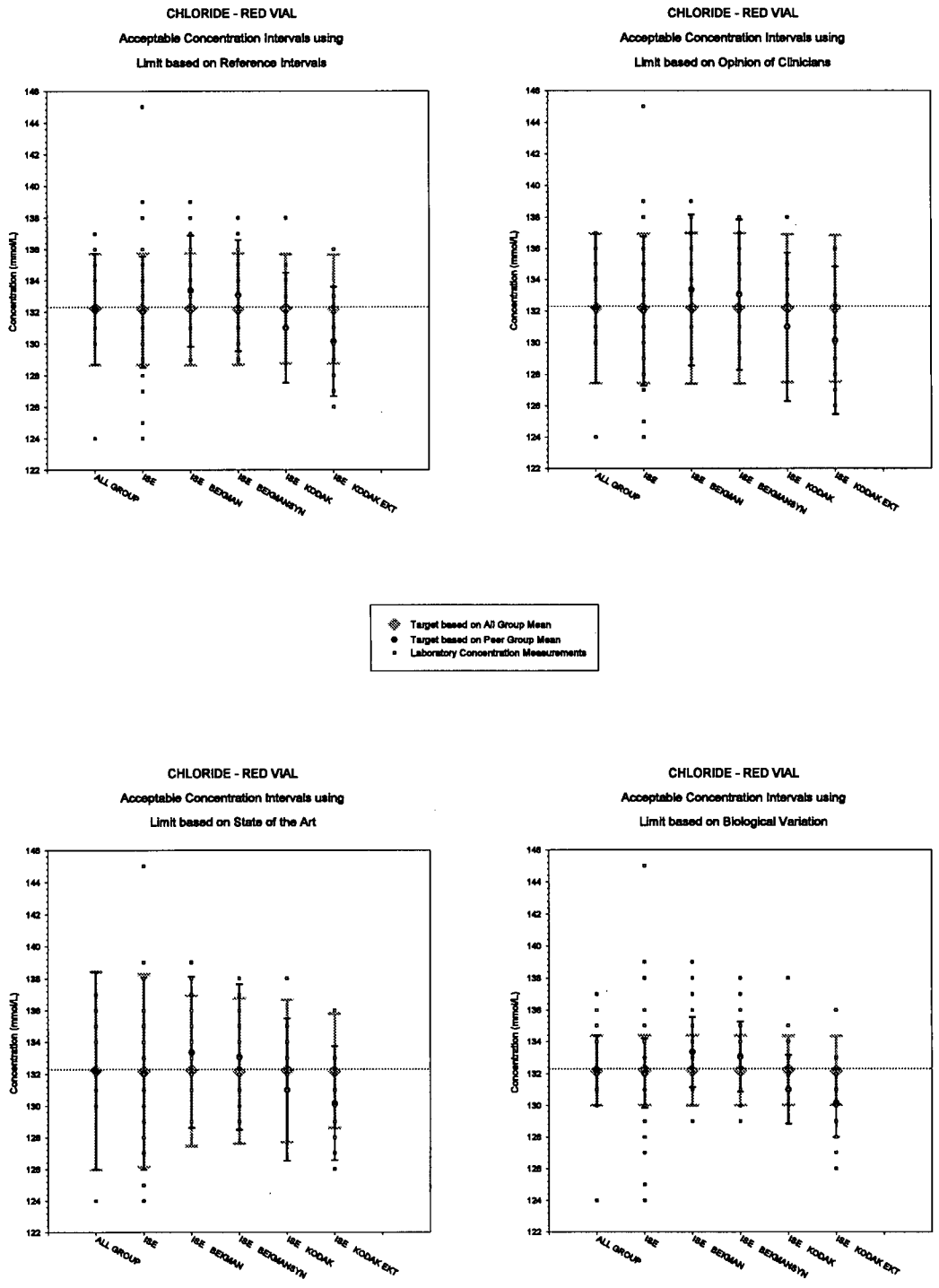


Figure A.15: Acceptable Concentration Intervals for the Red Vial of Chloride

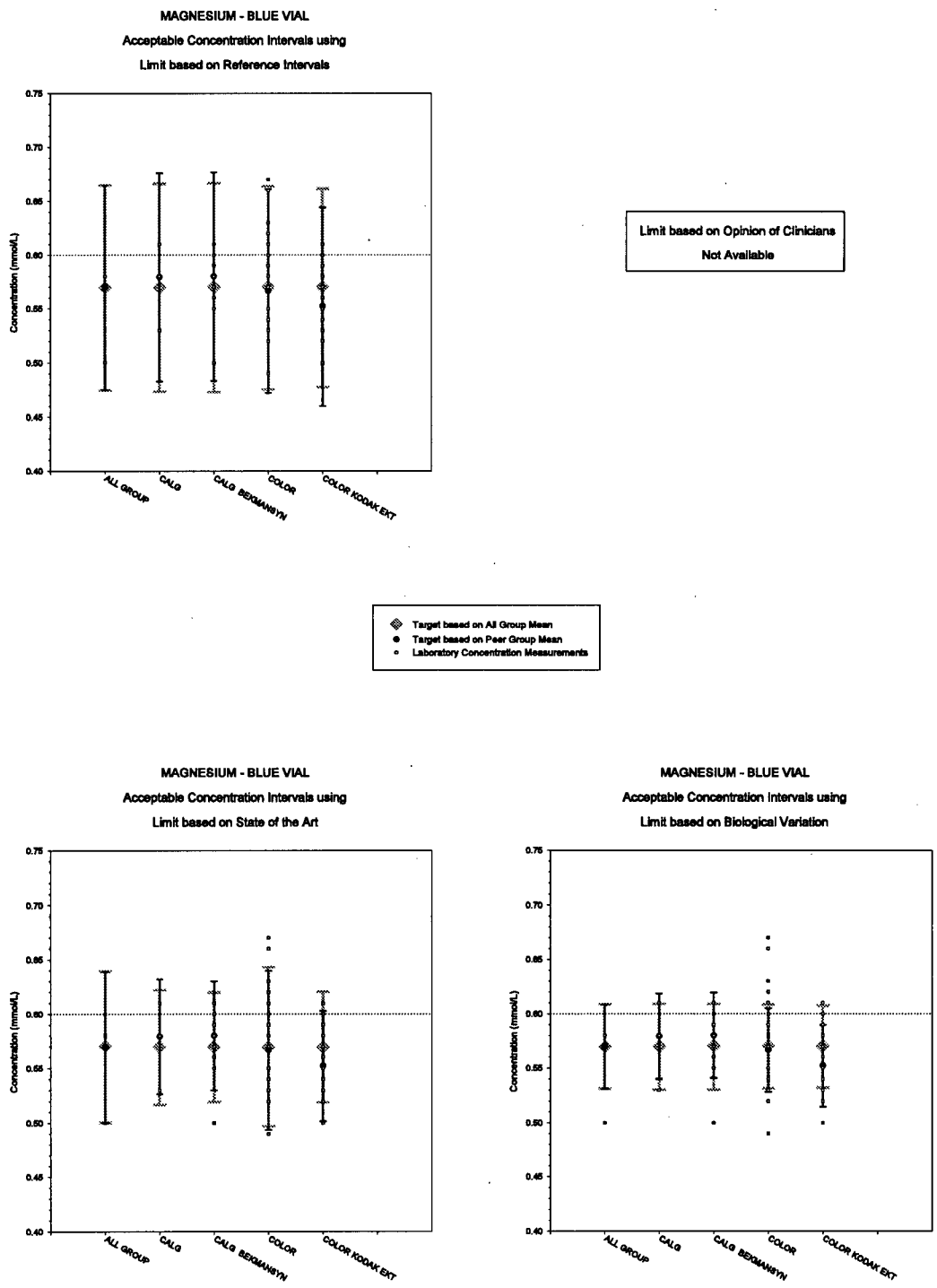


Figure A.16: Acceptable Concentration Intervals for the Blue Vial of Magnesium

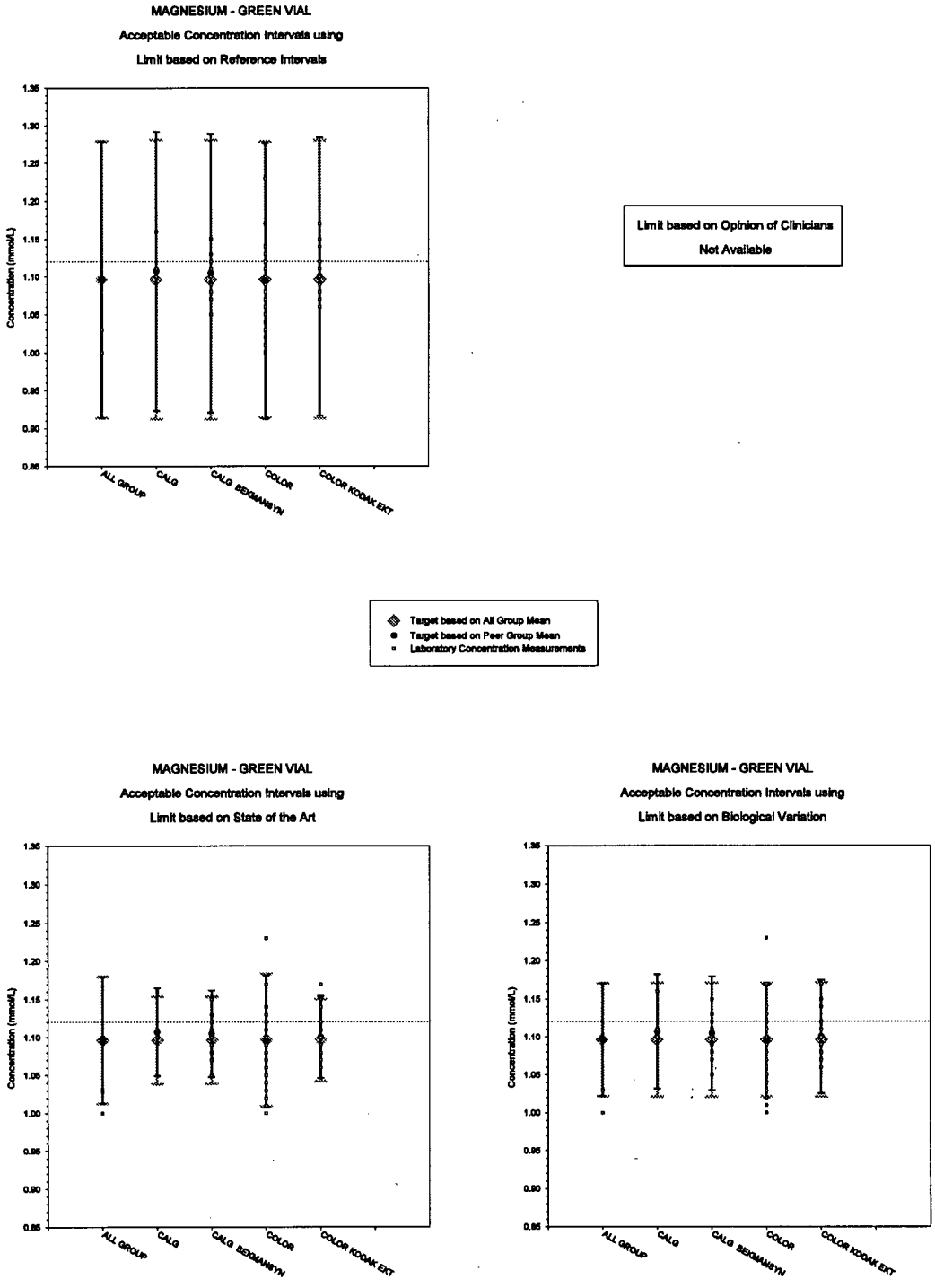


Figure A.17: Acceptable Concentration Intervals for the Green Vial of Magnesium

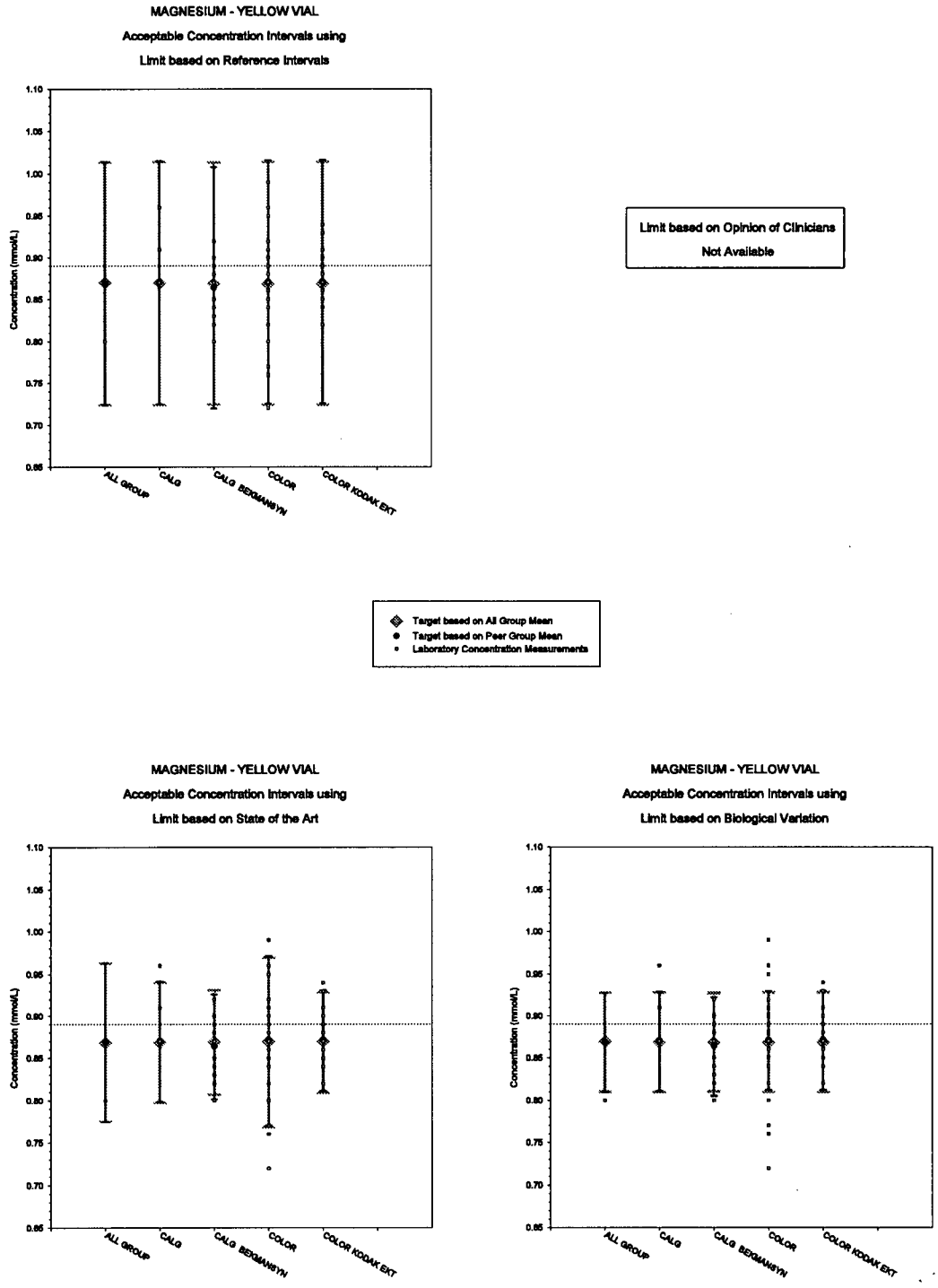


Figure A.18: Acceptable Concentration Intervals for the Yellow Vial of Magnesium

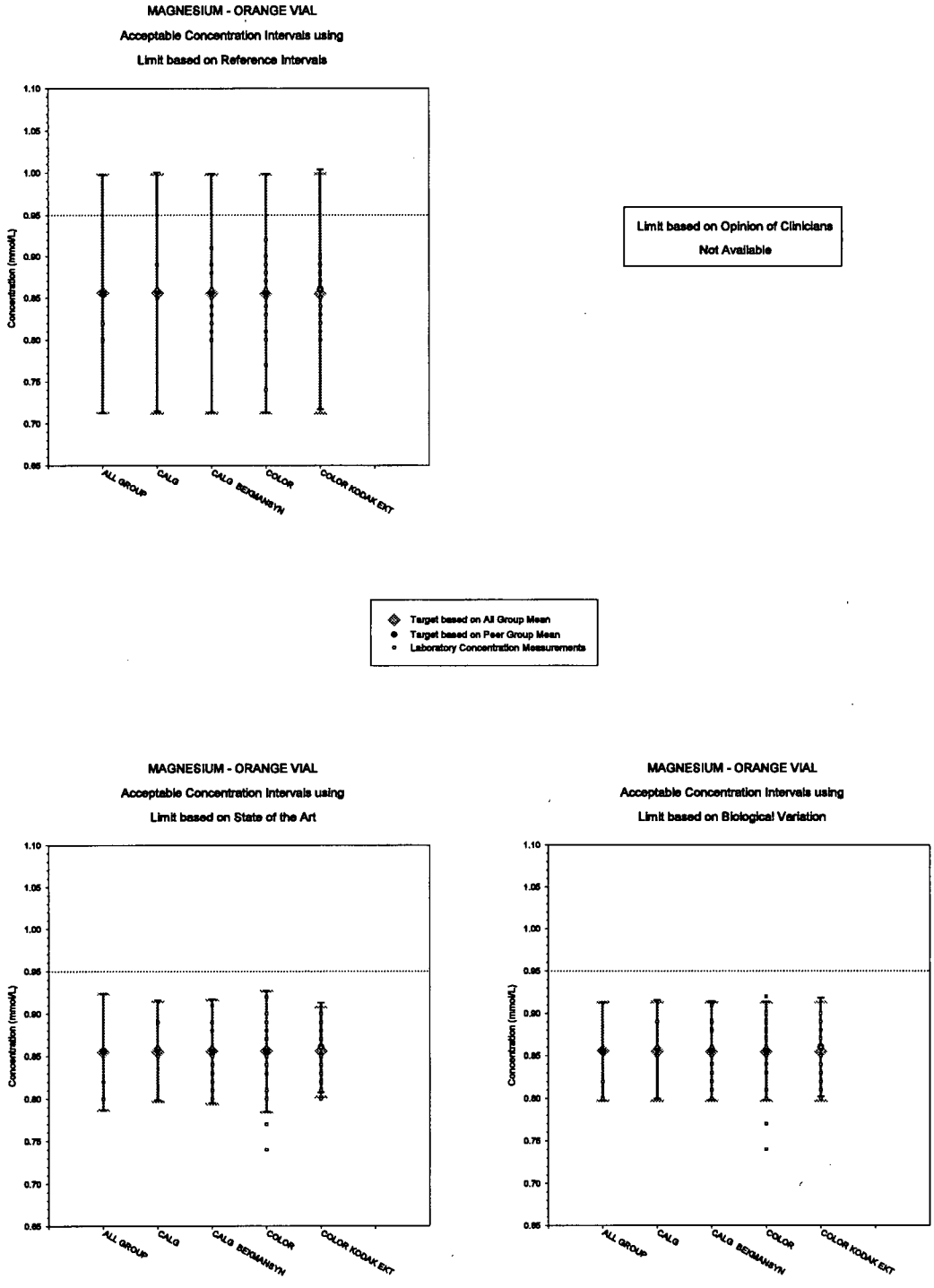


Figure A.19: Acceptable Concentration Intervals for the Orange Vial of Magnesium

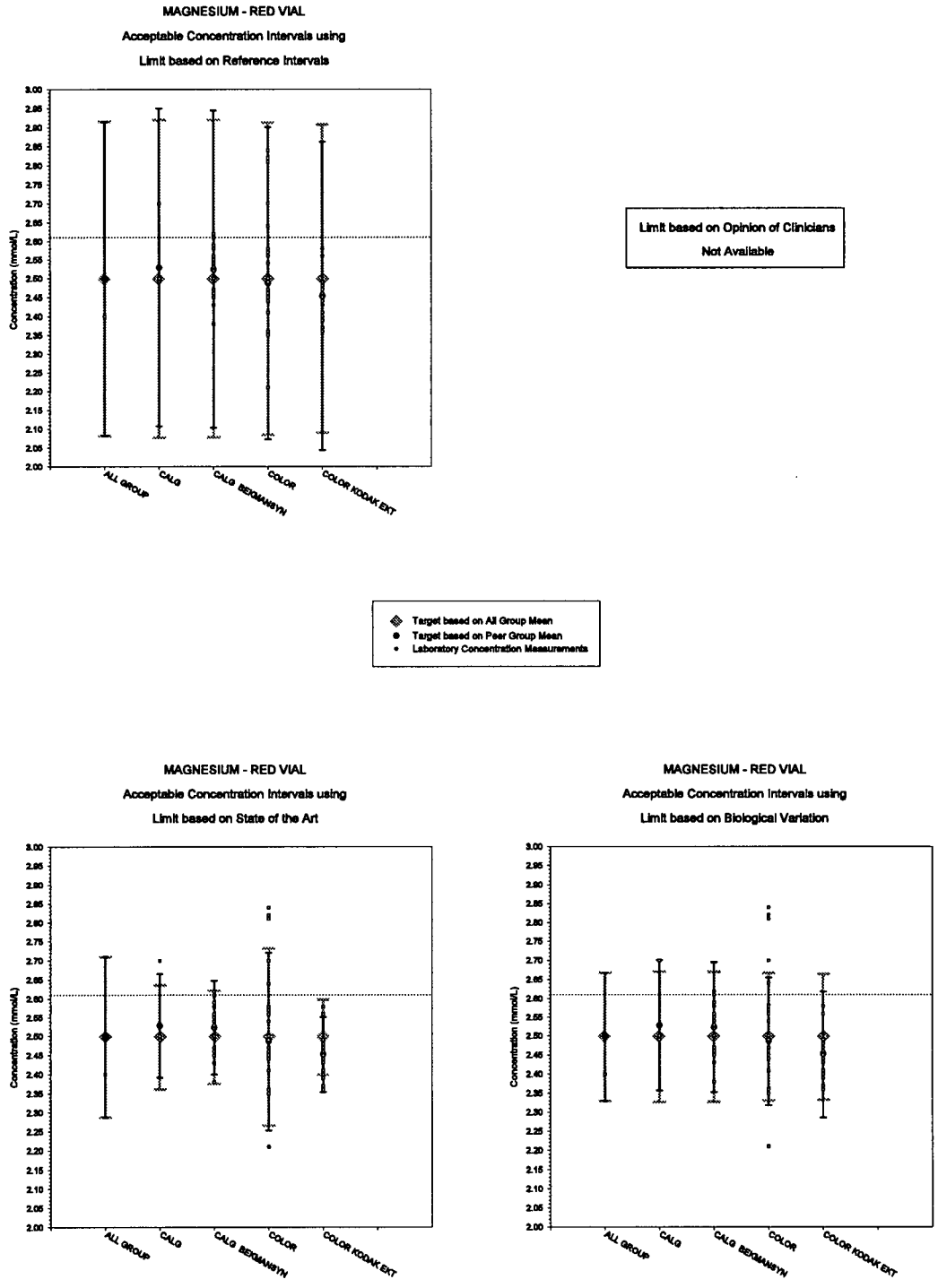


Figure A.20: Acceptable Concentration Intervals for the Red Vial of Magnesium

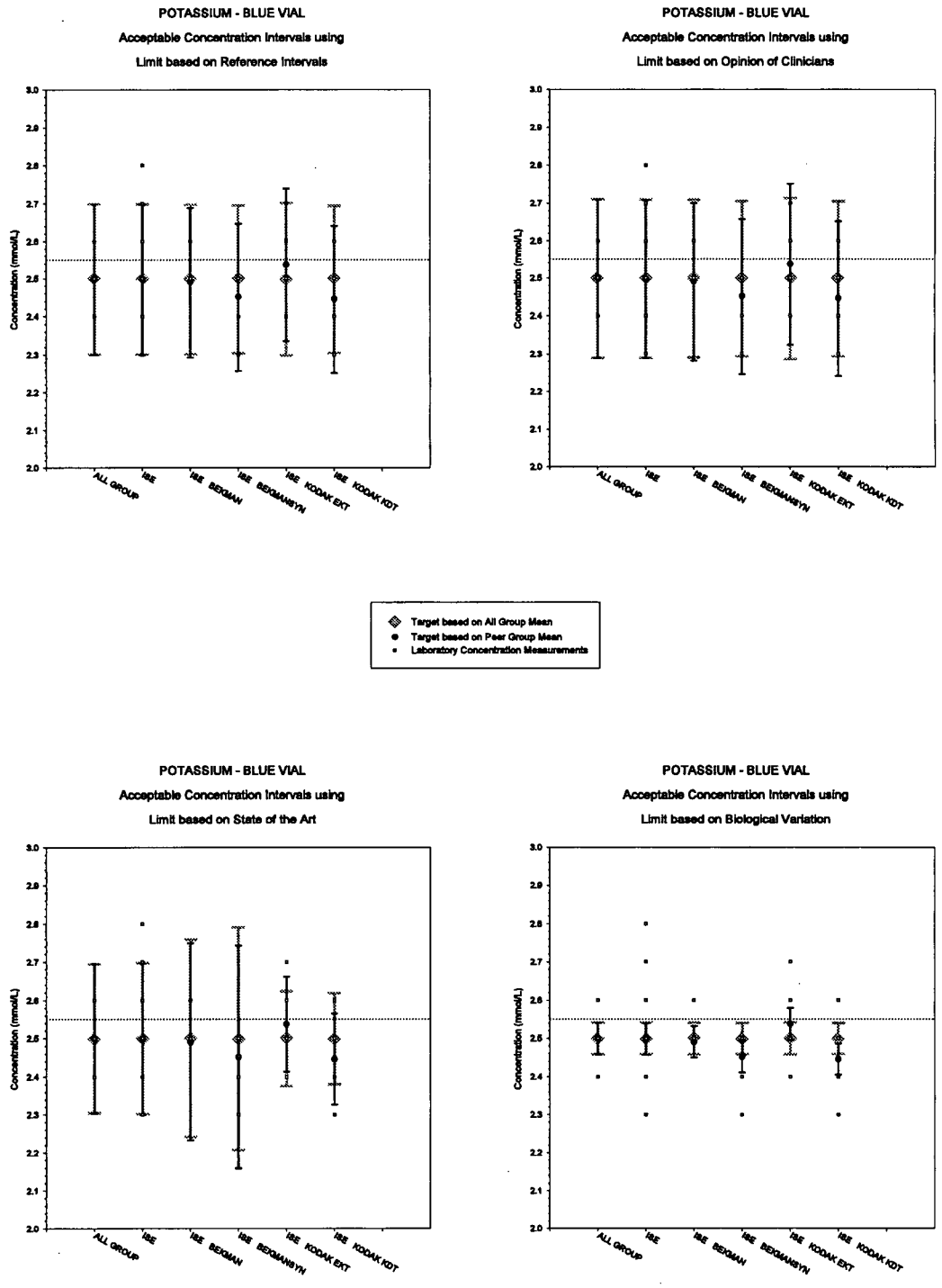


Figure A.21: Acceptable Concentration Intervals for the Blue Vial of Potassium

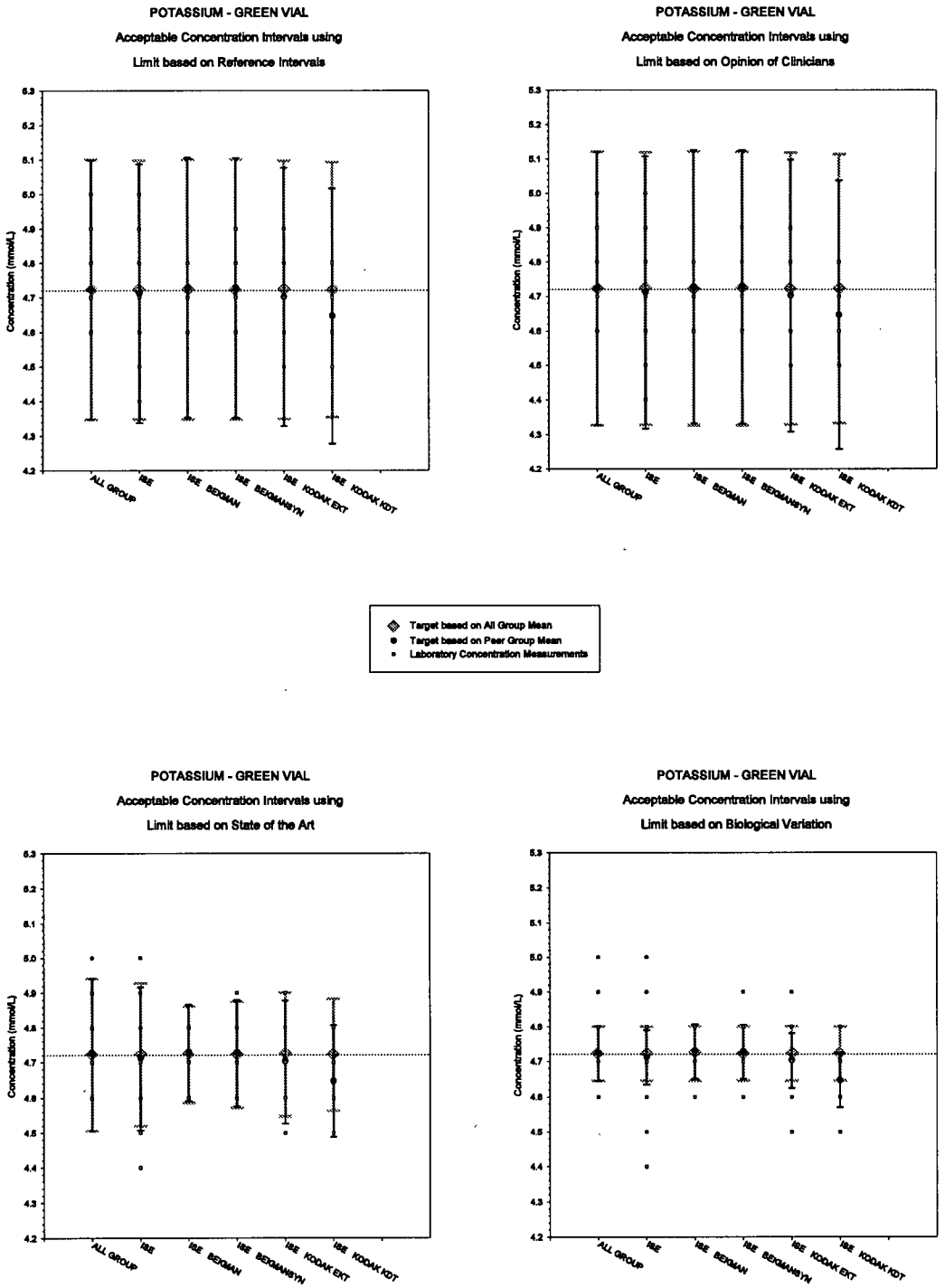


Figure A.22: Acceptable Concentration Intervals for the Green Vial of Potassium

