1ª ed

PDICSS

Statistical Methods for Assessing Agreement of Measurements: examples on obstetrics and gynecology

Cristina Maria Nogueira da Costa Santos

PROGRAMA DE DOUTORAMENTO EM
INVESTIGAÇÃO CLÍNICA E EM SERVIÇOS DE SAÚDE
3º CICLO DE ESTUDOS

ORIENTADORES:
João Francisco Montenegro de Andrade Lima Bernardes
John Martin Bland

MAR|2010

FMUP

SERVIÇO DE BIOESTATÍSTICA E INFORMÁTICA MÉDICA

DESIGN
CARLA AMARAL

EXECUÇÃO GRÁFICA
MEDISA – EDIÇÕES E DIVULGAÇÕES CIENCTÍFICAS, LDA

ENDEREÇO E CONTACTOS
AL. PROF. HERNÂNI MONTEIRO | 4200-319 PORTO . PORTUGAL
TEL: +351 225 513 622 | FAX: +351 225 313 623
E-MAIL: **SBIM@MED.UP.PT** | HTTP:// **SBIM.MED.UP.PT**

ISBN
978-989-20-2019-8

TIRAGEM
30 EXEMPLARES

DATA
MAIO 2010

Thesis prepared under the supervision of:

Professor Doutor João Francisco Montenegro de Andrade Lima Bernardes
Instituto de Engenharia Biomédica
Faculdade de Medicina da Universidade do Porto

Professor John Martin Bland
Department of Health Sciences, University of York

Professores Catedráticos:

Abel Vitorino Trigo Cabral
Alberto Manuel Barros da Silva
Altamiro Manuel Rodrigues Costa Pereira
Álvaro Jerónimo Leal Machado de Aguiar
António Albino Coelho Marques Abrantes Teixeira
António Carlos Freitas Ribeiro Saraiva
Daniel Filipe Lima Moura
Deolinda Maria Valente Alves Lima Teixeira
Francisco Fernandes Rocha Gonçalves
Isabel Maria Amorim Pereira Ramos
João Francisco Montenegro Andrade Lima Bernardes
Jorge Manuel Mergulhão Castro Tavares
José Agostinho Marques Lopes
José Carlos Neves da Cunha Areias
José Eduardo Torres Eckenroth Guimarães
José Manuel Pereira Dias de Castro Lopes
José Henrique Dias Pinto de Barros
José Luís Medina Vieira
José Manuel Lopes Teixeira Amarante
Manuel Alberto Coimbra Sobrinho Simões
Manuel António Caldeira Pais Clemente
Manuel Jesus Falcão Pestana Vasconcelos
Manuel Maria Paula Barbosa
Maria Amélia Duarte Ferreira
Maria Dulce Cordeiro Madeira
Maria Fátima Machado Henriques Carneiro
Maria Leonor Martins Soares David
Patrício Manuel Vieira Araújo Soares Silva
Rui Manuel Almeida Mota Cardoso
Rui Manuel Lopes Nunes

Professores Jubilados e Aposentados:

Abel José Sampaio da Costa Tavares
Alexandre Alberto Guerra Sousa Pinto
Amândio Gomes Sampaio Tavares
António Augusto Lopes Vaz
António Carvalho Almeida Coimbra
António Fernandes da Fonseca
António Fernandes Oliveira Barbosa Ribeiro Braga
António Germano Pina Silva Leal
António José Pacheco Palha
António Luís Tomé da Rocha Ribeiro
António Manuel Sampaio de Araújo Teixeira
Artur Manuel Giesteira de Almeida
Belmiro dos Santos Patrício
Cândido Alves Hipólito Reis
Carlos Rodrigo Magalhães Ramalhão
Cassiano Pena de Abreu e Lima
Daniel Santos Pinto Serrão
Eduardo Jorge Cunha Rodrigues Pereira
Fernando de Carvalho Cerqueira Magro Ferreira
Fernando Tavarela Veloso
Francisco de Sousa Lé
Henrique José Ferreira Gonçalves Lecour de Menezes
João Silva Carvalho
Joaquim Germano Pinto Machado Correia da Silva
José Augusto Fleming Torrinha
José Carvalho de Oliveira
José Fernando Barros Castro Correia
José Manuel Costa Mesquita Guimarães
Levi Eugénio Ribeiro Guerra
Luís Alberto Martins Gomes de Almeida
Manuel Augusto Cardoso de Oliveira
Manuel Machado Rodrigues Gomes
Manuel Teixeira Amarante Júnior
Maria da Conceição Fernandes Marques Magalhães
Maria Isabel Amorim de Azevedo
Mário José Cerqueira Gomes Braga
Serafim Correia Pinto Guimarães
Valdemar Miguel Botelho dos Santos Cardoso
Walter Friedrich Alfred Osswald

"The most elegant design of a clinical study will not overcome the damage caused by unreliable or imprecise measurement."

Fleiss, J (1986). The design and Analysis of Clinical Experiments. New York, John Wiley & Sons.

Acknowledgments

Any project certainly requires assistance and encouragement from others and this PhD project was not the exception.

Foremost, I would like to express my deeply gratitude towards my adviser, Professor João Bernardes, for his support, advice, encouragement and guidance. His exceptional scientist intuition gave me the ideas and he discussed with me all the work presented in this thesis.

I also wish to thank my co-adviser, Professor Martin Bland. I learned so much reading his extensive and extremely important work in statistical methods for assessing agreement and I hope to collaborate much more with him in the future.

I would like to thank the opportunity of joining the SisPorto research work group. The mensal meetings with Professor João Bernardes, Professor Diogo Ayres de Campos, Dr. Antónia Costa and Dr. Célia Costa were extremely important to me.

To Professor Altamiro da Costa Pereira, the head of Department of Biostatistics and Medical Informatics, I would like to acknowledge his support during this PhD project and all the different things I learned with him in the last decade. I learned so much with his inviting attitude towards science and towards the life in general. Moreover, my first steps in biostatistics and in assessment of agreement were done under his precious guidance.

I want to thank all my colleagues in the Department of Biostatistics and Medical Informatics; they provide me an excellent work environment and support. Work in a Department with people with so different backgrounds (physicians, computer scientists, mathematicians, etc…) was very helpful for

my PhD project. Many thank also to the administrative staff of the Department of Biostatistics and Medical Informatics for all the support.

To my co-authors also many thanks. I have the privilege to have as co-authors, both clinicians (Altamiro da Costa Pereira and the obstetricians from SisPorto Group mentioned above) and computer scientists/mathematicians (Luís Antunes and André Souto). Some of my papers would not be possible without this interdisciplinary team with different backgrounds and knowledge.

Many thank to my marvellous mother for the support and love always present. A word of appreciation to all my 4 brothers, and specially to my sister, an inspiring example of a woman.

To my kids Filipa and Lucas I thank their unconditional love and I am sorry for the lack of time and impatience during this PhD project. Finally, my deepest gratitude goes to my husband who, sometimes, seems to desire the success of this PhD project even more than myself. However, together, we have a more important and successful project: to build a solid family.

Abbreviations

ANOVA	-	Analisys of Variance
CCC	-	Concordance Correlation Coefficient
CI	-	Confidence Interval
CIN	-	Cervical Intraepithelial Neoplasia
CTG	-	Cardiotocography
FHR	-	Fetal Heart Rate
ICC	-	Intraclass Correlation Coefficient
K	-	Kappa statistics
LA	-	Limits of Agreement
PA	-	Proportions of Agreement
STAN	-	ST analyser
UAB	-	Umbilical Artery Blood

Resumo

A concordância na medição de uma determinada quantidade é fundamental para realizar um estudo de validade ou um ensaio clínico ou simplesmente para possibilitar uma eficiente prestação de cuidados. Uma vez que é impossível controlar ou eliminar todo o erro de medição é essencial medir o grau de concordância nas medições clínicas. Contudo, na literatura médica são encontrados com frequência estudos de concordância com análises incorrectas ou resultados mal interpretados.

Esta tese pretende 1) avaliar como é que a concordância tem sido medida em Obstetrícia e Ginecologia, 2) descrever as consequências da falta de concordância em medicina, 3) descrever as consequência do uso incorrecto ou da errada interpretação dos actuais métodos estatísticos para avaliar a concordância e, finalmente, 4) explorar se o estudo da complexidade pode contribuir para entender ou medir a concordância de medidas em Medicina.

Para avaliar como é que a concordância tem sido medida em Obstetrícia e Ginecologia foi realizada uma revisão (**Capítulo 4**). Esta revisão revela que medidas de concordância inapropriadas são, não raramente, usadas nas melhores revistas da área. Como o uso de uma só metodologia para avaliação da concordância pode levar a interpretações dos resultados erradas, o uso de mais do que um método para avaliar a concordância é recomendado.

O **Capítulo 4** descreve também as consequências da falta de concordância em medicina ilustrando como a falta de concordância pode influenciar dramaticamente, não só obviamente a prática clínica, mas também os resultados de estudos de investigação clínica.

Os **Capítulos 5 e 6** descrevem as consequências da má interpretação dos actuais métodos para avaliar a concordância considerando variáveis categóricas e contínuas, respectivamente. O **Capítulo 5** é baseado em três cartas ao Editor mostrando, com exemplos reais de estudos publicados, como o uso inapropriado de métodos estatísticos para avaliar a concordância em variáveis categóricas e outros problemas frequentemente encontrados, podem levar à

interpretação errada e inconsistente dos resultados. O **Capítulo 6** descreve como as metodologias mais recomendadas para avaliar a concordância, considerando variáveis contínuas, podem ser interpretadas de forma inconsistente podendo levar a conclusões erradas. As inconsistências observadas neste estudo podem ser explicadas pelas limitações dos métodos estatísticos usados, estas limitações são bem conhecidas e estão descritas na literatura, mas as suas consequências na interpretação dos resultados não tinham ainda sido demonstradas desta forma. Consequentemente, o uso de mais do que um método para avaliar a concordância, recomendado no Capítulo 4, pode não ser assim muito simples de interpretar.

Motivado pela necessidade de futura investigação e desenvolvimento de métodos estatísticos e pela necessidade de entender os resultados obtidos com os actuais métodos para avaliar a concordância, os **Capítulos 7 e 8** exploram com é que a complexidade pode contribuir para entender ou mesmo medir a concordância de medidas em Medicina. De facto, o **Capítulo 7** mostra como a análise da entropia, uma medida de complexidade ou incerteza, e a categorização das variáveis contínuas pode ajudar a entender os resultados tão discrepantes encontrados no estudo apresentado no Capítulo 6. No **Capítulo 8**, propomos uma nova medida de discordância baseada em informação que nos permite a comparação entre diferentes populações, uma limitação das outras medidas de concordância já existentes.

Como todos os métodos para avaliar a concordância têm importantes limitações, é a impressão global dada por diferentes formas de avaliar a concordância que nos pode dar uma ideia do grau de concordância de uma medição. Sempre que resultados de estudos de concordância são apresentados, é muito importante discutir a metodologia estatística usada e ter em conta as suas limitações quando se interpretam os resultados do estudo.

Esta tese mostra que a complexidade pode trazer não só uma melhor quantificação da concordância de medidas como também uma melhor compreensão dos problemas fundamentais relacionados com a avaliação da concordância em medicina.

Abstract

Agreement among measurements of the same quantity is fundamental for validity studies, randomised controlled trials and to ensure efficient delivery of health care. As it is impossible to control and eliminate all the measurement error it is essential to assess the degree of agreement in clinical measurements. However incorrect analysis or misleading results of agreement studies in medical literature is not rare.

This thesis intends 1) to assess how agreement has been measured in the Obstetrics and Gynecology, 2) to describe the consequences of lack of agreement in medicine, 3) to describe the consequences of misuse and misinterpretation of the current statistical strategies to assess agreement and finally 4) to explore whether complexity science and concepts such as entropy could contribute to our understanding of assessment of observer agreement in medicine.

To assess how agreement has been measured in the obstetrics and gynaecology literature a review was performed and presented in **Chapter 4**. This review clearly revealed that inappropriate measures of agreement continue to be used in world leading journals and, as the use of one single strategy to assess agreement can be misleading, the use of more than one strategy is recommended.

The **Chapter 4** also describes the consequences of lack of agreement in medicine illustrating how poor reproducibility can dramatically influence, not only obviously the clinical practice, but also the results of clinical research.

Chapters 5 and 6 describe the consequences of misuse and misinterpretation of the current statistical strategies to assess agreement considering categorical and continuous variables, respectively. **Chapter 5** is based in three letters to the editor showing, with real examples published in the literature, how the use of inappropriate methods for agreement assessment, considering categorical variables, and other common problems can lead to inconsistent and erroneous results. **Chapter 6** describes how the most recommended strategies to assess

agreement, considering continuous variables, can be interpreted providing inconsistent and misleading conclusions. The inconsistencies observed in this study may be explained by the limitations of statistical methods used, these limitations are well known and already described in the literature, but their consequences on the interpretation of results had not been demonstrated so clearly. Consequently, the use of more than one statistical strategy to assess agreement, recommended in chapter 4, can also be not straightforward.

Motivated by the need of further study and development of statistical methodology and the need to understand the results provide by the current statistical methods for assessing agreement, **Chapters 7 and 8** explore whether complexity science and concepts such as entropy could contribute to our understanding of assessment of observer agreement in medicine. In fact, **Chapter 7** shows how the entropy assessment, a measure of complexity or uncertainty, and the categorization of continuous variables helps to shed some light into the discrepant results of observer agreement found in the previous study presented in Chapter 6. In **Chapter 8**, we propose a new information-based approach to assess disagreement which allows the comparability of different trait populations, a limitation common to the current statistical strategies.

As all method for agreement measurement have important limitations, the overall impression given from several statistical approaches should gives one the feeling that a measure is reliable and to what degree. It is very important to discuss methods and specific drawbacks when interpreting the results of individual agreement study.

This thesis shows that complexity can provide not only better quantification of agreement but also better understanding of the underlying problems related to the measurement of agreement.

Contents

Contents

Acknowledgments ..... **i**

Abbreviations ..... **iii**

Resumo ..... **v**

Abstract ..... **vii**

Contents ..... **xi**

1 Introduction **3**

1.1 The different types of agreement 5
1.2 Statistical methods to assess agreement in continuous variables 6
1.3 Statistical methods to assess agreement in categorical variables 10
1.4 Complexity and agreement 12
1.5 References 13

2 Objectives **21**

3 Outline **25**

4 Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences **29**

4.1 Abstract 30
4.2 Introduction 31
4.3 Methods 31
4.4 Results 32
4.5 Discussion 34
4.6 References 35

5 Frequent problems in the design and analysis of agreement studies with categorical variables. **39**

5.1 Introduction 41
5.2 Design problems and inappropriate statistical methods 41
5.3 Incomplete analysis 42
5.4 Missing information 44
5.5 References 46
5.6 Appendix 1 49
5.7 Appendix 2 51
5.8 Appendix 3 53

6 The limits of agreement and the intra-class correlation coefficient may be inconsistent in assessing agreement. 57

6.1 Abstract 58
6.2 Introduction 59
6.3 Methods 59
6.4 Results 62
6.5 Discussion 66
6.6 Conclusions 67
6.7 Acknowledgments 68
6.8 References 68

7 Complexity and categorical analysis may improve the interpretation of agreement studies using continuous variables .. 73

7.1 Introduction 75
7.2 Complexity in clinical decisions 75
7.3 Entropy, a measure of complexity or uncertainty 76
7.4 Adding complexity and categorical analysis to the LA and ICC in prediction of neonatal outcomes based on FHR monitoring 76
7.5 Conclusions 78
7.6 References: 79

8 Assessment of disagreement: a new information based approach83

8.1 Abstract 84
8.2 Introduction 85
8.3 Methods 86
Most used strategies to assess agreement 86
The New Approach: Information-Based Measure of Disagreement87
Properties of Information-Based Measure of Disagreement 89
Inference of the information-based measure of disagreement 89

8.4 Results 90
Example - systolic blood pressure 90
Example - Apgar score at first minute 90
Simulation study 93
8.5 Discussion 94
8.6 References 96
8.7 APPENDIX 99
Properties of the proposed disagreement measure 99

9 General discussion and conclusions 105

9.1 General discussion 107
9.2 Conclusions 112
9.3 References 113

1. Introduction

1 Introduction

Which statistical method for assessing agreement is best depends on the aim of agreement study. This introductory chapter describes the different types of agreement studies and the respective statistical methods for assessing agreement both considering categorical and continuous variables. Moreover, this chapter discusses how observer agreement can be related with complexity in medical practice.