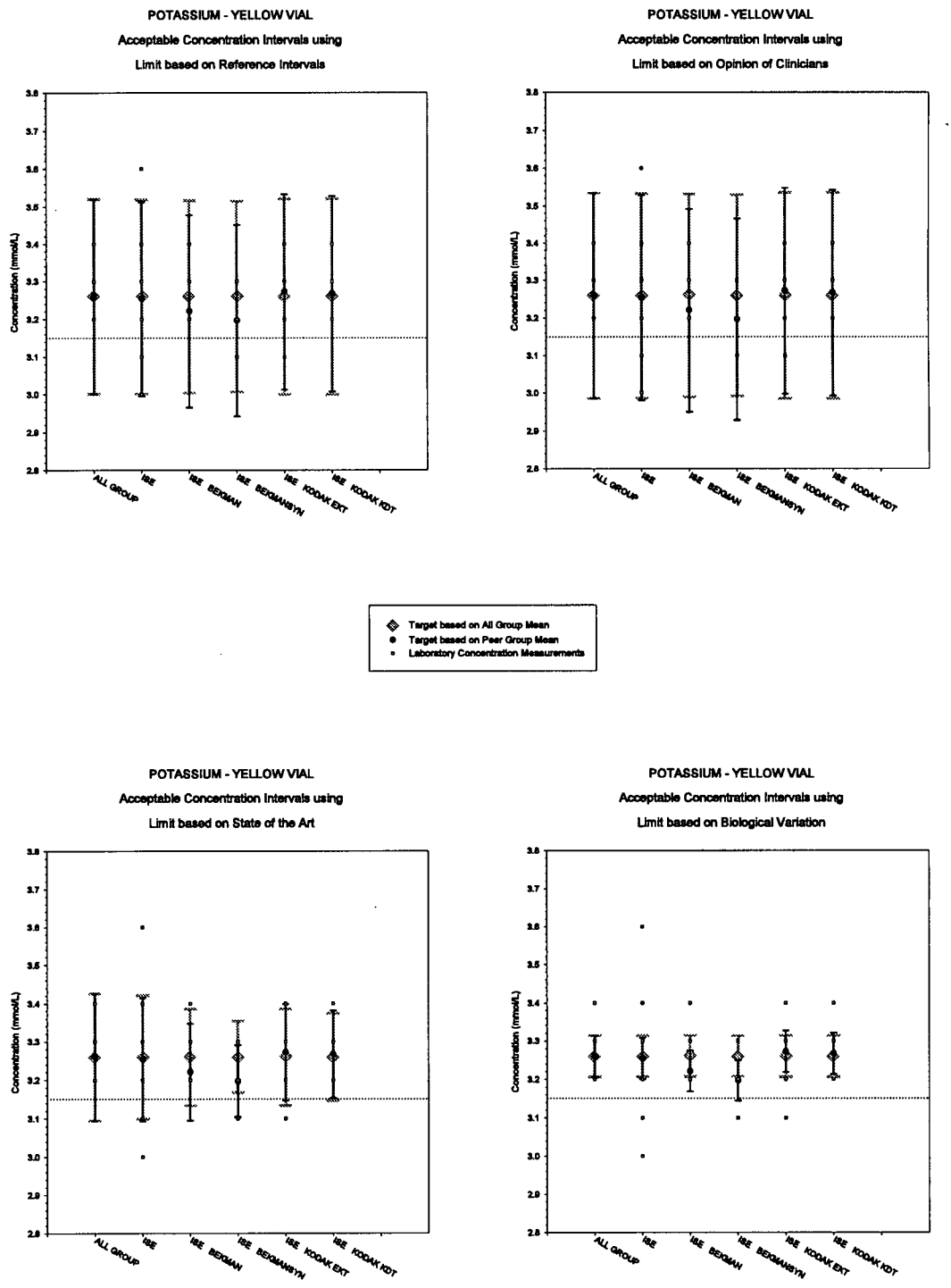


Figure A.23: Acceptable Concentration Intervals for the Yellow Vial of Potassium

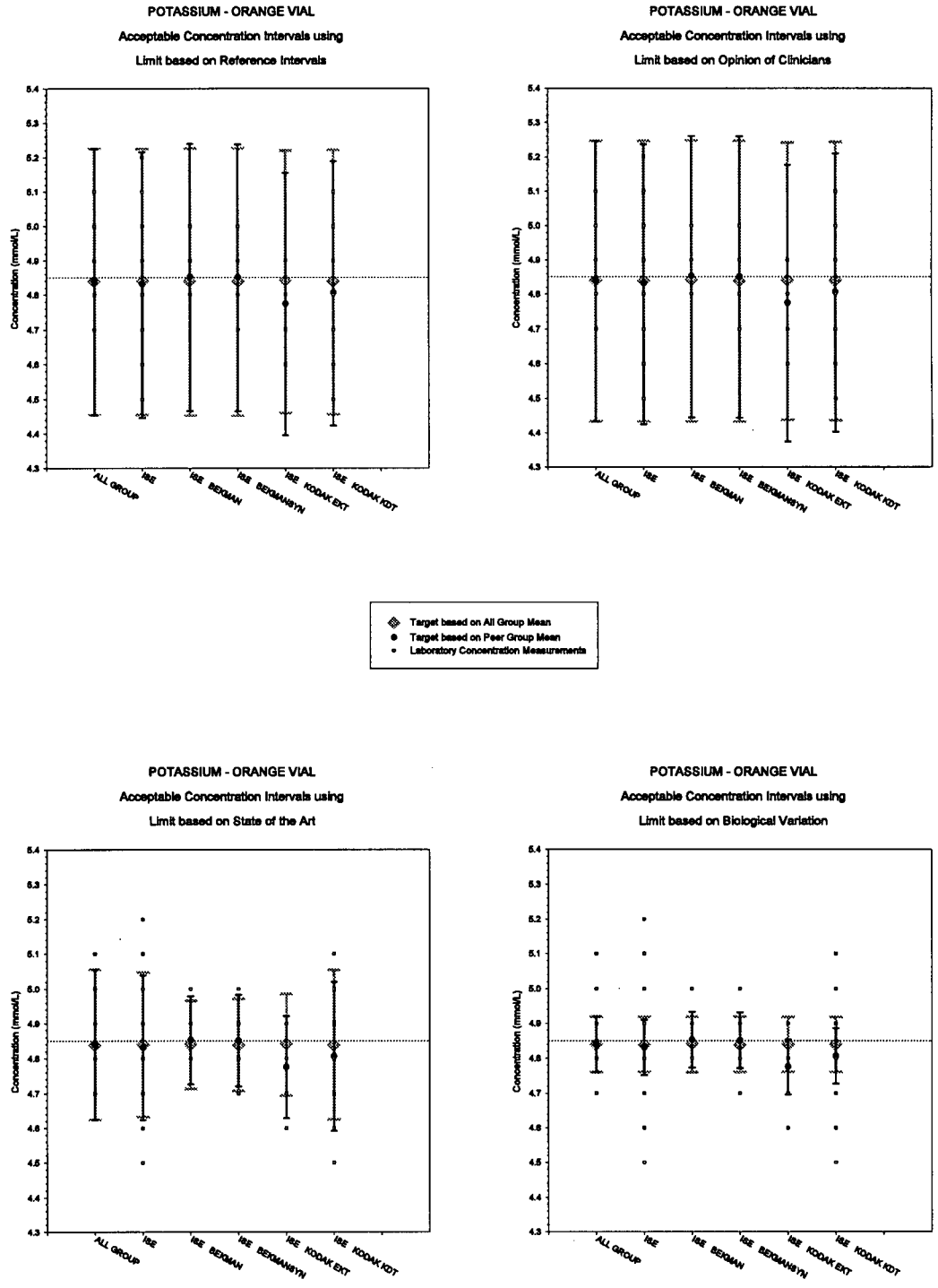


Figure A.24: Acceptable Concentration Intervals for the Orange Vial of Potassium

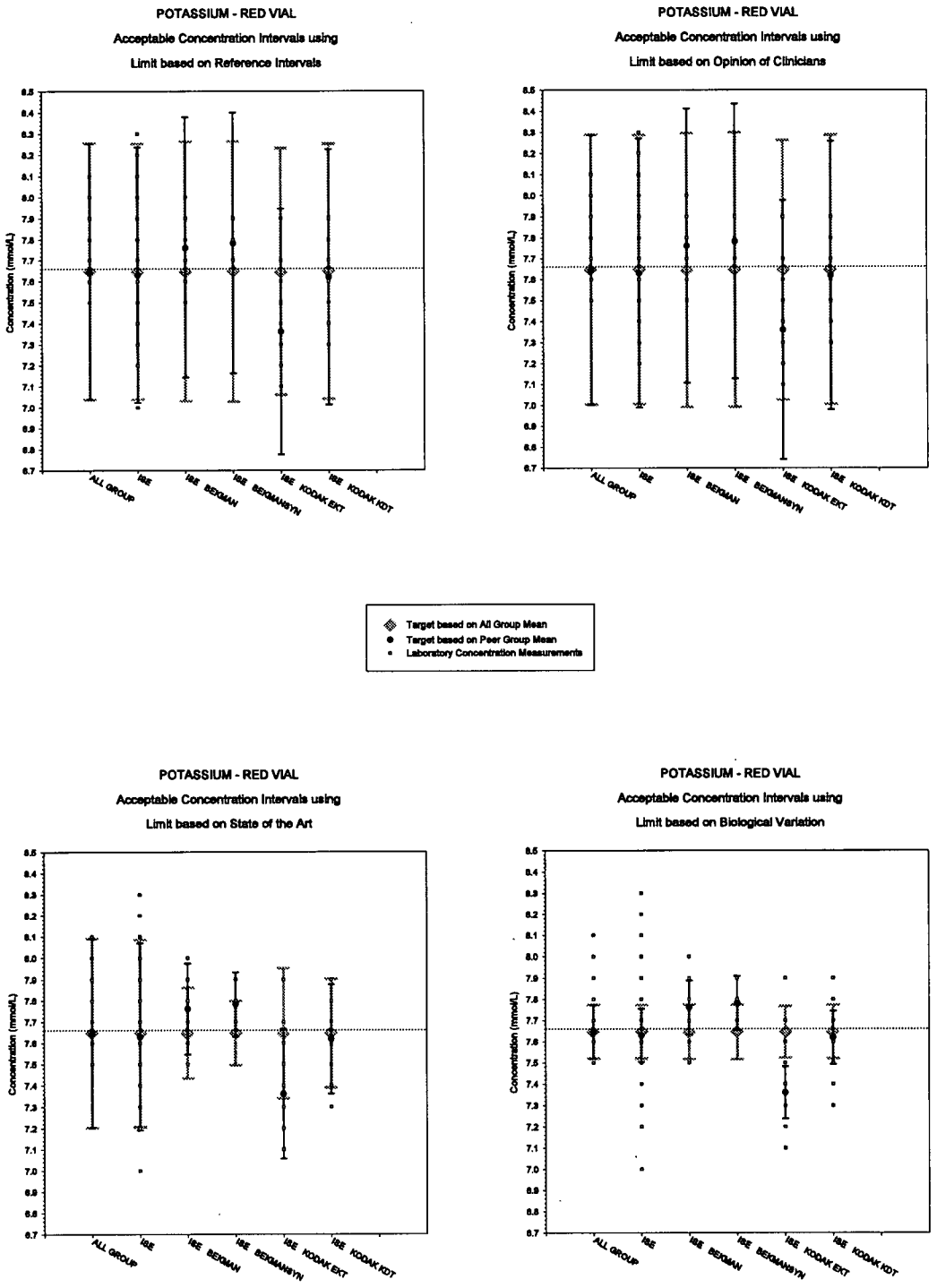


Figure A.25: Acceptable Concentration Intervals for the Red Vial of Potassium

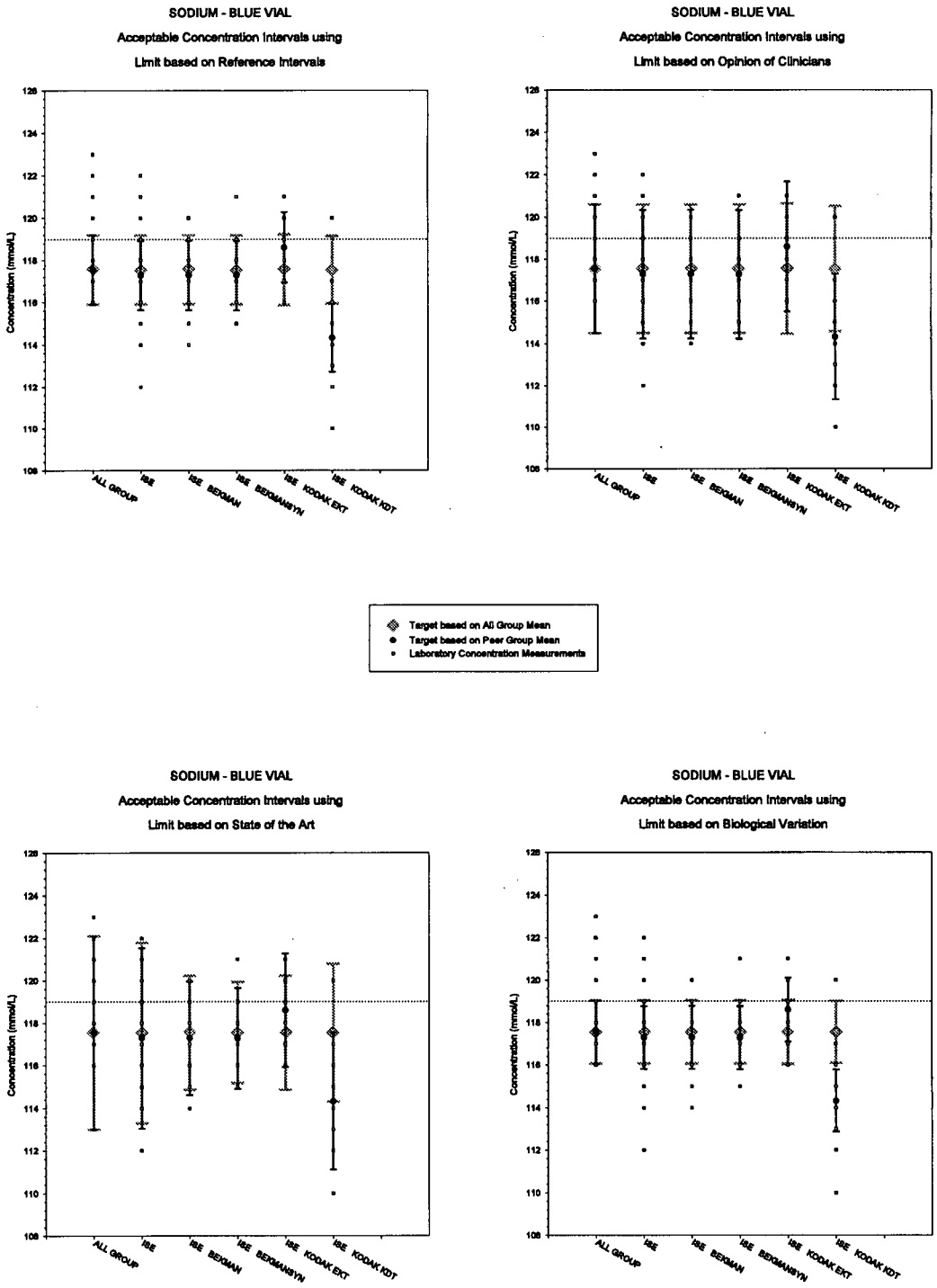


Figure A.26: Acceptable Concentration Intervals for the Blue Vial of Sodium

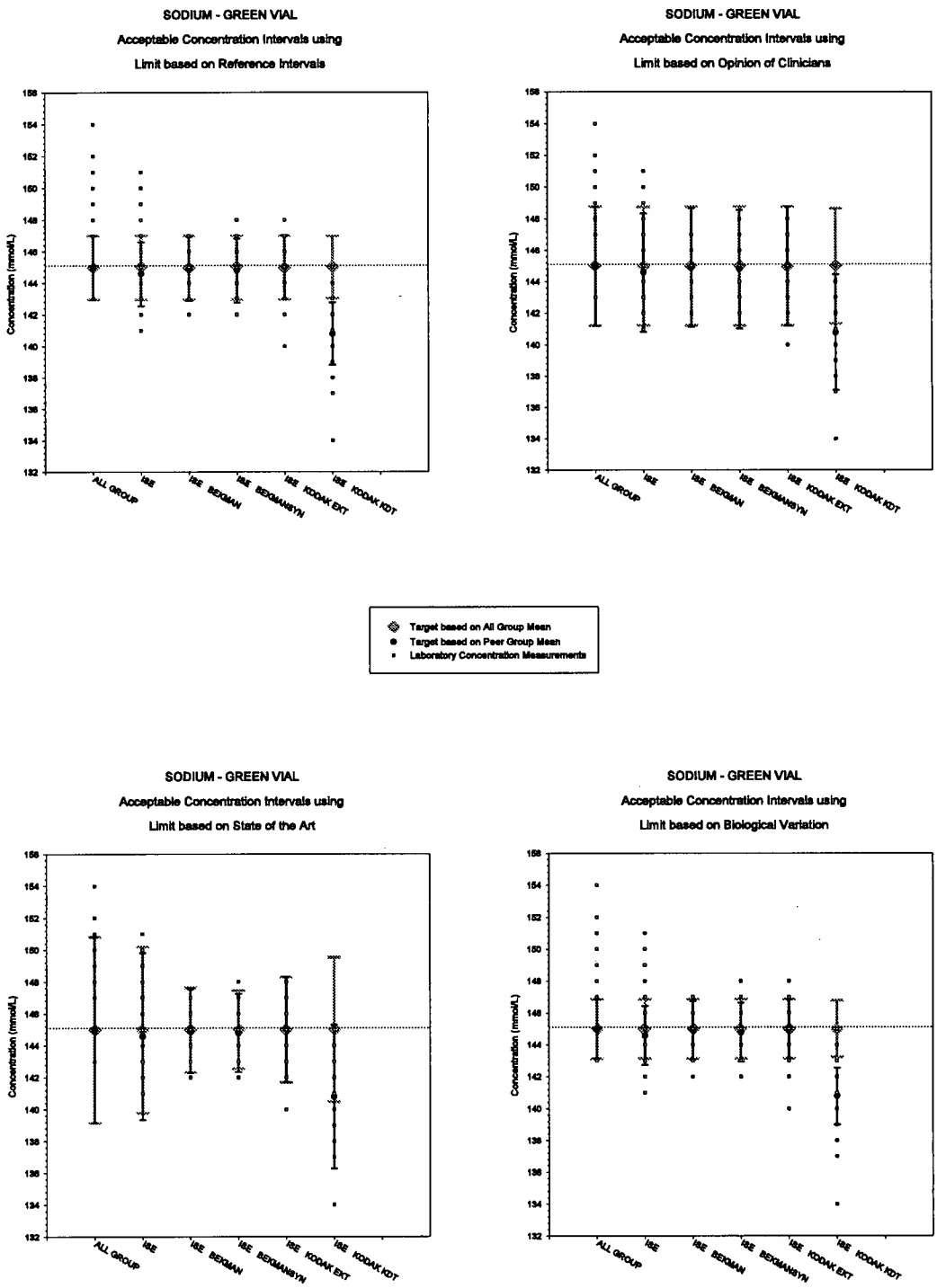


Figure A.27: Acceptable Concentration Intervals for the Green Vial of Sodium

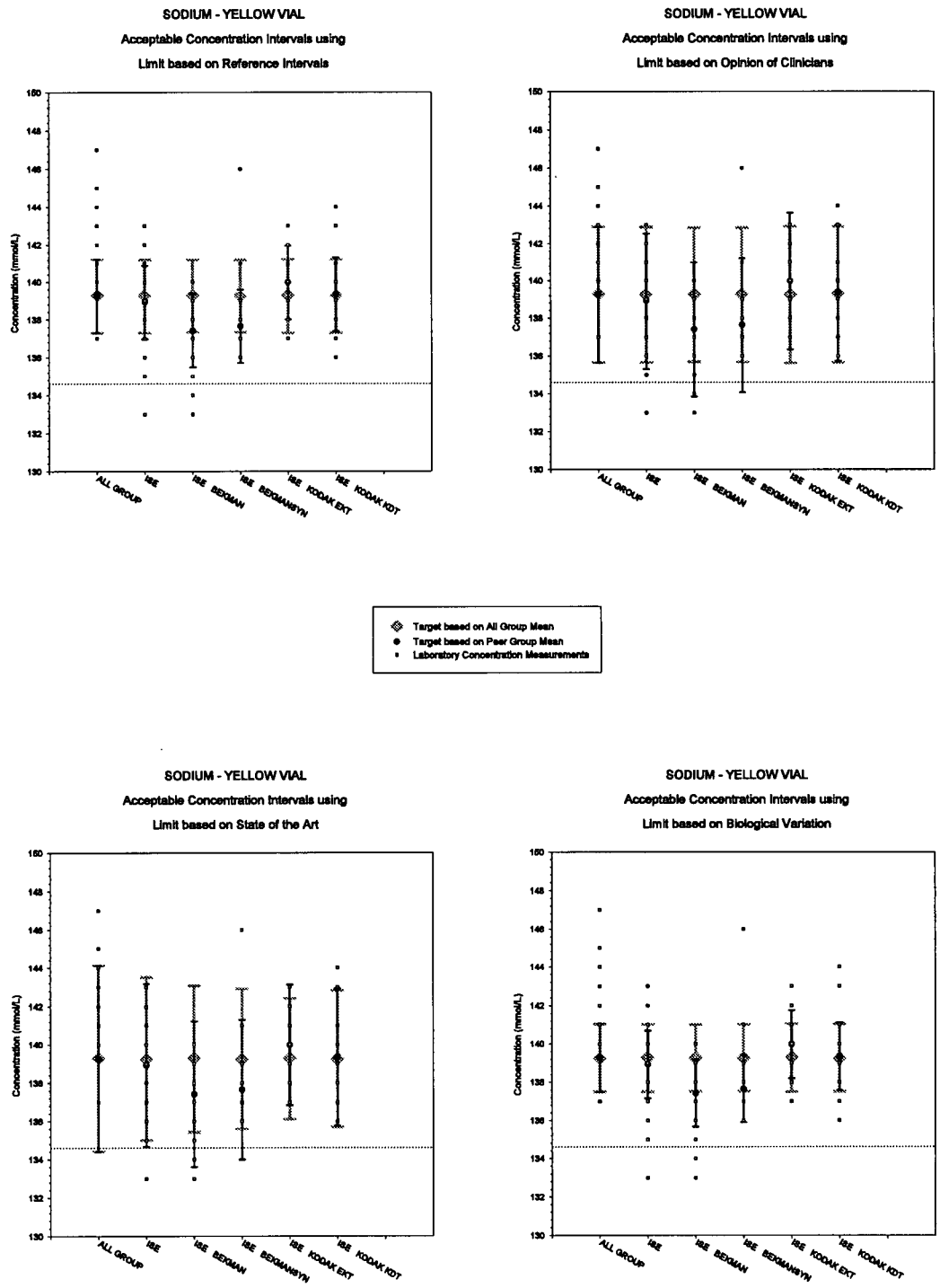


Figure A.28: Acceptable Concentration Intervals for the Yellow Vial of Sodium

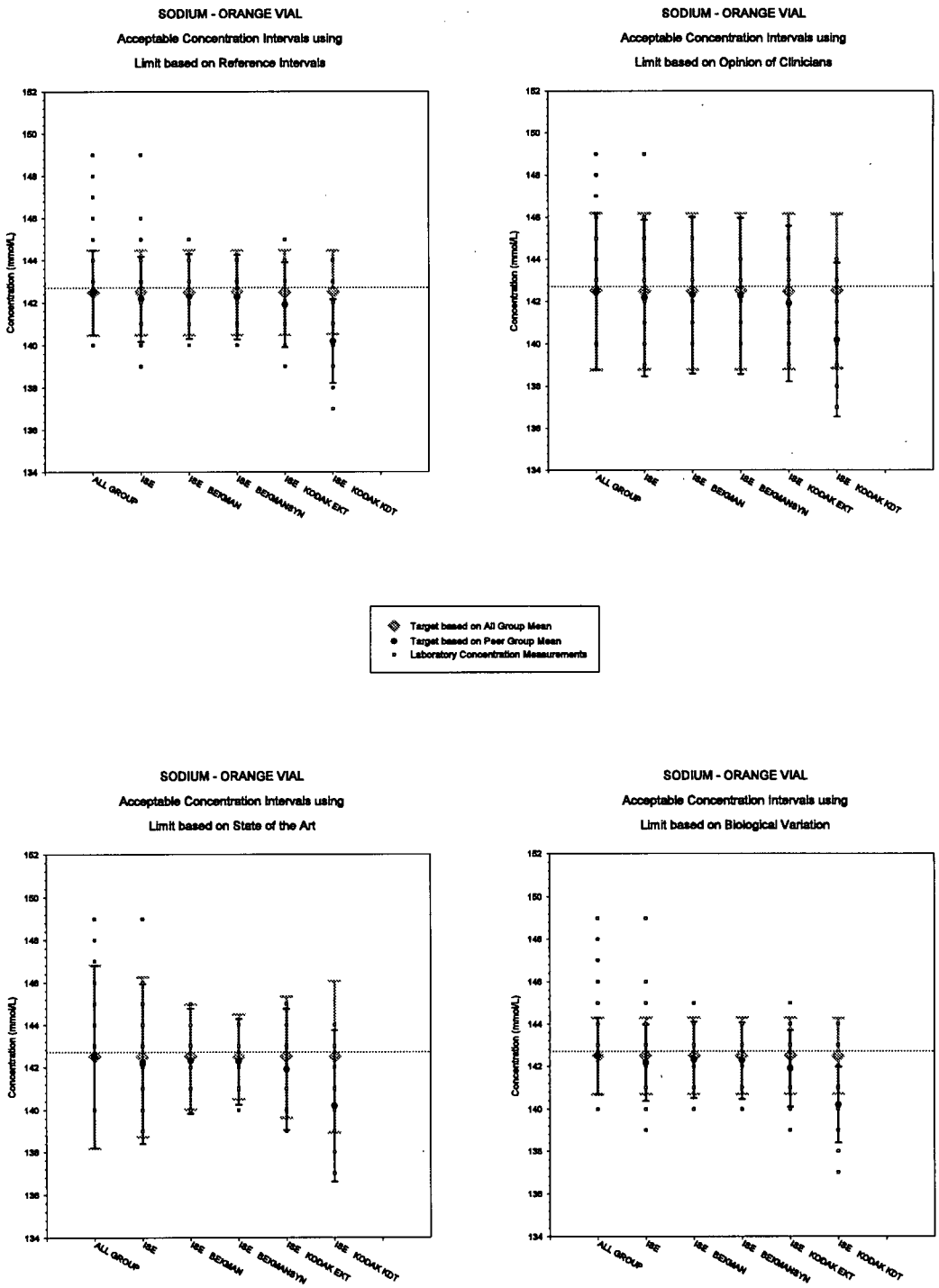


Figure A.29: Acceptable Concentration Intervals for the Orange Vial of Sodium

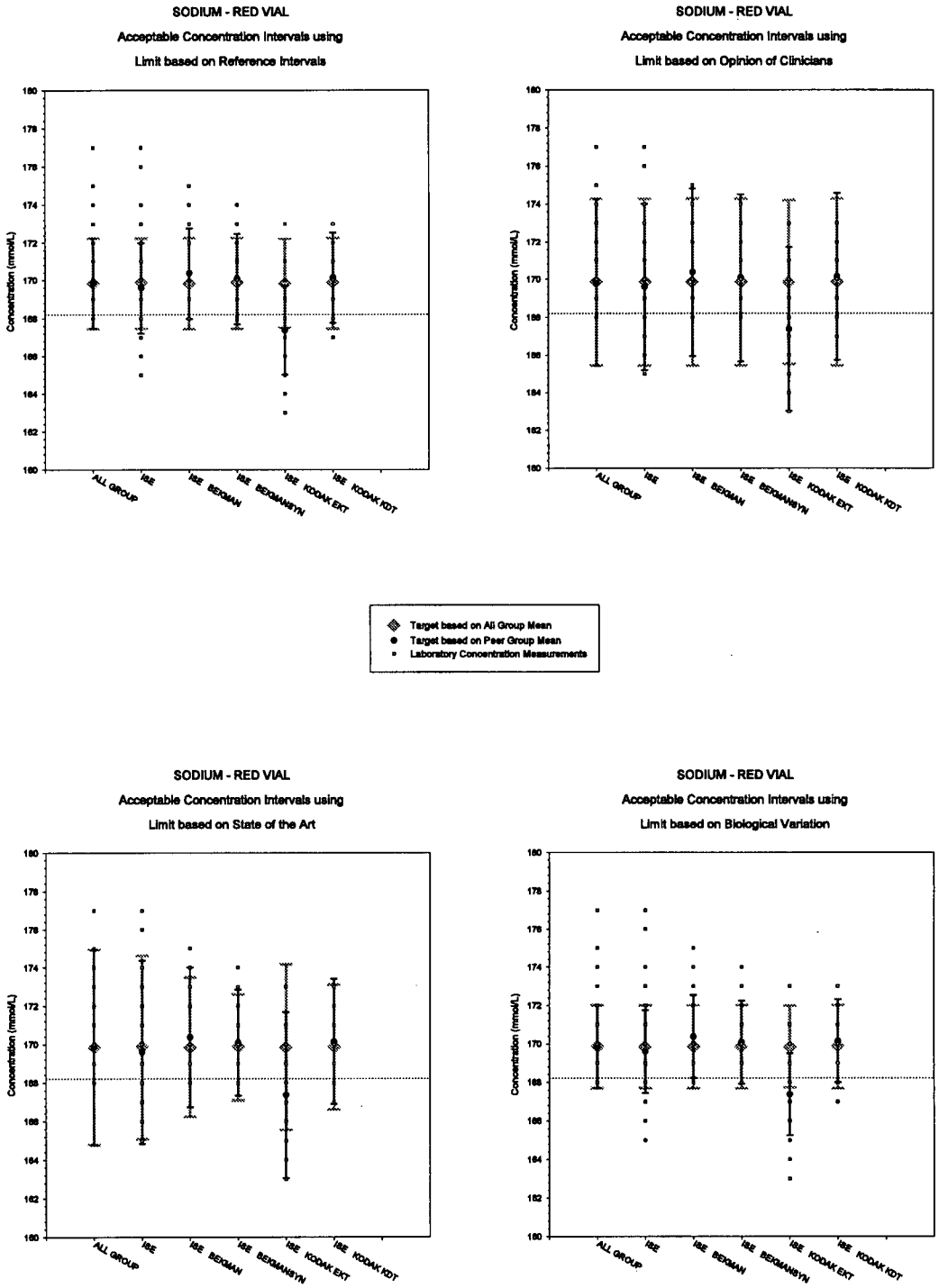


Figure A.30: Acceptable Concentration Intervals for the Red Vial of Sodium