1.1 The different types of agreement

The quality of the measurements taken by health professionals or by measurement devices is fundamental not only for clinical care but also for research (Shoukri 2005). Measurement of variables always implies some degree of error. Often, when an observer takes a measurement, the value obtained depends on several things such us the skills of the observer, his or her experience, the measurement instrument, and the observer's expectations (Bland 2000). Also, natural continuous variation in a biological quantity can be present (Bland 2000). For example, if two observers measure blood pressure consecutively in the same subject and under the same circumstances, the values obtained by the two observers can be different because of their different skills and experience, their different expectations, the measurement instrument, and also because of the natural continuous variation in blood pressure. When natural continuous variation in a biological quantity (**variation within the subject**) is present, it is outside the control of the observer. It is, however, possible to minimize the observer variability by training of observers, by the use of guidelines, and through automation, for example. In fact, it is almost always possible to reduce observer variability but it is impossible to eliminate it. Thus, the assessment of measurement agreement in different forms is an important issue in medicine.

It is important to assess how much different observers agree when they rate the same measurement (**inter-observer agreement**). Sometimes the interest is in the properties of the measurement method itself; for example, we might want to see whether a new measurement technique can be reproduced by a second observer. At other times, the focus may be on the observers rather than on the measurement method. For example, we may wish to evaluate the observer agreement of different obstetricians classifying cardiotocographic patterns as normal, suspicious, or pathological before and after a course explaining the guidelines for cardiotocography interpretation. It is also important to assess how much observations agree when the same observer repeats the same measurement (**intra-observer agreement**).

Sometimes, the aim is to assess the **agreement between two methods of clinical measurement**; for example, if a new, faster, or less invasive method agrees with the old one, i.e. if they do not disagree sufficiently to cause

problems in clinical interpretation, we can replace the old method with the new one or use both interchangeably.

But the aim could also be to test the **validity**, or the agreement between a measurement and the corresponding true value (gold standard), if it is known. When a new diagnostic test is developed, for example, it is essential to test it validity. When the gold standard is not known, the observer agreement provides an upper bound on the degree of validity present in the rating. In fact, if the observer agreement is good, then there is a possibility, but no means a guarantee, that the ratings do in fact reflect the gold standard. On the other hand, if the observer agreement is poor, then the usefulness of the ratings is severely limited and they cannot be considered valid (Fleiss 2003). Everyone is usually aware of the importance of a validity study before the introduction of a new measurement method or diagnostic test in medical practice. Also, everyone is aware of the importance of assessing agreement between two methods when we intend to replace an old method with a new one. However, although the clinical consequences of low observer agreement have been recognized for more than 50 years (Yerushalmy 1953), the importance of observer agreement studies is not always given due recognition. In general, improvement of observer agreement in clinical observations can have a great influence on the quality of healthcare (de Vet 2005).

1.2 Statistical methods to assess agreement in continuous variables

Several statistical methods to assess agreement between measurements in continuous scales can be found in the literature; however, all these statistical methods for assessment of agreement have important limitations and no method appears to be foolproof (Luiz 2005). Moreover, it is not uncommon to find some inappropriate analyses in leading journals. Bland and Altman have described the problems of some applications of correlation and regression methods to agreement studies, using examples from the obstetrics and gynaecology literature (Bland 2003), but other authors have found inappropriate analyses in papers dealing with other aspects of medicine (Gow 2008; Jakobsson 2005; Rothwell 2000; Winkfield 2003). Some authors recommended the use of a combination of more than one statistical method to assess agreement (Khan 2001; Luiz 2005) in order to overcome the limitations of each method, but this is not an ideal approach.

We can quantify **variation within the subject** by replicating readings of each subject and calculating the within-subject standard deviation ($\sqrt{s_w}$) by one-way analysis of variance (ANOVA) where s_w is the residual mean square in one-way ANOVA with the subjects being the "groups". The repeatability coefficient, which is $1.96\sqrt{2s_w^2}$, can be also computed. In this situation, we can also calculate the within-subject coefficient of variation, which is the within-subject standard deviation divided by the mean.

The intraclass correlation coefficient (ICC) is frequently claimed to be suitable for **assessment of observer agreement** (Bartko 1966; Shoukri 1996). Sometimes the ICC is also recommended for assessing agreement between two methods of clinical measurement (Lee et al. 1989), but other authors contest this recommendation, claiming that ICC is inappropriate for assessing the agreement between two measurement methods (Bland et al. 1990). The ICC is defined as the ratio of the variance between subjects to the total variance (Bartko 1966; Shrout 1979). These variances are derived from ANOVA and the ANOVA model depends on the agreement study design. If we have a pool of observers, and for each subject we select different observers randomly to rate this subject, this corresponds to a one-way ANOVA in which the subject is a random effect, and the observer is viewed as measurement error. On the other hand, if the same set of observers rates each subject, this corresponds to two-way ANOVA design in which both the subject and observers are separate effects. Here, observers are considered a random effect; this means that the observers in the study are considered to be a random sample from a population of potential observers. Finally, there is an ANOVA design similar to the former, but it is two-way in that the observers are "fixed"; the ICC estimates only apply to the observers selected in the study, since it does not permit generalization to other ones (Shrout 1979).

The ICC (from two-way models) that should be used for assessing agreement was defined by McGraw and Wong as the "ICC for agreement". We obtain the "ICC for consistency" or the "ICC for agreement" either by excluding or not excluding the observer variance from the denominator mean square, respectively (McGraw 1996). The systematic variability due to observers is irrelevant for "ICC for consistency" but it is not irrelevant for "ICC for agreement".

The ICC ranges from 0 (no agreement) to 1 (perfect agreement), but it can be negative. How such negative ICC values should be interpreted is quite unclear (Giraudeau 1996). The assumptions, multivariate normal distributions,

and equality of variances should be checked. An important limitation of ICC is that it is strongly influenced by the variance of the trait in the sample/population in which it is assessed (Muller 1994). This limitation can be illustrated with the flowing example. Suppose that we aim to assess the inter-observer agreement of a depression scale. When applied to a random sample of the adult population (a heterogeneous population with a high variance) the scale's inter-observer agreement may be high (high ICC); however, if the scale is applied to a very homogeneous population (with a low variance), such as patients hospitalized for acute depression, the scale's inter-observer agreement may be low (low ICC). Consequently, ICC values have no absolute meaning and the cut-off value of 0.75 proposed by Fleiss (Fleiss 1986)—which is often reproduced to signify a good agreement—has limited justification.

Lin's concordance correlation coefficient (CCC) is the Pearson coefficient of correlation, which assesses the closeness of the data to the line of best fit, modified by taking into account how far the line of best fit is from the 45-degree line through the origin (Lin 1989). Lin objected to the use of the ICC as a way of assessing agreement between methods of measurement and developed the concordance correlation coefficient (CCC). However, there are similarities between certain specifications of the ICC and the CCC (Nickerson 1997). Moreover, some limitations of the ICC, such as the limitation of comparability of populations described above, are also present in the CCC (Atkinson et al. 1997).

Bland claimed that the assessment of observer agreement is one of the most difficult areas in the study of clinical measurement, and he suggested assessing whether measurements taken on the same subject by different observers vary more than measurements taken by the same observer, and if so by how much, taking a sample of observers to make repeated observations on each of a sample of subjects. It is possible then to assess, by ANOVA, how much the variation between measurements on the subject is increased when different observers make the measurements (Bland 2004).

To assess **agreement between two methods of clinical measurement**, Bland and Altman proposed the limits of agreement approach (Bland 1986). This method is also sometimes used to assess agreement between different observers, and some authors even recommend the use of both approaches to assess observer agreement and to assess agreement between two methods of clinical measurement in continuous variables (Khan 2001; Luiz 2005). In fact, both approaches have different pros and cons. Unlike ICC, the limits of agreement distinguish between random error and bias. Limits of agreement can

be calculated based on the mean difference between the measurements of two methods in the same subjects and the standard deviation of these differences. Approximately 95% of these differences will lie between the mean differences ± 1.96 standard deviations of these differences. The limits of agreement approach depends on some assumptions about the data: that the mean and standard deviation of the differences are constant throughout the range of measurement, and that these differences are from an approximately normal distribution. An easy way to check these assumptions is a scatter plot of the difference against the average of the two measurements and a histogram of the differences (Bland 1986). Limits of agreement are expressed in terms of the scale of measurement; it is not possible to say how the limits of agreement should be to represent "good" agreement. Based on the limits of agreement, deciding whether agreement is acceptable or not is always a clinical, and not a statistical, judgement.

The limits of agreement approach still works in **validity** studies, when one method of measurement is a "gold standard". It tells us how far from the "gold standard" the measurements by the method are likely to be. By calculating differences as "method" minus "gold standard", we are able to say that the new method might exceed the "gold standard" by up to the upper limit and be less than the "gold standard" by down to the lower limit. However, plotting of difference against "gold standard" is misleading: the differences should be plotted against the mean between the method and the gold standard (Bland 1995). Often, the reason for measurement of agreement between one method and the true values is to assess a new diagnostic test comparing the test results with the true diagnosis. The test may be based on a continuous variable and the disease indicated if it is above or below a given level, or it may be a categorical variable such as positive or negative. In both cases, we are interested in knowing the proportion of disease positives that are test positive (the sensitivity) and the proportion of disease negatives that are test negative (the specificity). If the test is based on a continuous variable, a receiver operating characteristic (ROC) curve can be used, plotting the sensitivity against one minus the specificity for different cut-off points (Metz 1978; Zweig 1993). These ROC curves enable us to find the cut-off that best combines good sensitivity and good specificity, and from the area under the curve we can compare different tests.

1.3 Statistical methods to assess agreement in categorical variables

In many fields of medicine, it is common to study observer agreement or agreement between two methods concerning categorical variables. The proportion of overall agreement is the proportion of cases for which observers agree. Although this way of assessing agreement is informative and useful, taken by itself, it does not distinguish between agreement on positive ratings and agreement on negative ratings. However, we can compute observed agreement relative to each rating category individually (proportion of specific agreement). This strategy can be a useful and simple way to express agreement. Consider the ratings of two observers summarized in Table 1; the proportion of overall agreement is the proportion of cases for which observers 1 and 2 agree: (a+d)/(a+b+c+d).

The proportions of positive and negative agreement are calculated as 2a/(2a+b+c) and 2d/(2d+b+c), respectively. According to Fleiss (Fleiss 2003), these proportions of positive and negative agreement were originally proposed by Dice (Dice 1945) as a similarity index for use in ecological studies and can be interpreted as the probability that an observer rating that is positive (or negative) will be given a positive rating (or a negative one) by another, randomly chosen, observer.

Table 1.1: Summary of binary ratings by two observers (observer 1 and observer 2).

	Observer 2	
Observer 1	**+**	**-**
+	a	b
-	c	d

Chamberlain proposed quantifying positive agreement by the probability that both observers rate a subject positively given that at least one of the observers rates the subject positively: a/(a+b+c) (Chamberlain 1975). It is possible to generalize these formulae (the proportions of overall and specific agreement) to apply them not only to binary ratings, but also to ordered, or nominal ratings and to permit any number of observers, with potentially different numbers of observers or different observers for each case (Markus 1996; Uebersax 2000).

Kappa statistics was introduced by Cohen to provide a coefficient of agreement between two observers for categorical nominal variables (Cohen 1960). The definition of kappa presented by Cohen in 1960 is suitable for more than two nominal response categories. Cohen introduced the weighted kappa (Cohen 1968) to allow each kind of disagreement to be weighted according to the degree of disagreement between ordinal categories (for example, a disagreement suspicious-normal represents less disagreement than a disagreement pathological-normal). The equivalence of weighted kappa to the ICC (discussed in the previous section) under some circumstances was shown by Fleiss (Fleiss 1971). Fleiss also proposed an extension to kappa allowing the assessment of agreement between several observers that rate all or some of the subjects and classify them into more than two categories (Fleiss 1971).

The kappa coefficient aims to measure chance-corrected agreement. If the observers are in complete agreement, then kappa is equal to one. If there is no agreement between the observers other than what would be expected by chance, then kappa is equal to zero. If the observers are in disagreement then kappa is negative, being -1 when they totally disagree. However, kappa not only measures agreement but is also affected in complex ways by the presence of bias between observers and by the distribution of data across the categories that are used (prevalence) (Byrt 1993). As a consequence, we cannot compare the level of agreement with kappa from different populations. Thus, the commonly used tables presented Landis (Landis 1977), Fleiss (Fleiss 2003) or Rosner (Rosner 1990), to define kappa's ranges for good or bad agreement are not appropriate. Moreover, kappa statistics depends on the number of categories (Cook 2005; Landis 1977). As kappa has several important limitations, its interpretation without the limitations in mind can lead to misleading results and erroneous conclusions. Moreover, kappa pretends to assess the level of agreement between observers that is not due to chance, but it is questionable whether kappa is in fact "chance-corrected"; indeed, it is by no means clear how chance affects the decisions of actual observers and how one might correct for it. One can only apply valid chance correction using models of observer decision-making (Uebersax 1987). So valid chance correction becomes an exercise in modelling observer agreement (Agresti 1992; Uebersax 1992) rather than merely calculating a simple index (Uebersax 2009). But, even more than the doubt about whether kappa in fact corrects agreement for chance is the doubt about whether a correction for chance is needed for an agreement coefficient. In fact, different opinions have been put forward on this question (Fleiss 2003). Rogot and Goldberg emphasized the importance of contrasting

observed agreement with expected agreement (Rogot 1966); however, Goodman and Kruskal argued that agreement expected from chance need not cause much concern, and that the observed degree of agreement can usually be assumed to be in excess for chance (Goddman 1954). When the aim is to assess the validity of, for example, a new diagnostic test (agreement between a diagnostic test and a gold standard), we usually estimate the sensitivity (the proportion of true positives that are correctly identified by the test) and the specificity (the proportion of true negatives that are correctly identified by the test) (Altman et al. 1994). Uebersax has argued that with sensitivity and specificity, no adjustment is made for possible chance agreements between the diagnostic test and the gold standard and the same principle should apply to measuring agreement between two tests or observers (Uebersax 2009). In fact, the proportions of positive and negative agreement applied to measuring agreement between two tests or observers (observer agreement) are closely analogous to sensitivity and specificity applied to measuring agreement between a diagnostic test and its gold standard (validity) (Uebersax 2009).

1.4 Complexity and agreement

A complex system is a collection of individual agents that act in ways that are not always totally predictable, and whose actions are interconnected so that the action of one part changes the context for other agents (Plsek 2001). Complex and chaotic systems have been related with low certainty and low agreement. In fact, diagnostic uncertainty is commonly found in medicine, and diagnostic agreement between clinicians is often surprisingly poor, even with hard observable criteria (Wilson 2001).

Considering the lack of agreement in the literature about which methods of agreement are suitable for assessment of agreement in different circumstances and considering the misuses of statistical measures frequently found in the literature, further research must be done in this area. Perhaps complexity science and concepts such as entropy or uncertainty could contribute to our understanding of the low observer agreement in medicine.

1.5 References

Agresti, A (1992). "Modelling patterns of agreement and disagreement." Stat Methods Med Res 1(2): 201-218.

Altman, DG and Bland, JM (1994). "Diagnostic tests. 1: Sensitivity and specificity." BMJ 308(6943): 1552.

Atkinson, G and Nevill, A (1997). "Comment on the Use of Concordance Correlation to Assess the Agreement between Two Variables." Biometrics 53(775-777).

Bartko, JJ (1966). "The intraclass correlation coefficient as a measure of reliability." Psychol Rep 19(1): 3-11.

Bland, JM (2000). An introduction to medical statistics, Oxford University Press.

Bland, JM. (2004, May 2004). "How do I analyse observer variation studies?" Retrieved 06-01-2010, from http://www-users.york.ac.uk/~mb55/meas/observer.pdf.

Bland, JM and Altman, DG (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." Lancet 1(8476): 307-310.

Bland, JM and Altman, DG (1990). "A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement." Comput Biol Med 20(5): 337-340.

Bland, JM and Altman, DG (1995). "Comparing methods of measurement: why plotting difference against standard method is misleading." Lancet 346(8982): 1085-1087.

Bland, JM and Altman, DG (2003). "Applying the right statistics: analyses of measurement studies." Ultrasound Obstet Gynecol 22(1): 85-93.

Byrt, T, Bishop, J and Carlin, JB (1993). "Bias, prevalence and kappa." J Clin Epidemiol 46(5): 423-429.

Chamberlain, J, Rogers, P, Price, JL, Ginks, S, Nathan, BE and Burn, I (1975). "Validity of clinical examination and mammography as screening tests for breast cancer." Lancet 2(7943): 1026-1030.

Cohen, J (1960). "A coefficient of agreement for nominal scales." Educational and Psychological Measurement 20: 37-46.

Cohen, J (1968). "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit." Phychological Bulletin 70: 213-220.

Cook, R (2005). Kappa and its dependence on marginal rates. The Encyclopedia of Biostatistics. Armitage, P. New York, Wiley: 2166-2168.

de Vet, HC (2005). Observe Reliability and Agreement. Encyclopedia of Biostatistics. Armitage, Pc, T., John Wiley.

Dice, L (1945). "Measurements of the amount of ecologic association between species. " Ecology 26: 297-302.

Fleiss, J (1971). "Measuring Nominal Scale agreement among many raters." Psychological Bulletin 76(5): 378-382.

Fleiss, J (1986). Design and Analysis of Clinical Experiments. New York, John Wiley & Sons.

Fleiss, J and Cohen, J (1973). "The equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability." Educational and Psychological Measurements 33(613–619.).

Fleiss, J, Levin, B and Paik, M (2003). Statistical Methods for Rates and Proportions, John Wiley & Sons, Inc.

Giraudeau, B (1996). "Negative values of the intraclass correlation coefficient are not theoretically possible." J Clin Epidemiol 49(10): 1205-1206.

Goddman, L and Kruskal, W (1954). "Measures of association for cross classification." Journal of American Statistical Association. 49: 732-764.

Gow, RM, Barrowman, NJ, Lai, L and Moher, D (2008). "A review of five cardiology journals found that observer variability of measured variables was infrequently reported." J Clin Epidemiol 61(4): 394-401.

Jakobsson, U and Westergren, A (2005). "Statistical methods for assessing agreement for ordinal data." Scand J Caring Sci 19(4): 427-431.

Khan, KS and Chien, PF (2001). "Evaluation of a clinical test. I: assessment of reliability." BJOG 108(6): 562-567.

Landis, JR and Koch, GG (1977). "The measurement of observer agreement for categorical data." Biometrics 33(1): 159-174.

Lee, J, Koh, D and Ong, CN (1989). "Statistical evaluation of agreement between two methods for measuring a quantitative variable." Comput Biol Med 19(1): 61-70.

Lin, LI (1989). "A concordance correlation coefficient to evaluate reproducibility." Biometrics 45(1): 255-268.

Luiz, RR and Szklo, M (2005). "More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported." J Clin Epidemiol 58(3): 215-216.

Markus, H, Bland, JM, Rose, G, Sitzer, M and Siebler, M (1996). "How good is intercenter agreement in the identification of embolic signals in carotid artery disease?" Stroke 27(7): 1249-1252.

McGraw, K and Wong, S (1996). "Forming inferences about some intraclass correlation coefficients." Psychological Methods 1: 30-46.

Metz, CE (1978). "Basic principles of ROC analysis." Semin Nucl Med 8(4): 283-298.

Muller, R and Buttner, P (1994). "A critical discussion of intraclass correlation coefficients." Stat Med 13(23-24): 2465-2476.

Nickerson, C (1997). "A Note on "A Concordance Correlation Coefficient to Evaluate Reproducibility"." Biometrics 53: 1503-1507.

Plsek, PE and Greenhalgh, T (2001). "Complexity science: The challenge of complexity in health care." BMJ 323(7313): 625-628.

Rogot, E and Goldberg, ID (1966). "A proposed index for measuring agreement in test-retest studies." J Chronic Dis 19(9): 991-1006.

Rosner, B (1990). Fundamentals of biostatistics. Boston, PWS-Kent Publishing.

Rothwell, PM (2000). "Analysis of agreement between measurements of continuous variables: general principles and lessons from studies of imaging of carotid stenosis." J Neurol 247(11): 825-834.

Shoukri, M (2005). Agreement, Measurement of. E. Armitage, Pc, T., John Wiley.

Shoukri, M and Edge, V (1996). Statistical Methods for Health Sciences, CRC Press, Inc.

Shrout, PE and Fleiss, JL (1979). "Intraclass correlations: uses in assessing rater reliability." Psychol Bull 86(2): 420-428.

Uebersax, J. (2000, 02 Oct 2009). "Raw Agreement Indices." Retrieved January, 2010., from http://www.john-uebersax.com/stat/raw.htm

Uebersax, J. (2009). "The Myth of Chance-Corrected Agreement." Retrieved January 2010, from http://www.john-uebersax.com/stat/kappa2.htm.

Uebersax, JS (1987). "Diversity of decision-making models and the measurement of interrater agreement." Psychological Bulletin 101: 140-146.

Uebersax, JS (1992). "Modeling approaches for the analysis of observer agreement." Invest Radiol 27(9): 738-743.

Wilson, T, Holt, T and Greenhalgh, T (2001). "Complexity science: complexity and clinical care." BMJ 323(7314): 685-688.

Winkfield, B, Aube, C, Burtin, P and Cales, P (2003). "Inter-observer and intra-observer variability in hepatology." Eur J Gastroenterol Hepatol 15(9): 959-966.

Yerushalmy, J (1953). "The reliability of chest roentgenography and its clinical implications." Dis Chest 24(2): 133-147.

Zweig, MH and Campbell, G (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." Clin Chem 39(4): 561-577.

2. Objectives

2 Objectives

1 To assess how agreement has been measured in the obstetrics and gynaecology literature.

2 To describe the consequences of lack of agreement in medicine.

3 To describe the consequences of misuse and misinterpretation of the current statistical strategies to assess agreement.

4 To explore whether complexity science and concepts such as entropy could contribute to our understanding of assessment of observer agreement in medicine.

3. Outline

3 Outline

After the **Introduction** and **Objectives** presented above, and after the **Outline** presented here, the rest of the thesis is organized in five chapters that describe the work performed for the dissertation. A description of these chapters follows.

The **fourth chapter** is a literature review that assesses how agreement has been measured in the obstetrics and gynaecology literature. This chapter also illustrates how lack of agreement can influence different types of research studies (validity studies, randomized controlled trials, and meta-analyses) and thus obviously influence clinical practice and medico-legal decisions also.

In spite of the importance of agreement studies, there is little consensus about which statistical methods are best to analyze agreement in categorical variables. It is not uncommon to find misleading interpretations of results of agreement studies in the literature. The **fifth chapter** discusses some misleading analysis and design in agreement studies with categorical variables found in the obstetrics and gynaecology literature.

There are some discussions in the literature about the advantages and disadvantages of each method used to assess agreement in continuous variables. Some authors have recommended the combination of intra-class correlation coefficient and limits of agreement to assess observer agreement in continuous variables, to overcome the limitations of each method. The **sixth chapter** shows, with a practical example, how the interpretation of the results from these two strategies to assess agreement may be inconsistent. In fact, the combination of more than one strategy is not sufficient and all the agreement studies reports should also discuss the limitations of the statistical strategies used to assess agreement, thus permitting a more correct interpretation of results.

As the most commonly recommended measures of agreement for continuous variables, the intra-class correlation coefficient and limits of agreement may not be enough to assess agreement, and sometimes it is the

overall impression from several approaches that gives one the feeling of how a measure is reliable, and to what degree. The **seventh chapter** considers that the interpretation of agreement studies of continuous variables may be improved by analysis of their complexity and by their transformation into categorical variables with clinical significance.

Finally, in the **eighth chapter**, an information-based measure of disagreement is proposed. This measure, used as a complement to the approaches currently used to evaluate agreement, would be useful for comparison of the degree of disagreement between different populations with different characteristics.

Finally, the thesis ends with a **General Discussion and Conclusions**.

4. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences

4 Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences

Costa-Santos C, Costa-Pereira A and Bernardes J.

BJOG 2005;112(5): 667-669.

4.1 Abstract

A literature review was performed to assess how agreement has been measured in obstetrics and gynaecology. Inappropriate or controversial measures were often used. The possible consequences of these inadequacies on validity studies and randomised controlled trials are shown in case examples. The association of two measures, proportions of agreement and kappa for categorical variables, and limits of agreement and intraclass correlation coefficient for continuous variables, is encouraged in agreement studies, until new and better ways of assessing agreement are found.

4.2 Introduction

The importance of agreement studies in medicine has been stressed in leading medical journals.[1,2] Reproducibility studies assess the agreement of measurements obtained by two or more observers, or by one observer on two or more occasions under the same conditions. These are termed inter-observer and intra-observer agreement, respectively. Validity studies assess the agreement of measurements with a reference measure (the gold standard), and should only be performed after a reproducibility study. Randomised controlled trials should only be performed after adequate reproducibility and validity studies.[1–4] The important question is: how can researchers, doctors and lawyers interpret similar studies with results that disagree, and take decisions based on these results?

However, the importance of agreement studies is not always given due recognition. Moreover, in many cases, inappropriate or controversial measures of agreement are used, particularly regarding reproducibility and validity studies with the same scale for the test under scrutiny and the gold standard.[1–4]

In order to help clear up this problem, we decided to assess how agreement has been measured in reproducibility and validity studies with the same scale for the test under scrutiny and the gold standard in three world-leading journals in obstetrics and gynaecology. We also illustrate how poor reproducibility can dramatically influence the results of validity studies, randomised controlled trials and meta-analyses.

4.3 Methods

All abstracts from the American Journal of Obstetrics and Gynecology, the British Journal of Obstetrics and Gynaecology and Obstetrics and Gynaecology published between the beginning of 1997 and the end of 2001 were reviewed, excluding letters and editorials. The review was performed by the same statistician, who is experienced in agreement studies. Another statistician performed the same review for one random year for each journal and found the same results.

All reproducibility and validity studies with the same scale for the test under scrutiny and the gold standard were selected, excluding validity studies considering dichotomic variables (sensitivity and specificity studies).The full

text of all selected articles was analysed and discussed by two statisticians to discover what measures were used: kappa (k), proportions of agreement (PA), intraclass correlation coefficient (ICC), limits of agreement (LA) or any other measure.

k is the observed agreement (i.e. PA, minus the agreement expected by chance, divided by perfect agreement minus the agreement expected by chance). Thus, this measure provides measurement of agreement beyond chance in nominal scales. In ordinal scales, and when the discrepancy between some categories is much worse than between others, weighted k corrects the observed agreement not only for chance but also allows credit for partial agreement.[3] ICC is a measure of agreement for continuous variables, and measures both the variation between the observers and the variation between subjects. Thus, ICC is the proportion of the total variance that is due to the variation between the subjects.[3] To measure agreement in continuous variables, Bland and Altman recommended LA, which involves a scatter plot of difference between the measurements obtained by the two observers against the mean of the measurements for each subject, and LA is expressed as a range which encompasses 95% of these differences.[1,2]

To demonstrate how poor reproducibility can influence the results of validity studies, randomised controlled trials and meta-analysis, we studied two hypothetical case examples.

4.4 Results

The literature search in the three leading journals over the five-year period considered revealed 23 reproducibility or validity studies with the same scale for the test under scrutiny and the gold standard, 11 using categorical variables and 15 using continuous variables. Some studies involved both kinds of variables. The list of agreement studies analysed is available from the authors upon request. With categorical variables, agreement was measured with k only or weighted n only in 73% of cases, never with PA without any other measure, and with both measures in 18% of cases. With continuous variables, agreement was measured with ICC only in 33% of cases, in 13% only with LA and in 27% with both measures. In 9% of cases with categorical variables and 27% with continuous variables, clearly inadequate measures were used such as the Pearson correlation coefficient, Spearman correlation coefficient and qui-square.

In Table 4.1, an example illustrates how poor reproducibility can influence the result of a validity study. A commonly observed observer variation in a diagnostic test led to variations in the test sensitivity as wide as 0 to 100%. How can we interpret the results for these three similar trials in the same eight cases, to access the validity of this diagnosis test, with different results? Is sensitivity of this diagnosis test 0%, as suggested by trial 2 or 100%, as suggested by trial 3?

Table 4.1. Assessment of the same diagnostic test (positive or negative test) in three trials with the same eight cases. There is poor reproducibility among the three trials. With the final outcome in mind, in trial 1 the sensitivity is 50% (1/2) and false positive rate is 17% (1/6); in trials 2 and 3 sensitivities range between 0 (0/2) and 100% (2/2), respectively, and false positive rates range between 33% (2/6) and 0 (0/6).

Cases	Trial 1	Trial 2	Trial 3	Outcome
1	+	-	+	+
2	-	-	+	+
3	+	+	-	-
4	-	+	-	-
5	-	-	-	-
6	-	-	-	-
7	-	-	-	-
8	-	-	-	-

If, on the contrary, there was perfect reproducibility in the three trials, for example, trials 2 and 3 equal to trial 1, the sensitivity and the false positive rate would be always the same, 50% and 17%, respectively, in all trials.

In Table 4.2, an example illustrates how poor reproducibility in a diagnostic test A can lead to a variation in the relative risk in three trials comparing test A plus an 100% effective intervention with test B plus a 100% effective intervention as wide as 0.1 to 2. How can we interpret the results for the three similar trials? Is test A and a 100% effective intervention worse than test B and a 100% effective intervention, as suggested by trial 1 with a relative risk of 2? Or is test A and a 100% effective intervention better than test B and a 100% effective intervention, as suggested in trial 3 with a relative risk of 0.1?

If, on the contrary, there was perfect reproducibility in test A, as there was in test B, the relative risk would always be the same in the three trials.