All-Group Mean Target

	Reference Method Target: Pass	Fail	
Pass	446	73	519
Fail	33	186	219
	479	259	738

Overall Proportion of Agreement: 0.86 (SD≈0.013)

Peer Group Mean Target ARSDYE

	Reference Method Target: Pass	Fail	
Pass	407	84	491
Fail	72	175	247
	479	259	738

Overall Proportion of Agreement: 0.79 (SD≈0.015)

Peer Group Mean Target ARSDYE BEKMAN

	Reference Method Target: Pass	Fail	
Pass	355	92	447
Fail	124	167	291
	479	259	738

Overall Proportion of Agreement: 0.71 (SD≈0.017)

Peer Group Mean Target ARSDYE KODAK

	Reference Method Target: Pass	Fail	
Pass	369	115	484
Fail	110	144	254
	479	259	738

Overall Proportion of Agreement: 0.70 (SD≈0.017)

Peer Group Mean Target ARSDYE KODAK EKTSYS

	Reference Method Target: Pass	Fail	
Pass	359	116	475
Fail	120	143	263
	479	259	738

Overall Proportion of Agreement: 0.68 (SD≈0.017)

Peer Group Mean Target CRESO

	Reference Method Target: Pass	Fail	
Pass	386	75	461
Fail	93	184	277
	479	259	738