Table 4.2. Assessment of an outcome prevalence after diagnostic test A plus a 100% effective intervention and after a diagnosis test B plus a 100% effective intervention in three trials. There is no observer agreement in test A among the three trials, whereas there is perfect agreement in test B. The relative risks between test A plus effective intervention and test B plus effective intervention range between 0.1 and 2, in the three trials.

Trial	Test A + effective intervention	Test B + effective intervention	Relative risk
	Outcome prevalence	Outcome prevalence	
1	20/100	10/100	2
2	10/100	10/100	1
3	1/100	10/100	0.1

4.5 Discussion

Our results show that the k and ICC presented without any other measure are the most common ways used to assess agreement in categorical and continuous variables, respectively. In fact, these measures are advocated by some authors as being the best measures of agreement. However, they have been strongly contested by other authors.[1,2] k and ICC measure association rather then agreement[1,2] and are affected by prevalence of the categories in assessment and by observer bias, which may result in potentially misleading interpretations.5 On the other hand, LA and PA measure agreement, are not affected by differences in prevalence of the categories in the assessment and may show existing observer bias.[1,2] However, they do not take the variation between subjects into account, and their interpretation depends on subjective comparison with the range of the measurements encountered in clinical practice.

Our results also show that the correlation coefficient (Spearman), qui-square test and Pearson correlation coefficient are sometimes used as measures of agreement. It has been demonstrated repeatedly that they measure association, not agreement.[1–4] A complete disagreement can exist, having a perfect association with a correlation coefficient of 1. Our examples also show how observer disagreements commonly observed in reproducibility studies may lead to discrepancies in validity studies and randomised controlled trials. Under-estimation, ignorance or, even worse, wrong assumptions about this situation may have obvious implications for research and clinical practice, and also

medico legal consequences. Indeed, how can we interpret the results from most of the studies regarding the sensitivity and specificity of some of the most common screening and diagnostic tests in obstetrics and gynaecology? What is, for example, the sensitivity and specificity of colposcopy and of cervical cytology in detection of CIN II and III? Mitchell et al.[6] found a variation in sensitivity and specificity of colposcopy ranging from 87% to 100% and 26 to 87%, respectively, in nine published studies using similar methods. The same authors found a variation in sensitivity and specificity of cervical cytology from 0 to 93% and 0 to 100%, respectively, in 26 published studies using similar methods.[6] And, what are, for example, the relative risks of poor outcomes generally supposed to be avoided by interventions based on cardiotocography and mammography? Are they really close to 1, as suggested by most meta-analyses,[7,8] or are they just the mean result of discrepant results from the different randomised clinical trial included in meta-analyses?

Agreement studies, involving both reproducibility and validity, should be performed and interpreted in an adequate manner, and until better ways of assessing agreement are developed, we recommend for categorical variables the use of both k and PA for each category and overall, and for continuous variables, the use of both ICC and LA in reproducibility and validity studies with the same scale for the test under scrutiny and the gold standard.

4.6 References

1. Bland JM, Altman DG. Statistical methods for assessing two methods of clinical measurement. Lancet 1986;1:307– 310.

2. Grant JM. The fetal heart rate is normal, isn't it? Observer agreement of categorical assessments. Lancet 1991;337:215– 218.

3. Khan K, Chien P. Evaluation of a clinical test. I: Assessment of reliability. Br J Obstet Gynaecol 2001;108:562– 567.

4. Chien P, Khan K. Evaluation of a clinical test. II: Assessment of validity. Br J Obstet Gynaecol 2001;108:568– 572.

5. Feinstein AR, Cicchetti DV. High agreement but low kappa. I: The problems of two paradoxes. J Clin Epidemiol 1990;43:543– 549.

6. Mitchell MF, Cantor SB, Ramanujam N, Tortolero-Luna G, Richards-Kortum R. Flourescence spectroscopy for diagnosis of squamous intraepithelial lesions of the cervix. Obstet Gynecol 1999;93:462–470.

7. Thacker SB, Stroup DF, Peterson HB. Efficacy and safety of intrapartum electronic fetal monitoring: an update. Obstet Gynecol 1995;86:613–620.

8. Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet 2001;358(9290):1340– 1342.

5. Frequent problems in the design and analysis of agreement studies with categorical variables.

5 Frequent problems in the design and analysis of agreement studies with categorical variables.

This chapter is based on 3 letters to the Editor. Each letter is a different example of a common problem found in observer agreement studies considering categorical variables. The letters refer to studies which assess observer agreement in electronic foetal monitoring and they indentify design problems, the use of inappropriate statistical methods, incomplete analysis and missing information.

5.1 Introduction

In spite of the importance of agreement studies, it is not uncommon to find in the medical literature agreement studies with serious shortcomings in the design, statistical methods, and interpretation of results. This chapter is based on 3 letters to the Editor (Costa-Santos 2008-a; Costa-Santos 2008-b; Costa-Santos 2009) focusing on different common problems found in observer agreement studies with categorical variables in obstetrics and gynaecology. All the letters to the Editor identify problems in studies that assess observer agreement in electronic foetal monitoring.

Electronic foetal monitoring is a method for examining the condition of a foetus by noting any unusual changes in its heart rate. Electronic foetal monitoring is performed late in pregnancy, or continuously during labour to ensure normal delivery of a healthy baby.

Cardiotocography (CTG) is a technique for electronic monitoring of foetuses by recording foetal heartbeat and uterine contractions, which can be used to identify signs of foetal distress. The appendix of this chapter contains the 3 letters.

5.2 Design problems and inappropriate statistical methods

Frequently, the problems of agreement studies begin in the study design. For example, in the study by Larma et al. (Larma 2007), which aimed to determine whether electronic foetal monitoring could identify foetuses with metabolic acidosis and hypoxic-ischemic encephalopathy, the last hour of electronic foetal monitoring before delivery was reviewed by 3 obstetricians who were blinded as to the outcome. The results of this study suggest good to high observer agreement in the detection and classification of decelerations and late decelerations. However, this result can be misleading due to the study design (Costa-Santos 2008-a). In fact, the number of decelerations in a cardiotocographic tracing within an hour is not the ideal way of measuring the observer agreement in the detection and classification of decelerations. Actually, two observers can identify the same number of decelerations in a cardiotocographic tracing (suggesting an excellent degree of agreement) but the

decelerations identified can be in different segments (not such good agreement). Consequently, the observer agreement in the detection of deceleration may have been overestimated in the study by Larma et al. (Larma 2007). In fact, these results contrast with observations showing fair to poor observer agreement in deceleration detection. A previous study assessed the agreement over the identification of each deceleration rather than the total number of decelerations identified and the results showed fair to poor observer agreement in deceleration detection (Bernardes 1997).

Moreover, in the study by Larma et al. (Larma 2007), inappropriate statistical methods to assess agreement were used (Costa-Santos 2008-a). The Pearson correlation coefficient was used to assess observer agreement on the number of decelerations, but the Pearson correlation coefficient assesses association rather than agreement, and high values of a Pearson correlation coefficient do not necessarily reflect good agreement—as the Pearson correlation coefficient does not detect systematic bias between observers. The relevant letter to the editor (Costa-Santos 2008-a) is reproduced in Appendix 1.

5.3 Incomplete analysis

It is also common to find agreement studies with an appropriate design but with an incomplete analysis. This is the case in the study by Ojala et al. (Ojala 2008) on inter-observer agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies (Costa-Santos 2008-b). This article reports proportions of agreement (PA) and kappa statistics (k) of 0.56 to 0.59 and 0.47 to 0.48, respectively, in agreement of classification of cardiotocography (Table 5.1). It also reports PA and k values of 0.80 to 0.85 and 0.47 to 0.60, respectively, in agreement of clinical decision making (Table 5.2). The authors concluded that interobserver agreement among experienced obstetricians in the classification of cardiotocography and clinical decision making according to STAN guidelines may be "moderate at best". However, the interobserver agreement among experienced obstetricians may be even worse. According to our estimations based on tables provided in the article (Ojala 2008) (Tables 5.1 and 5.2), if we analyzed the most critical clinical decisions rather than just total agreement, we found PA and k values for agreement on pre-terminal tracings of 0.0 and -0.0025 to -0.0044, respectively, and PA and k to be in agreement on interventional labor management (0.42 to 0.53 and 0.33 to 0.46, respectively). In that study, the agreement in most critical

clinical decisions is clinically even more important than the overall agreement. In fact, the k and PA values specific for each category are especially informative, but they are unfortunately seldom reported (Costa-Santos 2008-b). The relevant letter to the editor (Costa-Santos 2008-b) is reproduced in Appendix 2.

Table 5.1: Table from the study by Ojala et al. (Ojala 2008) study providing the interobserver agreement between three observers in classifying the cardiotocographic recordings as normal, intermediary, abnormal, or preterminal according to automated fetal electrocardiogram ST-analysis guidelines.

Table II. The interobserver agreement between three observers in classifying the cardiotocographic recordings as normal, intermediary, abnormal, or preterminal according to automated fetal electrocardiogram ST-analysis guidelines (4).

	Normal	Intermediary	Abnormal	Preterminal	Total
Observer 1 versus Observer 2†					
Observer 1			Observer 2		
Normal	48	28	1	0	77
Intermediary	10	52	10	0	72
Abnormal	2	29	18	1	50
Preterminal	0	0	1	0	1
Total	60	109	30	1	200
Observer 1 versus Observer 3‡					
Observer 1			Observer 3		
Normal	55	15	7	0	77
Intermediary	18	27	27	0	72
Abnormal	5	8	30	7	50
Preterminal	0	0	1	0	1
Total	78	50	65	7	200
Observer 2 versus Observer 3§					
Observer 2			Observer 3		
Normal	51	6	3	0	60
Intermediary	27	40	38	4	109
Abnormal	0	4	23	3	30
Preterminal	0	0	1	0	1
Total	78	50	65	7	200

†Total agreement: 59%; weighted kappa: 0.47 (95% CI: 0.38–0.56).
‡Total agreement: 56%; weighted kappa: 0.47 (95% CI: 0.37–0.56).
§Total agreement: 57%; weighted kappa: 0.48 (95% CI: 0.40–0.56).
CI, confidence interval.

Reprinted from Acta Obstet Gynecol Scand, 87, Ojala, K, Makikallio, K, Haapsamo, M, Ijas, H and Tekay, A, Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies, 536-540, Copyright (2008), with permission from Taylor & Francis.

Table 5.2: Table from the study by Ojala et al. (Ojala 2008) study providing the interobserver agreement between three observers in managing labor either conservatively or interventionally following STAN clinical guidelines.

Table III. The interobserver agreement between three observers in managing labor either conservatively or interventionally following STAN clinical guidelines (4).

	Conservative	Interventional	Total
Observer 1 versus Observer 2†			
Observer 1		Observer 2	
Conservative	145	15	160
Interventional	15	25	40
Total	160	40	200
Observer 1 versus Observer 3‡			
Observer 1		Observer 3	
Conservative	138	22	160
Interventional	7	33	40
Total	145	55	200
Observer 2 versus Observer 3§			
Observer 2		Observer 3	
Conservative	133	27	160
Interventional	12	28	40
Total	145	55	200

†Total agreement: 85%; kappa: 0.53 (95% CI: 0.38–0.68).
‡Total agreement: 86%; kappa: 0.60 (95% CI: 0.47–0.73).
§Total agreement: 80%; kappa: 0.47 (95% CI: 0.33–0.61).
CI, confidence interval; STAN, automated fetal electrocardiogram ST-interval analysis.

Reprinted from Acta Obstet Gynecol Scand, 87, Ojala, K, Makikallio, K, Haapsamo, M, Ijas, H and Tekay, A, Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies, 536-540, Copyright (2008), with permission from Taylor & Francis.

5.4 Missing information

Some agreement studies do not provide sufficient information about the data analysis methodology and the results become difficult to interpret. For example, Chauhan et al. (Chauhan 2008) showed interesting differences and consequences in interobserver agreement in assessment of fetal heart rate compared to other studies (Ayres-de-Campos 2004; Bernardes 1997). In fact, as the former authors claim, interobserver variation in FHR assessment is

excessive and undesirable. However, Chauhan et al. found very low kappa values (near 0) in FHR baseline estimation using the National Institute of Child Health and Human Development guidelines (Table 5.3), whereas in several other studies using the International Federation of Gynecology and Obstetrics guidelines, we have consistently found very high values (near 1)—namely, when nonreassuring FHR tracings were also analyzed in the final minutes of labour (Ayres-de-Campos 2004).

Table 5.3: Table from the study by Chauhan et al. (Chauhan 2008) study providing the interobserver variability of intrapartum foetal heart rate tracing.

TABLE 1
Interobserver variability of intrapartum FHR tracing

Agreement	Kappa coefficient	Early labor[a]	Before delivery[b]
Poor	0.00-0.19	Baseline	Baseline
		Accelerations	Accelerations
		Bradycardia	Bradycardia
		Beat-to-beat variability, decreased	Beat-to-beat variability, decreased
		Beat-to-beat variability, absent	—
		Recurrent prolonged decelerations	Recurrent prolonged decelerations
		Variable deceleration with slow return	Variable decelerations with slow return
		—	Recurrent late decelerations
Fair	0.20-0.29	Recurrent late decelerations	Beat-to-beat variability, absent
		Recurrent severe variable decelerations	
Moderate	0.30-0.43	—	Recurrent severe variable decelerations
Good	0.44-0.59	Tachycardia	Tachycardia
Substantial	0.60-0.80	—	—
Almost perfect	0.81-1.00	—	—

[a] FHR tracing of 1-hour duration, before the onset of periodic decelerations. [b] FHR tracing of 1-hour before birth.

Chauhan. Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes. Am J Obstet Gynecol 2008.

Reprinted from American Journal of Obstetrics and Gynecology, 199, Suneet P. Chauhan, Chad K. Klauser, Thomas C. Woodring, Maureen Sanderson, Everett F. Magann, John C. Morrison, Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability, e621-625, Copyright (2008), with permission from Elsevier.

These differences may result from the use of different guidelines for FHR assessment, but are probably also due to different categorization of variables and/or different statistical methods for assessment of observer variation (Costa-Santos 2009). Due to the lack of information provided in the article by Chauhan et al., we cannot understand the source of the differences between the results of this study and other studies. In fact, in spite of the fact that baseline was a continuous variable, it was analyzed with methods appropriate for categorical variables (kappa statistics). The kappa statistics depends on the number of categories; a variable with more

categories will provide lower kappa values (Cook 2005; Landis 1977). However, the article does not provide information on how many and what categories were used to categorize the baseline variable.

The relevant letter to the editor (Costa-Santos 2009) is reproduced in Appendix 3.

5.5 References

Ayres-de-Campos, D, Bernardes, J, Marsal, K, Nickelsen, C, Makarainen, L, Banfield, P, Xavier, P and Campos, I (2004). "Can the reproducibility of fetal heart rate baseline estimation be improved?" Eur J Obstet Gynecol Reprod Biol 112(1): 49-54.

Bernardes, J, Costa-Pereira, A, Ayres-de-Campos, D, van Geijn, HP and Pereira-Leite, L (1997). "Evaluation of interobserver agreement of cardiotocograms." Int J Gynaecol Obstet 57(1): 33-37.

Chauhan, SP, Klauser, CK, Woodring, TC, Sanderson, M, Magann, EF and Morrison, JC (2008). "Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability." Am J Obstet Gynecol 199(6): 623 e621-625.

Cook, R (2005). Kappa and its dependence on marginal rates. The Encyclopedia of Biostatistics. Armitage, P. New York, Wiley: 2166-2168.

Costa-Santos, C, Bernardes, J and Ayres-De-Campos, D (2008-a). "Assessment and reporting of the reproducibility of fetal heart rate monitoring." Am J Obstet Gynecol 198(3): 343; author reply 343-344.

Costa-Santos, C and Bernardes, J (2008-b). "Interobserver agreement according to STAN guidelines." Acta Obstet Gynecol Scand 87(11): 1260.

Costa-Santos, C and Bernardes, J (2009). "Differences and consequences in interobserver variation in intrapartum fetal heart rate assessment." Am J Obstet Gynecol 200(6): e8.

Landis, JR and Koch, GG (1977). "The measurement of observer agreement for categorical data." Biometrics 33(1): 159-174.

Larma, JD, Silva, AM, Holcroft, CJ, Thompson, RE, Donohue, PK and Graham, EM (2007). "Intrapartum electronic fetal heart rate monitoring and the identification of metabolic acidosis and hypoxic-ischemic encephalopathy." Am J Obstet Gynecol 197(3): 301 e301-308.

Ojala, K, Makikallio, K, Haapsamo, M, Ijas, H and Tekay, A (2008). "Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies." Acta Obstet Gynecol Scand 87(5): 536-540.

5.6 Appendix 1

LETTERS TO THE EDITORS www.AJOG.org

Assessment and reporting of the reproducibility of fetal heart rate monitoring

TO THE EDITORS: We read with interest the paper by Larma et al ("Intrapartum electronic fetal heart rate monitoring and the identification of metabolic acidosis and hypoxic-ischemic encephalopathy").[1] However, we are concerned with a continuing uncertainty regarding reproducibility assessment and reporting.[2] Larma et al reported Pearson correlation coefficients of 0.79 for observers' agreement on the total number of decelerations and of 0.66 for observers' agreement on the number of late decelerations, suggesting good to high reproducibility in detection and classification of these features.[1]

This is in contrast with previous observations putting in evidence fair to poor reproducibility in deceleration detection, with a proportion of agreement (PA) of 0.51 (95% confidence interval [CI], 0.46 to 0.56) and a kappa statistic (K) of 0.49, for observers' agreement on decelerations, in general, and a PA of 0.24 (95% CI, 0.11 to 0.37) and a K of 0.21 for observers' agreement on late decelerations, in particular.[3] Because the Pearson correlation coefficient measures association rather then agreement,[4] the values reported by Larma et al do not necessarily reflect good agreement between observers. Indeed, if, for example, observer A systematically detects 10 times more decelerations in each tracing than observer B, the Pearson correlation coefficient between observer A and B will be 1, reflecting a perfect association but a total or an almost total disagreement.

Agreement over the identification of each deceleration, rather than the total number of decelerations identified, is surely an important issue for assessing the reproducibility of this parameter.[2,3] Use of statistical measures of agreement (proportions of agreement, kappa statistic, limits of agreement) rather than measures of association (correlation coefficient) is another crucial aspect when evaluating interobserver reproducibility.[2-4] ■

Cristina Costa Santos, MSc
Bioestatística e Informática Médica
Faculdade de Medicina da Universidade do Porto
Alameda Hernani Monteiro
Porto 4200-319, Portugal
csantos@med.up.pt

João Bernardes, MD, PhD
Diogo Ayres-De-Campos, MD, PhD
Departamento de Ginecologia e Obstetrícia
Faculdade de Medicina da Universidade do Porto
Alameda Hernani Monteiro
Porto 4200-319, Portugal

REFERENCES

1. Larma JD, Silva AM, Holcroft CJ, Thompson RE, Donohue PK, Graham EM. Intrapartum electronic fetal heart rate monitoring and the identification of metabolic acidosis and hypoxic-ischemic encephalopathy. Am J Obstet Gynecol 2007;197:301.e1-8.
2. Bernardes J, Ayres-de-Campos D, Costa-Pereira A. Further concerns about the National Institute of Child Health and Human Development guidelines for interpretation of electronic fetal heart rate monitoring. Am J Obstet Gynecol 2001;184:1587-8.
3. Bernardes J, Costa-Pereira A, Ayres-de-Campos D, van Geijn HP, Pereira-Leite L. Evaluation of interobserver agreement of cardiotocograms. Int J Gynaecol Obstet 1997;57:33-7.
4. Grant JM. The fetal heart rate trace is normal, isn't it? Observer agreement of categorical assessments. Lancet 1991;337:215-8.

© 2008 Mosby, Inc. All rights reserved. doi: 10.1016/j.ajog.2007.11.037

Costa-Santos C, Bernardes J, Ayres de Campos D. Assessment and reporting of the reproducibility of fetal heart rate monitoring. *Letter to the Editors*. **Am J Obstet Gynecol.** 2008;198(3):343

5.7 Appendix 2

Acta Obstetricia et Gynecologica. 2008; 87: 1260

informa
healthcare

LETTER TO THE EDITOR

Interobserver agreement according to STAN guidelines

Sir,

We read with interest the paper by Ojala et al. on interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies (1). The paper alerts clinicians and researchers to the continuing need of using intrapartum automated fetal electrocardiography in a reproducible way, something that one may expect to be improved by training and automatization (2). In fact, although, the effectiveness of the Fetal electrocardiogram ST Analyser (STAN) technology in fetal surveillance improvement during labour has been proved in several studies (3), this may be lost, if we can not use it in a reproducible way (1,4). The article by Ojala et al. shows that interobserver agreement among experienced obstetricians in the classification of cardiotocography and clinical decision making according to STAN guidelines may be 'moderate at best', considering total agreement (1). This may be even worse, if we analyse the most critical clinical decisions rather than just total agreement, with Proportions of Agreement (PA) and Kappa statistics (κ) for agreement on pre-terminal tracings of 0.0 and −0.0025 to −0.0044, respectively, and PA and κ in agreement on interventional labor management of 0.42–0.53 and 0.33–0.46, respectively, with large confidence intervals, according to our estimations based on Tables II and III in the article by Ojala et al. (1).

Cristina Costa Santos and João Bernardes

Address for correspondence
Biostatistics and Medical Informatics Department
and Obstetrics and Gynecology Department
Faculty of Medicine
University of Porto
Portugal
E-mail: csantos@med.up.pt

References

1. Ojala K, Mäkikallio K, Haapsamo M, Ijäs H, Tekay A. Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies. Acta Obstet Gynecol Scand. 2008;87:536–40.
2. Ayres-de-Campos D, Sousa P, Costa A, Bernardes J. Omniview-SisPorto 3.5 – a central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis. J Perinat Med. 2008;36:260–4.
3. Neilson JP. Fetal electrocardiogram (ECG) for fetal monitoring during labour. Cochrane Database Syst Rev. 2006;3:CD 000116.
4. Costa Santos C, Costa Pereira A, Bernardes J. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences. BJOG. 2005;112:667–9.

Costa-Santos C, Bernardes J. Interobserver agreement according to STAN guidelines. *Letter to the Editors*. **Acta Obstet Gynecol** Scand. 2008; 87:1260.

5.8 Appendix 3

LETTERS TO THE EDITORS www.AJOG.org

Differences and consequences in interobserver variation in intrapartum fetal heart rate assessment

TO THE EDITORS: We read with interest the article by Chauhan et al,[1] which showed interesting differences and consequences in interobserver agreement in fetal heart rate (FHR) assessment compared with other studies.[2,3]

We agree with the authors that interobserver variation in FHR assessment is excessive and undesirable. However, Chauhan et al[1] found very low (near 0) Kappa values in FHR baseline estimation using the National Institute of Child Health and Human Development guidelines (Table 1), whereas in several other studies, using the International Federation of Gynecology and Obstetrics guidelines, we have consistently found very high (near 1) values—namely, when nonreassuring FHR tracings were also analyzed in the final minutes of labor.[2] A similar situation occurred with accelerations assessment and the opposite with decelerations.[1,3] These differences may result from the use of different guidelines for FHR assessment, different categorization of variables, and/or different statistical methods for observer variation assessment, which still remain a source of controversy.[4] Further research should contribute to clarify these issues.

Chauhan et al[1] also showed consequences of poor interobserver agreement in FHR assessment, leading to dramatic discrepancies in prediction of neonatal outcome, sometimes ranging from 0-100%, as occurred with the sensitivity in prediction of base excess ≥ 12 mmol/L (Table 2). These and other consequences—namely, on randomized controlled trials and metaanalysis—should always be kept in mind in clinical practice and research.[4] ■

Cristina Costa Santos, MSc
Biostatistics and Medical Informatics Department
Faculty of Medicine
University of Porto
Al. Prof. Hernani Monteiro, 4200-319 Porto, Portugal

João Bernardes, MD, PhD
Obstetrics and Gynecology Department
Faculty of Medicine
University of Porto
Porto, Portugal

REPLY DECLINED

REFERENCES

1. Chauhan SP, Klauser CK, Woodring TC, Sanderson M, Magann EF, Morrison JC. Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability. Am J Obstet Gynecol 2008;199:623.e1-5.

2. Ayres-de-Campos D, Bernardes J, Marsal K, et al. Can the reproducibility of fetal heart rate baseline estimation be improved? Eur J Obstet Gynecol Reprod Biol 2004;112:49-54.

3. Bernardes J, Costa-Pereira A, Ayres-de-Campos D, van Geijn HP, Pereira-Leite L. Evaluation of interobserver agreement of cardiotocograms. Int J Gynaecol Obstet 1997;57:33-7.

4. Costa Santos C, Costa Pereira A, Bernardes J. Agreement studies in obstetrics and gynecology: inappropriateness, controversies and consequences. BJOG 2005;112:667-9.

© 2009 Mosby, Inc. All rights reserved. doi: 10.1016/j.ajog.2008.12.013

Santos C C, Bernardes J. Differences and consequences in interobserver variation in intrapartum fetal heart rate assessment. *Letter to the Editors*. Am J Obstet Gynecol 2009;200(6):e8.

6. The limits of agreement and the intra-class correlation coefficient may be inconsistent in assessing agreement.

6 The limits of agreement and the intra-class correlation coefficient may be inconsistent in assessing agreement.

Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A and Costa C.

J Clin Epidemiol 2010 (in press).

6.1 Abstract

Objective: To compare the interpretation of agreement in the prediction of neonatal outcome variables, using the limits of agreement (LA) and the intraclass correlation coefficient (ICC).

Study Design and Setting: Three obstetricians were asked to predict neonatal outcomes independently based on the evaluation of intrapartum cardiotocographic tracings. Interobserver agreement was assessed with the LA and the ICC, and the results obtained were interpreted by six clinicians and six statisticians on a scale that established agreement as very poor, poor, fair, good, or very good.

Results: Interpretation of the LA results was less consensual than the ICC results, with proportions of agreement of 0.36 (95% confidence interval [CI]: 0.28-0.44) vs. 0.63 (95% CI: 0.54-0.73), respectively. LA results suggested a fair to good agreement among obstetricians, whereas interpretation of ICC results suggested a poor to fair agreement. LA results were more plausible with reality, suggesting that obstetricians predicted neonatal outcomes better than randomly generated values, whereas it was not always the case with the ICC.

Conclusions: LA and ICC can provide inconsistent results in agreement studies. Accordingly, in the absence of better strategies to assess agreement, both should be used for this purpose, but their results need to be interpreted with caution keeping their respective limitations in mind.

6.2 Introduction

Observer disagreement on the interpretation of diagnostic tests and clinical decisions remains an important problem in Medicine [1,2]. The influence that this aspect may have on the validity and effectiveness of clinical interventions has been well demonstrated in the past [2,3]. However, inappropriate statistical methods to assess agreement and misleading interpretations of agreement results are frequently found in the scientific literature [3,4].

The intraclass correlation coefficient (ICC) is claimed to be suitable for the assessment of observer agreement [1]; Bland and Altman proposed the limits of agreement (LA) to assess agreement between methods of measurement [5]. The ICC and LA, or a combination of both, have been the most commonly used strategies for assessing agreement on continuous variables in obstetrics and gynecology [3]. The ICC is supported by the analysis of variance (ANOVA) [6]. The main limitation of this method resides in its strong dependence on the variance of the assessed population. Higher ICC values may, thus, be obtained when applied to a more heterogeneous population as compared with a more homogeneous one despite similar levels of agreement [7]. Consequently, the ICC values cannot be said to translate an absolute level of agreement, and the cutoff value of 0.75, proposed by Burdock et al. to signify good agreement [8], has very limited justification. The LA, on the other hand, are not influenced by the variance of the assessed population, but interpretation of the significance of its results is highly subjective, as it relies on individuals' understanding of the clinical importance of the obtained range [5]. The aim of this study was to compare the interpretation of agreement results using the ICC and the LA in the clinical prediction of widely used neonatal outcome variables, the Apgar score and the umbilical artery blood (UAB) pH.

6.3 Methods

Three obstetricians were asked to evaluate 72 intrapartum cardiotocographic tracings independently and, from them, predict 1- and 5-minute Apgar scores and UAB pH. All obstetricians had more than 5 years' experience in cardiotocography (CTG) interpretation and had previously authored scientific papers on intrapartum fetal heart rate monitoring. The 72 tracings were presented to the obstetricians with information on maternal age, parity,

gestational age, underlying pathology, and drugs administered during labor. Tracings were consecutively selected from a preexisting database of the term singleton gestations, delivered vaginally, with at least 60 minutes of tracing length, less than 15% signal loss, an interval of less than 5 minutes to delivery, and no known fetal malformations. In each case, only the last 60 minutes before birth were evaluated. All tracings had been previously acquired with Sonicaid (Oxford Instruments, Abingdon, UK) or Hewlett-Packard (Palo Alto,CA,USA) fetal monitors using an internal fetal scalp electrode and an external uterine activity detector.

What is new?

- Important inconsistencies in interpretation of intraclass correlation coefficient and limits of agreement results were found, when these statistics were used to assess agreement in prediction of Apgar score and umbilical artery blood pH.

- The inconsistencies observed in this study may largely be explained by the intrinsic dependence of the intraclass correlation coefficient on variance, and the subjectivity that occurs in interpretation of the limits of agreement. These limitations have been identified before, but their consequences on the interpretation of results had not been demonstrated in this fashion.

- The interpretation of agreement results on continuous variables, using intraclass correlation coefficient and limits of agreement, should be performed with caution, as it may show considerable inconsistencies. The limitations of these methods should be kept in mind when considering the clinical application of the results of agreement studies.

Apgar scores were evaluated by the caregivers responsible for immediate neonatal support. The umbilical cord was doubly clamped immediately after birth, and samples were obtained within 30 minutes to evaluate artery blood gas values. The Apgar score is widely used in developed countries for assessing the clinical status of the newborn determining whether or not there is a need for immediate medical care. It ranges between 0 and 10 and is based on the analysis of the newborn's skin color, response to stimulus, breathing, muscle tonus, and heart rate [9]. The UAB pH is the most objective, currently available measure of fetal oxygenation. A low UAB pH establishes the presence of respiratory and/or metabolic acidemia occurring during labor and delivery, representing a higher risk of perinatal death or neurological injuries from hypoxia [10].

To simulate an observer who estimated Apgar scores and UAB pH by chance, we randomly generated Apgar score values between 0 and 10 and pH values within the range obtained in these newborns (7.049 - 7.390).

Agreement between pairs of obstetricians and among all three of them was assessed by the ICC (two-way random effects) [11] and the LA. Agreement among individual obstetricians and real outcome values (Apgar and UAB pH) was assessed using the same statistics, as was the case with randomly generated numbers.