Overall Proportion of Agreement: 0.77 (SD≈0.015)

Figure A.31: Overall Agreement between the All-group Mean Target (as well as the Peer Group Mean Targets) and the Reference Method Target for Calcium

All-Group Mean Target

	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	452	22	474
Fail	42	381	423
	494	403	897

Overall Proportion of Agreement: 0.93
(SD≈0.009)

Peer Group Mean Target ISE

	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	438	47	485
Fail	56	356	412
	494	403	897

Overall Proportion of Agreement: 0.88
(SD≈0.011)

Peer Group Mean Target ISE BEKMAN

	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	316	88	404
Fail	178	315	493
	494	403	897

Overall Proportion of Agreement: 0.70
(SD≈0.015)

Peer Group Mean Target ISE BEKMAN SYNSYS

	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	401	70	471
Fail	93	333	426
	494	403	897

Overall Proportion of Agreement: 0.82
(SD≈0.013)

Peer Group Mean Target ISE KODAK

	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	309	140	449
Fail	185	263	448
	494	403	897

Overall Proportion of Agreement: 0.64
(SD≈0.016)

Peer Group Mean Target ISE KODAK EKTSYS

	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	247	168	415
Fail	247	235	482
	494	403	897

Overall Proportion of Agreement: 0.54
(SD≈0.017)

Figure A.32: Overall Agreement between the All-group Mean Target (as well as the Peer Group Mean Targets) and the Reference Method Target for Chloride

All-Group Mean Target

	Reference Method Target: Pass	Fail	
Pass	235	105	340
Fail	16	37	53
	251	142	393

Overall Proportion of Agreement: 0.69
(SD≈0.023)

Peer Group Mean Target CALG

	Reference Method Target: Pass	Fail	
Pass	243	101	344
Fail	8	41	49
	251	142	393

Overall Proportion of Agreement: 0.72
(SD≈0.023)

Peer Group Mean Target CALG BEKMAN SYNSYS

	Reference Method Target: Pass	Fail	
Pass	241	97	338
Fail	10	45	55
	251	142	393

Overall Proportion of Agreement: 0.73
(SD≈0.022)

Peer Group Mean Target COLOR

	Reference Method Target: Pass	Fail	
Pass	232	109	341
Fail	19	33	52
	251	142	393

Overall Proportion of Agreement: 0.67
(SD≈0.024)

Peer Group Mean Target COLOR KODAK EKTSYS

	Reference Method Target: Pass	Fail	
Pass	241	97	338
Fail	10	45	55
	251	142	393

Overall Proportion of Agreement: 0.64
(SD≈0.024)

Figure A.33: Overall Agreement between the All-group Mean Target (as well as the Peer Group Mean Targets) and the Reference Method Target for Magnesium

All-Group Mean Target	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	313	249	562
Fail	93	431	524
	406	680	1086

Overall Proportion of Agreement: 0.68
(SD≈0.014)

Peer Group Mean Target ISE	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	313	198	511
Fail	93	482	575
	406	680	1086

Overall Proportion of Agreement: 0.73
(SD≈0.013)

Peer Group Mean Target ISE BEKMAN	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	360	182	542
Fail	46	498	544
	406	680	1086

Overall Proportion of Agreement: 0.79
(SD≈0.012)

Peer Group Mean Target ISE BEKMAN SYNSYS	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	360	110	470
Fail	46	570	616
	406	680	1086

Overall Proportion of Agreement: 0.86
(SD≈0.011)

Peer Group Mean Target ISE KODAK EKTSYS	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	175	258	433
Fail	231	422	653
	406	680	1086

Overall Proportion of Agreement: 0.55
(SD≈0.015)

Peer Group Mean Target ISE KODAK KDT60	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	255	168	423
Fail	151	512	663
	406	680	1086

Overall Proportion of Agreement: 0.71
(SD≈0.014)

Figure A.34: Overall Agreement between the All-group Mean Target (as well as the Peer Group Mean Targets) and the Reference Method Target for Potassium

All-Group Mean Target	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	429	216	645
Fail	72	372	444
	501	588	1089

Overall Proportion of Agreement: 0.74
(SD≈0.013)

Peer Group Mean Target ISE	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	373	205	578
Fail	128	383	511
	501	588	1089

Overall Proportion of Agreement: 0.69
(SD≈0.014)

Peer Group Mean Target ISE BEKMAN	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	395	198	593
Fail	106	390	496
	501	588	1089

Overall Proportion of Agreement: 0.72
(SD≈0.014)

Peer Group Mean Target ISE BEKMAN SYNSYS	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	420	214	634
Fail	81	374	455
	501	588	1089

Overall Proportion of Agreement: 0.73
(SD≈0.013)

Peer Group Mean Target ISE KODAK EKTSYS	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	406	104	510
Fail	95	484	579
	501	588	1089

Overall Proportion of Agreement: 0.82
(SD≈0.012)

Peer Group Mean Target ISE KODAK KDT60	Reference Method Target: Pass	Reference Method Target: Fail	
Pass	129	279	408
Fail	372	309	681
	501	588	1089

Overall Proportion of Agreement: 0.40
(SD≈0.015)

Figure A.35: Overall Agreement between the All-group Mean Target (as well as the Peer Group Mean Targets) and the Reference Method Target for Sodium