ICC assesses agreement by comparing the variance of different measurements of the same subject made by different observers with the total variance across all measurements and subjects [6]. These variances are derived from ANOVA, and the ANOVA model (one way or two way with random or fixed effects) depends on the agreement study design. The ICC ranges from 0 (no agreement) to 1 (perfect agreement), but it can also be negative. The interpretation of negative ICC values remains unclear.

LA is calculated based on the mean difference between two measurements in the same subjects and the standard deviation of these differences. Approximately 95% of these differences will lie between the mean differences ±1.96 standard deviations of these differences [5].

In the second part of the study, interpretation of the meaning of the obtained ICC and LA results was performed independently by 12 experts whowere asked to attribute a score in a 5-point Likert scale, reflecting the following: 1-very poor, 2-poor, 3-fair, 4-good, or 5-very good agreement. This panel of experts was constituted by six clinicians, all with ongoing or concluded PhD or MSc studies in the area of cardiotocography and six biostatisticians with a background in mathematics, all with published papers in medical statistics. All experts were familiar with the concepts and numerical ranges of

the Apgar score, UAB pH, ICC, and LA, but as a safeguard, they received a written reminder of these at the time of evaluation stating: "The Apgar score is a scoring system to assess the clinical status of a newborn infant. It is usually assessed at the first and fifth minutes after birth and can range from zero to ten. Umbilical artery blood is collected immediately after birth and the obtained pH is an objective indicator of fetal acidemia, usually ranging from 7.05 to 7.40. When lower than 7.05 it is associated with an important fetal acidemia. When limits of agreement are considered to assess interobserver agreement, approximately 95% of the differences between two observers will lie between their lower and upper limits. For example, if the limits of agreement between observer A and B in prediction of an Apgar score are [-2;2], we can say that in 95% of cases observer A predicted between 2 points less and 2 points more than observer B. The intraclass correlation coefficient can assess agreement between two or more observers. Perfect agreement is represented by a coefficient of one, while zero represents total disagreement (although it can also be negative)." To further reduce bias, ICC and LA results for each of the assessed variables were provided to experts separately, in random order, and without any further information.

Agreement in interpretation of the ICC and LA results was evaluated with the proportions of overall agreement and 95% confidence intervals (CI). The proportions of overall agreement were obtained with generalized formulas for any number of raters with potentially different numbers of raters or different raters for each case [12], and the 95% CIs were obtained using nonparametric bootstrap [13].

The average value of experts' interpretation of agreement results was calculated to compare the evaluation of ICC with the evaluation of LA. Wilcoxon rank test was performed to compare these two interpretations, and the median value of agreement was computed for both. A significance level of 0.05 was chosen.

6.4 Results

Mean maternal age was 28 years, with a standard deviation of 6 years. Mean gestational age at delivery was 40.2 weeks, with a standard deviation of 1.4 weeks. Eighty-eight percent of cases had no identifiable underlying pathology. Mean newborn weight was 3,503 g, with a standard deviation of 505 g.

Apgar scores at 1 and 5 minutes were evaluated in all 72 cases, but valid UAB pH values were available only in 51 cases. Although there were 21 missing values in UAB pH, there were no significant differences in gestational age (p=0.295, t-test), birth weight (p=0.696, t-test), and maternal age (p=0.818, t-test) in the groups with and without UAB pH.

Apgar scores at 1 minute ranged between 4 and 10, with a variance of 0.962, and Apgar scores at 5 minutes ranged from 8 to 10, with a variance of 0.441. The variances of randomly generated Apgar scores were 7.7 at 1 minute and 7.8 at 5 minutes. The variances of 1-minute Apgar scores predicted by obstetricians A, B, and C were 0.7, 0.5, and 0.4, respectively, whereas those of 5-minute Apgar scores were 0.7, 0.3, and 0.3, respectively. UAB pH values ranged from 7.049 to 7.390, with a variance of 0.007. The variances of UAB pH values predicted by obstetricians A, B, and C were 0.005, 0.003, and 0.004, respectively. The variance of randomly generated UAB pH values was 0.009.

Interobstetrician, obstetrician-outcome, and random-outcome agreement results are presented in Tables 6.1 and 6.2, where the six clinicians' (Table 6.1) and six biostatisticians' (Table 6.2) interpretations of the obtained ICC and LA values are also included.

Among clinicians, the proportion of overall agreement in interpretations of ICC results was significantly higher (0.628; 95% CI: 0.540-0.730) than for LA results (0.359; 95% CI: 0.276-0.441). Similar results were obtained among biostatisticians (0.749, 95% CI: 0.641-0.851 vs. 0.438, 95% CI: 0.352-0.530).

Significant differences were found between the average interpretation of ICC and LA results, both by clinicians and by biostatisticians. In fact, both attributed a lower median agreement for ICC results than for LA results (for clinicians, 2.5 vs. 3.5, p<0.001; and for biostatisticians, 2.8 vs. 3.7, p=0.001).

Table 3 shows the agreement between obstetricians' prediction of Apgar scores and UAB pH and real outcomes assessed by ICC and LA with respective 95% CIs. LA results show that obstetricians predict 1- and 5-minute Apgar values significantly better than a random estimation (Table 6.3). On the other hand, ICC results suggest that these predictions are no better than those obtained by chance (Table 6.3).

Table 6.1: Interobstetrician, obstetrician-outcome, and random-outcome agreement in prediction of Apgar scores and UAB pH, considering ICC and LA, and percentage of **clinicians** (n=6) who chose each point of the Likert scale in interpretation of agreement results (higher percentages are in bold).

Agreement in prediction of	Agreement measure: ICC	Interpretation of ICC results: very poor	poor	fair	good	very good	Agreement measure: LA	Interpretation of LA results: very poor	poor	fair	good	very good
1-min Apgar score												
Inter-obstetrician:												
A–B	0.69			33	**67**		[-1.09,0.93]				**67**	33
A-C	0.53		17	**83**			[-0.93,1.63]			33	**67**	
B-C	0.52		17	**83**			[-0.98,1.79]			33	**67**	
A-B-C	0.57						-					
Obstetrician-outcome:												
A-outcome	0.17	**50**	**50**				[-2.06,2.06]			**50**	**50**	
B-outcome	0.29	17	**83**				[-2.01,2.01]			**50**	**50**	
C-outcome	0.06	**100**					[-2.42,2.62]		17	**83**		
Random -outcome	0.04	**100**					[-3.16,8.88]	**83**	17			
5-min Apgar score												
Inter-obstetrician:												
A–B	0.66			33	**67**		[-0.82,0.96]				**67**	33
A-C	0.40		33	**67**			[-0.84,1.87]			33	**67**	
B-C	0.47		17	**83**			[-0.87,1.76]			33	**67**	
A-B-C	0.48						-					
Obstetrician-outcome:												
A-outcome	0.18	**50**	**50**				[-1.48,1.48]			17	**67**	17
B-outcome	0.11	**67**	33				[-1.62,1.62]			33	**50**	17
C-outcome	0.00	**100**					[-2.08,2.08]			**50**	**50**	
Random -outcome	0.01	**100**					[-1.42,10.0]	**83**	17			
Umbilical artery blood												
pH Inter-obstetrician:												
A–B	0.79				**67**	33	[-0.08;0.08]		33	17		**50**
A-C	0.70				**100**		[-0.09;0.11]	17	17	17	17	**33**
B-C	0.57		17	**83**			[-0.10;0.12]	17	17	17	**33**	17
A-B-C	0.69						-					
Obstetrician-outcome:												
A-outcome	0.35		**50**	**50**			[-0.15;0.20]	17	**33**	17	**33**	
B-outcome	0.32		**50**	**50**			[-0.15;0.19]	17	**33**	17	**33**	
C-outcome	0.14	**67**	33				[-0.16;0.21]	17	**33**	17	**33**	
Random -outcome	-0.20	33	**67**				[-0.32;0.24]	**50**	17	17	17	

Abbreviations: UAB, umbilical artery blood; ICC, intraclass correlation coefficient; LA, limits of agreement; A, obstetrician A; B, obstetrician B; C, obstetrician C.

Table 6.2: Interobstetrician, obstetrician-outcome, and random-outcome agreement in prediction of Apgar scores and UAB pH, considering ICC and LA, and percentage of **biostatisticians** (n=6) who chose each point of the Likert scale in interpretation of agreement results (higher percentages are in bold).

Agreement in prediction of	Agreement measure: ICC	Interpretation of ICC results: very poor	poor	fair	good	very good	Agreement measure: LA	Interpretation of LA results: very poor	poor	fair	good	very good
1-min Apgar score												
Inter-obstetrician:												
A–B	0.69				**100**		[-1.09,0.93]			33	17	**50**
A-C	0.53			**100**			[-0.93,1.63]			33	**50**	17
B-C	0.52			**100**			[-0.98,1.79]			33	**50**	17
A-B-C	0.57						-					
Obstetrician-outcome:												
A-outcome	0.17	17	**83**				[-2.06,2.06]			**67**	33	
B-outcome	0.29		**67**	33			[-2.01,2.01]			**67**	33	
C-outcome	0.06	33	**67**				[-2.42,2.62]		33	**67**		
Random -outcome	0.04	33	**67**				[-3.16,8.88]	**100**				
5-min Apgar score												
Inter-obstetrician:												
A–B	0.66				**100**		[-0.82,0.96]				33	**67**
A-C	0.40		17	**83**			[-0.84,1.87]			33	**67**	
B-C	0.47		17	**83**			[-0.87,1.76]			33	**67**	
A-B-C	0.48						-					
Obstetrician-outcome:												
A-outcome	0.18	**50**	**50**				[-1.48,1.48]				**83**	17
B-outcome	0.11	**50**	**50**				[-1.62,1.62]			33	**83**	17
C-outcome	0.00	**100**					[-2.08,2.08]			**83**	17	
Random -outcome	0.01	**100**					[-1.42,10.0]	**83**	17			
Umbilical artery blood												
pH Inter-obstetrician:												
A–B	0.79				**50**	**50**	[-0.08;0.08]				33	**67**
A-C	0.70				**100**		[-0.09;0.11]			17	**50**	33
B-C	0.57			**100**			[-0.10;0.12]			17	**50**	33
A-B-C	0.69						-					
Obstetrician-outcome:												
A-outcome	0.35			**100**			[-0.15;0.20]		17	17	**50**	17
B-outcome	0.32			**100**			[-0.15;0.19]		17	17	**50**	17
C-outcome	0.14	17	**83**				[-0.16;0.21]		17	**33**	**33**	17
Random -outcome	-0.20	**67**	33				[-0.32;0.24]	**67**	17	17		

Abbreviations: UAB, umbilical artery blood; ICC, intraclass correlation coefficient; LA, limits of agreement; A, obstetrician A; B, obstetrician B; C, obstetrician C.

Table 6.3: Agreement between real outcomes and obstetrician and random predictions for 1- and 5-min Apgar scores and for UAB pH, assessed by the ICC and the LA with respective 95% CIs.

	ICC (95%CI)	**Mean difference/bias** (95%CI)	**LA**	**(Lower LA 95%CI)**	**(Lower LA 95%CI)**
Agreement between					
1-min Apgar outcome and:					
Obstetrician A prediction	0.17 (-0.07;0.38)	0.0 (-0.10;0.10)	[-2.06,2.06]	(-2.88;-1.24)	(1,24;2.88)
Obstetrician B prediction	0.29 (-0.06;0.49)	0.0 (-0.09;0.09)	[-2.01,2.01]	(-2.81;-1.21)	(1.21;2.81)
Obstetrician C prediction	0.06 (-0.14;0.27)	0.1 (-0.02;0.22)	[-2.42,2.62]	(-3.43;-1.41)	(1.61;3.63)
Random generated Apgar	0.04 (-0.05;0.17)	2.86 (2.58;3.14)	[-3.16,8.88]	(-5.57;-0.75)	(6.47;11.29)
5-min Apgar outcome and:					
Obstetrician A prediction	0.18 (-0.03;0.38)	0.0 (-0.07;0.07)	[-1.48,1.48]	(-2.07;-0.89)	(0.89;2.07)
Obstetrician B prediction	0.11 (-0.10;0.32)	0.0 (-0.08;0.08)	[-1.62,1.62]	(-2.27;-0.97)	(0.97;2.27)
Obstetrician C prediction	0.00 (-0.22;0.23)	0.0 (-0.10;0.10)	[-2.08,2.08]	(-2.91;-1.25)	(1.25;2.91)
Random generated Apgar	0.01 (-0.040.10)	4.29 (4.02;4.56)	[-1.42,10.0]	(-3.70;0.86)	(7.72;12.28)
UAB pH and:					
Obstetrician A prediction	0.35 (0.10:0.39)	0.025 (0.02;0.03)	[-0.15;0.20]	(-0.22;-0.08)	(0.13;0.27)
Obstetrician B prediction	0.32 (0.06;0.54)	0.02 (0.01;0.03)	[-0.15;0.19]	(-0.22;-0.08)	(0.12;0.26)
Obstetrician C prediction	0.14 (0.10;0.57)	0.025 (0.02;0.03)	[-0.16;0.21]	(-0.23;-0.09)	(0.14;0.28)
Random generated pH values	-0.20 (-0.47;0.07)	-0.04 (-0.05;-0.03)	[-0.32;0.24]	(-0.43;-0.21)	(0.13;0.35)

6.5 Discussion

Important inconsistencies in interpretation of ICC and LA results by both clinicians and biostatisticians were observed when these statistics were used to assess agreement in prediction of Apgar score and UAB pH. Interpretation of LA results was less consensual than of those of ICC. The greater consistency in interpretation of ICC results may have been related to the existence of cutoff values, established by some authors for this measure, such as the 0.75 value set by Burdock et al. to signify good agreement [8]. It is possible that these cutoff values were kept in mind by experts when interpreting results. The intrinsic subjectivity of interpretation of the LA depending on a clinical rather than a statistical interpretation may also have contributed to the lower consistency among experts in the interpretation of this measure [3-5]. LA results suggested a fair to good interobstetrician agreement, whereas interpretation of the ICC suggested a poor to fair agreement.

A preexisting database of 72 cases, with good-quality signals, monitored until very close to delivery was used to maximize obstetricians' predictive capacity by lessening possible confounding factors. It is unlikely that the fact that UAB pH values were missing in 21 cases had any influence in the analysis, as Apgar scores and UAB pH results were analyzed separately, and there were

no differences in gestational age, birth weight, and maternal age among the groups with and without UAB pH values.

Apgar scores displayed a skewed distribution, but the LA approach assumptions were met both for Apgar scores and UAB pH, as the differences approximated a normal distribution, and the mean and standard deviation of the differences were constant throughout the range of measurements [5]. Moreover, although assumptions of normality for ANOVA models may not be satisfied, the ANOVA procedure is generally robust and permits the estimation of the ICC [14].

To avoid bias in interpretation of the LA and ICC results, these were independently provided in a random order to six experienced clinicians and six experienced biostatisticians, without any further information. Three of these clinicians were the obstetricians involved in the prediction of the neonatal outcome variables, an aspect that may have introduced bias into the analysis. However, we feel that this is unlikely, as the predictions were performed more than 2 months before the interpretation of agreement results, and they had no information on their personal scores. Moreover, the interpretation of LA and ICC results was highly consistent among clinicians and statisticians.

LA results seem more plausible than the ICC results, as they suggest, although not reaching statistical significance, that obstetricians predict 5-minute Apgar scores better than 1-minute Apgar scores (Table 6.3). The latter are intuitively less predictable because of their greater variance and clinical instability during delivery. LA results are also clinically more plausible, as they suggest that experienced obstetricians predict outcome variables better than randomly generated values, which was not the case with ICC results.

The inconsistencies between ICC and LA can largely be explained by the intrinsic dependence of the ICC on variance [8]. Indeed, when Apgar values were randomly generated, variances were much higher (7.7 and 7.8 for 1- and 5-minute Apgar, respectively) than when observers predicted them (variances ranging from 0.3 to 0.7).

6.6 Conclusions

The interpretation of agreement studies on continuous variables using ICC and LA should be performed with caution, as considerable inconsistencies can be found between these two measures. A possible approach to overcome this problem is the use of both measures to assess agreement. It is the overall

impression from several approaches on different populations and setups that gives one the feeling that a measure is reliable and to what degree. As each method for agreement measurement has its own limitations, it is important to discuss method specific drawbacks when interpreting the results of individual studies. Agreement remains a difficult concept to represent mathematically, and further development of statistical methods needs to be considered.

6.7 Acknowledgments

Alberto Freitas, Armando Teixeira Pinto, Cláudia Camila Dias, José Cliver Albuquerque, Luís Antunes, Orquídea Ribeiro, Paula Pinto, and Pedro Rodrigues are acknowledged for their collaboration in the interpretation of the results of agreement statistics.

6.8 References

[1] de Vet H. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. 1st edition. Chichester, UK: Wiley; 1999. p. 3123-7.

[2] Ayres-de-Campos D, Bernardes J, Costa-Pereira A, Pereira-Leite L. Inconsistencies in expert's classification of cardiotocograms and subsequent clinical decision. Br J Obstet Gynaecol 1999;106:1307-10.

[3] Santos CC, Pereira CA, Bernardes J. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences.Br J Obstet Gynaecol 2005;112:667-9.

[4] Luiz RR, Szklo M. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. J Clin Epidemiol 2003;56:963-7.

[5] Bland JM, Altman DG. Statistical methods for assessing two methods of clinical measurement. Lancet 1986;1:307-10.

[6] Bartko JJ. The intraclass correlation coefficient as a measure of reliability.Psychol Rep 1966;19:3-11.

[7] Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465-76.

[8] Burdock E, Fleiss JL, Hardesty AS. A new view of interobserver agreement. Person Psychol 1963;16:373-84.

[9] Casey BM, Mclntire DD, Leveno KJ. The continuing value of the Apgar score for the assessment of newborn infants. N Engl J Med 2001;344:467-71.

[10] van de Riet JE, Vandenbussche FP, Le Cessie S, Keirse MJ. Newborn assessment and long-term adverse outcome: a systematic review. Am J Obstet Gynecol 1999;180:1024-9.

[11] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-7.

[12] Fleiss JU, Davies M. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378-82.

[13] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141-64.

[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

7. Complexity and categorical analysis may improve the interpretation of agreement studies using continuous variables.

7 Complexity and categorical analysis may improve the interpretation of agreement studies using continuous variables

Costa-Santos C, Bernardes J, Antunes L, Ayres-de-Campos D.

(submitted for publication)

This chapter discusses how the interpretation of agreement studies considering continuous variables can be improved by the analysis of complexity and by the categorization of continuous variables into categories with clinical significance. Another aspect discussed in this chapter is the importance of the introduction of complexity analysis into medical research and decision making.

7.1 Introduction

Disagreement over clinical decisions may have important research, clinical and medico-legal consequences (1). However, the ideal statistical measure of agreement has yet to be established, and use of more than one measure has been proposed (1-3). Even when this solution is adopted, inconsistent results can be found in assessment of complex continuous variables with the most widely used measures - the limits of agreement (LA) and the intra-class correlation coefficient (ICC) (4). When these were used to assess agreement in prediction of umbilical artery blood pH (UAB pH) and Apgar scores, based on the interpretation of fetal heart rate (FHR) tracings, LA results suggested a fair to good agreement, whereas ICC suggested it was poor to fair (4). LA results were judged to be more plausible with reality but their interpretation was less consensual, whereas the opposite occurred with the ICC.

In this commentary we propose the addition of complexity analysis and transformation into categorical variables in studies evaluating agreement over continuous variables, as a way to improve the interpretation of results. For this purpose a reappraisal of the previously cited study evaluating inter-observer agreement in prediction of neonatal UAB pH and Apgar scores was performed, using a higher sample size.

7.2 Complexity in clinical decisions

A complex system is a collection of individual agents which act in ways that are not totally predictable and whose actions are interconnected, so that the action of one part changes the context for other agents (5). A complex system behaves in a non-linear fashion, i.e. small changes in input may lead to large changes in outcome (6). Consequently, in complex medical systems, it is difficult to establish predictions or to develop simple cause-effect models (7). In such cases, clinical judgment is known to involve a high degree of uncertainty, frequently leading to a poor agreement among clinicians over the diagnosis and management, even when hard observable criteria are present. Uncritical adherence to guidelines in this setting may lead to more harm than good (5), and other approaches to deal with such uncertainty need to be used, such as non-linear models, intuition or scan of patterns (7).

7.3 Entropy, a measure of complexity or uncertainty

In information theory, entropy is a measure of the uncertainty, or complexity (8). A high value of entropy signifies that a large amount of information is needed to describe an outcome about which there is a great deal of uncertainty, i.e., complexity is high. On the other hand, a null value of entropy signifies that no information is required to describe the outcome, about which there is no uncertainty. As an example, we may consider the random picking of a ball from a box containing several coloured balls. If all the balls are of different colours, then there is maximal uncertainty about the colour of the chosen sample. On the other hand, if all the balls are of the same colour, then there is no uncertainty about the colour of the chosen ball.

The nature and meaning of entropy differs from that of variance and the coefficient of variation. It is a measure of uncertainty or complexity, whereas variance describes a linear measure of dispersion and the coefficient of variation describes a normalized measure of dispersion, calculated as the ratio of the standard deviation to the mean, so that it is a dimensionless number.

7.4 Adding complexity and categorical analysis to the LA and ICC in prediction of neonatal outcomes based on FHR monitoring

In clinical care, decisions based on interpretation of FHR signals are a good example of interventions with the characteristics of a complex system. Several studies have evidenced a low observer agreement over FHR analysis (9) and over clinical decisions based on its interpretation (10). It is therefore natural that agreement over the prediction of neonatal outcomes - UAB pH and Apgar score - based on FHR analysis is also low (4).

From a pre-existing database of term singleton pregnancies delivered vaginally, 152 intrapartum FHR tracings with at least 60 minutes, less than 15% signal loss and less than five minutes interval to delivery were consecutively selected. All tracings had been acquired with Sonicaid® (Oxford Instruments, Abingdon, UK) or Hewlett-Packard® (Hewlett-Packard, Palo Alto, California,

USA) fetal monitors, using an internal fetal scalp electrode. Apgar scores were evaluated by the caregivers responsible for immediate neonatal support. The umbilical cord was double clamped immediately after birth and samples was obtained to evaluate umbilical artery and vein blood pH. Neonatal outcome values were available in all cases.

Two experienced clinicians were asked to evaluate the last 60 minutes of each FHR tracing independently, using the International Federation of Ginecology and Obstetrics guidelines (11), and to predict newborn UAB pH and Apgar scores. Entropy, variance and coefficient of variation of the real UAB pH and Apgar score values were calculated and the interobserver agreement in prediction of these outcomes was assessed, using the LA, the ICC, and after categorisation of UAB pH values in three categories (≤7.10, 7.11-7.19, and ≥7.20) and Apgar scores in three categories (≤3, 4-6, and ≥7), using the the kappa statistics (K) and proportions of agreement (PA) (12). Results are displayed in Table 1.

Analysis of entropy and categorization, as described above, helps to shed some light into the results of observer agreement and the discrepancies between LA and ICC. The LA are greatly dependent on clinical, rather than statistical, interpretation and a difference of 0.21 (the LA range) in UAB pH is a clinically very meaningful difference. A UAB pH value of 7.00 is associated with a large risk to the newborn, while a value of 7.21 depicts a normal situation. So the LA results for UAB pH represent a clinically much more important disagreement than a difference of 2.99 (the LA range) for the 1-min Apgar. The clinical interpretation of LA seems therefore to be consistent with that of entropy analysis. Less entropy (or uncertainty) exists in prediction of Apgar scores, as more agreement between observers is found (Table 7.1). On the other hand, ICC results are in contradiction with clinical common sense, as they are extremely low for prediction of 10-min Apgar scores, due to the low variance of the variable. Lower ICC values are obtained when applied to a more homogeneous population, as compared with a more heterogeneous one, despite similar levels of agreement (13). For prediction of UAB pH, despite the low variance of this variable, the differences between predictions were extremely small, providing a relatively high ICC. However, those small differences are clinically very meaningful, as previously referred.

Assignment of continuous variables to the three previously described categories, and analysis with the PA, led to an interpretation of results that was consistent with the LA and entropy. On the other hand, K was misleading due to its strong dependence on prevalence (14). At 5 and 10 minutes almost all

newborns had an Apgar score higher or equal to 7, and there was an almost perfect interobserver agreement on prediction of these, however K was 0, suggesting that agreement was the same as expected by chance.

Table 7.1: Entropy, variance and coefficient of variation (CV) of real UAB pH and Apgar score values. Limits of agreement (LA) and intraclass correlation coefficient (ICC) on interobserver agreement in prediction of UAB pH and Apgar score values by two clinicians. Proportions of agreement (PA) and kappa statistic (K), on interobserver agreement in the same prediction of UAB pH divided into three categories (≤7.10, 7.11-7.19, and ≥7.20) and Apgar scores divided into three categories (≤3, 4-6, and ≥7,); (n=152).

	Real values			**Agreement between clinicians**			
	entropy	variance	CV	PA	kappa	ICC	LA
UAB pH	4.727	0.006	1%	0.80	0.55	0.70	[-0.12;0.09]
Apgar at:							
minute 1	1.558	0.935	11%	0.90	0.52	0.62	[-1.39;1.60]
minute 5	1.258	0.500	7%	0.97	0.00	0.59	[-1.18;1.30]
minute 10	0.748	0.195	4%	0.99	0.00	-0.02	[-0.70;0.73]

7.5 Conclusions

These results suggest that interpretation of agreement studies using continuous variables may be improved by analysis of complexity and by their transformation into categorical variables, with cutoff values chosen according to clinical significance. This should be taken into consideration in future studies of agreement involving complex continuous variables. The overall impression obtained from all these approaches and the awareness of each method's limitations, offers a more rigorous interpretation of the degree of agreement.

Another aspect that needs to be considered is the introduction of complexity analysis into medical research and decision making. While in simple problems with low uncertainty it is reasonable to use linear models and uncomplicated protocols to help guide clinical diagnosis and decisions (7), more complex situations pose a higher level of uncertainty for clinical judgement, which frequently has to rely on intuition. In these scenarios use of non-linear models (15) and mining of patterns (16) may be a helpful way to deal with higher levels of complexity.

7.6 References:

1 Costa-Santos C, Costa Pereira A, Bernardes J. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences. BJOG. 2005;112:667-9.

2 Khan KS, Chien PF. Evaluation of a clinical test. I: assessment of reliability. BJOG .2001;108: 562-7.

3 Luiz RR, Szklo M. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. J Clin Epidemiol. 2005;58: 215-6.

4 Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C. The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. J Clin Epidemiol. 2010; (in press)

5 Plsek P, Greenhalgh T. The challenge of complexity in health care. BMJ. 2001; 323:625-8.

6 Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. BMJ. 2008;336:1281-3.

7 Wilson T, Holt T, Greenhalgh T. Complexity Science: Complexity and Clinical Care. 2001; BMJ 23:685-8.

8 Shannon Î. A Mathematical Theory of Communication. Bell System Technical Journal. 1948;27:379-423, 623-656.

9 Bernardes J, Costa-Pereira A, Ayres-de-Campos D, van Geijn HP, Pereira-Leite L. Evaluation of interobserver agreement of cardiotocograms. Int J Gynecol Obstet 1997;57(1):33-37.

10 Ayres-de-Campos D, Bernardes J, Costa-Pereira A, Pereira-Leite L. Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision. Br J Obstet Gynaecol. 1999;106:1307-10.

11 FIGO subcommittee on Standards in Perinatal Medicine: Guidelines for the use of fetal monitoring. Int J Gynaecol Obstet 1987;25;159-67.

12 Grant JM. The fetal heart rate is normal, isn't it ? Observer agreement of categorical assessments. The Lancet 1991;337:215-218.

13 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46:423-9.

14 Müller R, Büttner P. A critical discussion of intraclass correlation coefficients.Stat Med. 1994;13:2465-76.

15 Gonçalves H, Rocha AP, Ayres-de-Campos D, Bernardes J. Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. Med Biol Eng Comput. 2006; 44:847-55.

16 Costa-Santos C, Bernardes J, Vitányi P, Antunes L. Clustering Fetal Heart Rate Tracings by Compression. Proceedings of 19th IEEE International Symposium of Computer-Based Medical Systems 2006; 685-90

8. Assessment of disagreement:
a new information based approach

8 Assessment of disagreement: a new information based approach

Costa-Santos C, Antunes L, Souto A, Bernardes J.

Ann Epidemiol 2010 (in press).

8.1 Abstract

Purpose: Disagreement on the interpretation of diagnostic tests and clinical decisions remains an important problem in Medicine. As no strategy to assess agreement seems to be fail-safe to compare the degree of agreement, or disagreement, we propose a new information based approach to assess disagreement.

Methods: The sum over all logarithms of possible outcomes of a variable is a valid measure of the amount of information contained in the variable. So the proposed measure is based on the amount of information contained in the differences between two observations. This measure is normalized and satisfies the flowing properties: it is a metric, scaled invariant and it has differential weighting.

Results: Two real examples and a simulation were used to illustrate the usefulness of the proposed measure to compare the degree of disagreement.

Conclusions: Used as complement to the existing methods, our approach can be useful to compare the degree of disagreement among different populations.

8.2 Introduction

Measurement is essential both for clinical care and for epidemiologic research; however measurement always implies some degree of variability[1]. Ideally the only source of variability in measurements should be the variability within subjects, however often observer variability and other sources of variability are also present.

In clinical care, diagnosis often depends on measurements, and disagreement in these measurements may have obvious implications for clinical practice and may also have medico-legal consequences[2]. Before the introduction, in clinical practice, of a new diagnostic method, it is essential to assess the agreement between the new method and the traditional one and it is also fundamental to know whether the new method can be reproduced by a second observer. In research, disagreement in measurements may lead to discrepant results in validity or randomized controlled studies with the same objectives and with the same methods, consequently misleading and heterogeneous results in meta-analysis will be found[2]. As it is impossible to control the various sources of variation of a measurement, agreement studies have a very important role. Despite the importance of agreement studies, misleading and sometimes inappropriate measures of agreement have been used in leading medical literature[2,3]. Considering the limitations of current strategies to assess agreement and as no strategy seems to be fail-safe to compare the degree of agreement among different populations, agreement studies should be interpreted with caution and possibly combined with the use of different strategies to assess agreement always with the limitations of these strategies in mind[2,4].

In this article, we propose a new approach to assess agreement based on information theory.

8.3 Methods

Most used strategies to assess agreement

The Intraclass Correlation Coefficient (ICC) is claimed to be suitable for observer agreement assessment5. The ICC is defined as the ratio of the variance between subjects, to the total variance[6,7]. These variances are derived from analyses of variance (ANOVA). Fleiss and Shrout present different kinds of ICC derive from different ANOVA models, and the ANOVA model depends on the study design[6]. One-way random effects model should be used when the subjects are deemed a random sample of subjects to be evaluated by the observers. The focus of interest is the difference of the individual observer's rating from the average rating of the observers for the ith subject[6].

In two-way models the observers are deemed an important factor in the ICC computation. In two-way random effects model, not only the subjects are deemed random, but the observers are deemed a random effect as well. In two-way mixed model each subject is evaluated by k observers, who are the only observers of interest, in this case, the observers are a fixed effect while the subject ratings are a random effect[6].

The ICC (from two-way models) that should be used for assessing agreement was defined by McGraw and Wong as the 'ICC for agreement'. We obtain the 'ICC for consistency' or the 'ICC for agreement' excluding or not the observer variance from the denominator mean square, respectively[8]. The systematic variability due to observers is irrelevant for 'ICC for consistency' but it is not for 'ICC for agreement'.

The ICC ranges from 0 (no agreement) to 1 (perfect agreement), however it can be negative, how such negative ICC values should be interpreted is quite unclear. The ICC assumptions, multivariate normal distributions and equality of variances, should be checked.

An important limitation of ICC is that it is strongly influenced by the variance of the trait in the population in which it is assessed[9]. This limitation can be illustrated with the following example: suppose that we aim to assess whether a depression scale can be reproduced by a second observer when applied to a random sample of the adult population (a heterogeneous population, with high variance) the scale's ICC may be higher than when the scale is applied to a very homogeneous population (with low variance), such as patients hospitalized for acute depression. This ICC characteristic has also been considered by some authors as an advantage, for it would make the discordance

relative to the magnitude of measurements[10], however comparability across populations is not possible with ICC. Consequently, ICC values have no absolute meaning, and the cut-off value of 0.75 proposed by Burdock[7] and Lee[11], and often reproduced to signify a good agreement, has limited justification.

Lin's Concordance Correlation Coefficient (CCC) is the Pearson Correlation of Coefficient, which assesses the closeness of the data to the line of best fit, modified by taking into account how far the line of best fit is from the 45º line through the origin[12]. Lin objected to the use of the ICC as a way to assess agreement between methods of measurement and developed the concordance correlation coefficient (CCC). However, there are similarities between certain specifications of the ICC and the CCC[13]. Some limitations of ICC, like the limitation of comparability of population described above, are also present in CCC[14].

Bland and Altman propose the limits of agreement to assess agreement between methods of measurement[15]. Limits of agreement can be calculated based on the mean difference between the measurements of two methods in the same subjects and the standard deviation of these differences. Approximately 95% of these differences will lay between the mean differences $\pm$ 1.96 standard deviation of these differences. The limits of agreement approach depends on some assumptions about the data: the mean and standard deviation of the differences are constant throughout the range of measurement and these differences are approximately Normally distributed[15]. Limits of agreement are expressed in terms of the scale of measurement and the decision whether a degree of agreement is acceptable or not is always a clinical, not statistical, judgment.

Other approaches[16,17], have been proposed for assessing agreement; however all of them are also limited when the aim is to compare the agreement in different populations with different trait characteristics.

The New Approach: Information-Based Measure of Disagreement

Entropy, introduced by Shannon, can be described as the average amount of information contained in a variable[18]. A high value of entropy means that a large amount of information is needed to describe an outcome variable about which we have high uncertainty. The sum over all logarithms of possible

outcomes of the variable is a valid measure of the amount of information, or uncertainty, contained in a variable. Consider that we aim to measure disagreement between measurements obtained by observer A (variable A) and observer B (variable B). Also, consider for variable A, a vector A that can take the range of non-negative values ($a_1,\ldots,a_n$) and for variable B, a vector B that can take the range of non-negative values ($b_1,\ldots,b_n$). The intuition behind our definition is that the disagreement between A and B is related to the differences between them, with the minimum reached when A and B are identical. So, it is then natural to consider:

$$\sum_{i=1}^{n} \log_2 |a_i - b_i|$$

the amount of information contained in the differences between observers A and B. By adding 1 to the differences, we avoid the behavior of the logarithmic function between 0 and 1, while keeping a quite natural close relation to the notion of entropy where Shannon considered the log of the inverse of a probability, i.e., the log of a value always greater or equal to 1. Now, in order to get a value between 0 and 1 we normalize the amount of information contained in the differences to obtain the following measure of disagreement.

Considering two vectors A = ($a_1,\ldots,a_n$) and B = ($b_1,\ldots,b_n$) with non-negative components.

The information-based measure of disagreement is defined as:

$$d(A,B) = \frac{1}{n}\sum_{i=1}^{n} \log_2\left(\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1\right)$$

with the convention $\frac{|0-0|}{\max\{0,0\}} = 0$.

This coefficient equals 0 when the observers agree, or when there is no disagreement, i.e., when $a_i = b_i$. In this case we say that there is no information in the differences between observers A and B. As observer A and B's measurements disagree, the amount of information in the differences between observers A and B increases and the information-based measure of disagreement increases, tending to 1, or, in other words, the distance between the observers increases.

The convention $\frac{|0-0|}{\max\{0,0\}} = 0$ has a natural interpretation since if both observers rate 0, they agree, so the contribution of this observation to the final sum of disagreement is $\log_2(0+1) = 0$.

Properties of Information-Based Measure of Disagreement

The information-based measure of disagreement, d(A,B), is a metric, i.e., follows the properties:

- is non-negative, d(A,B) is always equal to or greater than zero;
- is zero if and only if the observers rate exactly the same measurements, d(A,B)=0 if and only if A=B;
- is symmetric, d(A,B)=d(B,A) and
- the triangular inequality is verified, d(A,B)≤d(A,C)+d(C,B).

The information-based measure of disagreement is scale invariant, d(A,B)=d(kA,kB) with k a non-zero constant. It also has differential weighting, i.e., a difference found between high values contribute less to the information-based measure of disagreement than the same difference found between low values. The appendix contains the proof of each property.

Inference of the information-based measure of disagreement

In the absence of any distributional assumption, the most obvious inference method is the non parametric bootstrap. Each of the M bootstrap samples is taken from $a_{i\ (i=1,\ldots,n)}$, and $b_{i\ (i=1,\ldots,n)}$, where a_i and b_i are the measurements rated by the observers A and B respectively, on the subject i, and provides an estimate of the proposed information-based measure of disagreement using $d(A,B)=\frac{1}{n}\sum_{i=1}^{n}\log_2\left(\frac{|a_i-b_i|}{\max\{a_i,b_i\}}+1\right)$. A confidence interval for the proposed measure of disagreement can be obtained from the percentiles of the empirical distribution of the M estimates.

8.4 Results

Example - systolic blood pressure

Bland and Altman present the example of measurements of systolic blood pressure of 85 individuals, by two observers (observer J and observer R) with sphygmomanometer, and one other measurement, by a semiautomatic device (device S)[19]. Luiz et al re-analyze the data and also observe, with a graphical approach, a greater agreement between the two observers than between the observers and the semiautomatic device[16]. Using our information-based measure of disagreement; we also obtained a significantly more disagreement between each observer and the semiautomatic device than between the two observers (Table 8.1).

Table 8.1. Disagreement between observers J and R, using a sphygmomanometer, and between each observer and the semiautomatic device S, in measurement of the systolic blood pressure of 85 individuals.

Disagreement between:	Information Based Measure of Disagreement	95% confidence intervals
Observer J and observer R	0.019	[0.013;0.029]
Observer J and S	0.158	[0.134;0.182]
Observer R and S	0.160	[0.136;0.186]

Example - Apgar score at first minute

The Apgar score is widely used in developed countries for assessing the clinical status of the newborn, determining whether or not there is a need for immediate medical care. It is usually assessed at the first and fifth minutes after birth and ranges between zero and ten. It is based on the analysis of the newborn's skin color, response to stimulus, breathing, muscle tonus and heart rate.

Cardiotocography is the most common method used for fetal monitoring during labor. It registers the heart rate of the fetus and contractions of the uterus. The value of fetal monitoring procedure is in its ability to predict newborn outcome, as the prediction of newborn outcome during labor can decrease perinatal mortality and morbidity.

Three obstetricians were asked to evaluate 72 intrapartum cardiotocographic tracings independently and through them predict the Apgar score at the first minute. After birth, the true values of the Apgar scores were evaluated by the caregivers responsible for immediate neonatal support. To simulate an observer who estimated Apgar by chance, we randomly generated Apgar score values between zero and ten.

Table 8.2 presents the true Apgar scores evaluated after birth for 72 newborns and the Apgar scores predicted by the three obstetricians based on the intrapartum cardiotocographs tracings. The obstetricians were blinded for the neonatal outcome and evaluate tracings independently. Table 2 also presents a random estimation of the Apgar scores created by a generation of random values between zero and ten.

Table 8.3 presents the agreement, assessed by the ICC, and the disagreement, assessed with the information-based measure of disagreement, between each obstetrician´s predictions and Apgar evaluated by the caregiver for neonatal support and between randomly generated Apgar and evaluated by the caregiver for neonatal support.

As ICC is a measure of agreement, higher values correspond to better agreement. However, as the proposed information-based measure of disagreement is a measure of disagreement, higher values correspond to a higher disagreement, i.e., a worse agreement. ICC results suggest that one obstetrician predicted Apgar scores no better than the random generation of values. This is an artifact caused by the dependence of the ICC on variance. When Apgar scores were randomly generated, variance is much higher (7.7) than when obstetricians predict them (variances ranging from 0.4 and 0.7).

On the other hand, the information-based measure of disagreement demonstrates evidence that the Apgar score predictions by the obstetricians were significantly better, i.e. there was less disagreement with Apgar scores evaluated by the caregivers responsible for neonatal support, than by the random generation of Apgar scores (Table 8.3).

Table 8.2. Apgar score at first minute evaluated by the caregivers responsible for immediate neonatal support (Apgar) and Apgar score predicted by three obstetricians (A, B and C) and randomly estimated (random) for 72 newborns.

		Apgar predicted by						Apgar predicted by			
Newborn	**Apgar**	A	B	C	random	**Newborn**	**Apgar**	A	B	C	random
1	9	9	9	9	1	**37**	9	9	9	9	1
2	9	9	9	9	1	**38**	10	9	9	9	2
3	8	9	9	9	8	**39**	8	9	9	9	2
4	9	9	9	9	8	**40**	9	9	8	8	9
5	9	9	9	9	5	**41**	9	6	9	9	7
6	9	9	9	9	2	**42**	9	7	8	8	6
7	9	9	9	9	7	**43**	9	7	9	8	6
8	9	8	9	9	4	**44**	7	7	9	8	1
9	8	9	9	9	4	**45**	7	8	8	8	2
10	9	9	9	8	5	**46**	8	8	9	9	9
11	9	8	9	9	3	**47**	8	8	8	8	8
12	4	7	6	7	1	**48**	10	8	8	8	8
13	9	7	7	7	6	**49**	9	8	9	9	6
14	8	8	9	9	7	**50**	9	8	9	9	4
15	9	9	9	8	1	**51**	9	7	8	8	9
16	9	7	9	8	7	**52**	9	9	9	9	3
17	9	9	9	9	2	**53**	9	7	8	8	8
18	9	8	8	8	4	**54**	9	8	9	9	3
19	9	8	8	8	9	**55**	9	8	9	8	1
20	8	7	7	7	7	**56**	9	8	9	9	2
21	8	8	8	8	4	**57**	8	8	8	8	4
22	7	7	6	8	6	**58**	8	9	9	9	3
23	9	9	9	9	6	**59**	9	7	8	8	3
24	10	7	7	8	3	**60**	9	7	8	8	9
25	10	8	8	9	4	**61**	9	9	9	9	5
26	9	7	8	8	5	**62**	8	7	7	7	4
27	9	8	8	9	1	**63**	9	8	9	8	2
28	9	7	8	9	5	**64**	9	9	9	9	9
29	9	8	9	9	1	**65**	9	9	9	9	1
30	8	9	9	9	1	**66**	10	8	9	8	4
31	8	8	9	9	1	**67**	8	9	9	9	9
32	9	9	9	9	9	**68**	8	8	9	8	4
33	9	9	9	9	6	**69**	10	9	9	9	8
34	5	9	9	9	1	**70**	8	9	9	9	7
35	8	9	9	9	1	**71**	10	8	9	8	7
36	9	9	9	9	1	**72**	9	8	9	8	4

Table 8.3. Agreement assessed with the ICC and respective 95% confidence intervals (in brackets), and disagreement, assessed with the information-based measure of disagreement and respective 95% confidence intervals (in brackets), between each obstetrician´s predictions and Apgar scores evaluated by the caregiver for neonatal support and between randomly generated Apgar predictions and Apgar scores evaluated by the caregiver for neonatal support.

	Agreement (ICC)	**Information-based Measure of Disagreement**
Obstetrician A and Apgar	0.06 [-0.14;0.27]	0.13 [0.11;0.17]
Obstetrician B and Apgar	0.29 [0.06;0.49]	0.10 [0.07;0.13]
Obstetrician C and Apgar	0.17 [-0.07;0.38]	0.11 [0.08;0.13]
Random and Apgar	0.04 [-0.05;0.17]	0.54 [0.47;0.61]

Note that ICC is a measure of agreement (higher values correspond to better agreement) while the proposed information-based measure of disagreement is a measure of disagreement (higher values correspond to worse agreement).

Simulation study

To assess the performance of the proposed measure we used simulations in different scenarios. From Uniform distributions between 0 and 100 we generated the observer A's ratings for sample sizes of 10, 25, 50, 100, 500 and 5000. In a first scenario, Observer B's ratings were created by adding a systematic difference to observer A's ratings: Observer B's ratings were equal to observer A's ratings plus 10 units. In a second scenario, Observer B's ratings were created by adding a random difference to Observer A's ratings: Observer B's ratings were equal to observer A's ratings plus a quantity generated by a Normal distribution with mean 0 and standard deviation 5. In a third scenario, Observer B's ratings were created by adding both random and systematic differences to Observer A's ratings: Observer B's ratings were equal to observer A's ratings plus a quantity generated by a Normal distribution with mean 10 and standard deviation 5. We assumed as 0 all generated negative Observer B's ratings.

A thousand bootstrap samples were taken from data for each scenario and each sample size to provide the estimates of the proposed information-based measure of disagreement. Table 8.4 presents the mean and standard deviation of those 1000 estimates (standard error). Table 8.4 also presents the bias, i.e., the difference between the original and the mean of 1000 estimates of the information-based measure of disagreement. The point estimates of the

proposed information-based measure of disagreement had very small bias and standard error even for small sample sizes.

Table 8.4. A thousand bootstrap samples were taken from simulated data for each of three scenarios (systematic, random or systematic and random difference between observers) and each sample size (10, 25, 50, 100, 500, 5000) to provide estimates of the information based measure of disagreement. The table presents the mean, bias and standard error (se) of those estimates.

Inter observer difference	**Sample size**					
	10	25	50	100	500	5000
Systematic						
mean	0.286742	0.260622	0.267393	0.321260	0.283476	0.303583
bias	-0.000240	-0.000355	-0.000099	-0.000153	0.000059	0.000013
se	0.063482	0.033186	0.024173	0.020150	0.007960	0.002773
Random						
mean	0.145643	0.166753	0.179945	0.151991	0.147059	0.172590
bias	0.000402	-0.000170	-0.000355	-0.000342	-0.000173	0.000050
se	0.091367	0.041682	0.031509	0.017342	0.007857	0.002955
Both						
mean	0.285104	0.252425	0.268723	0.288005	0.295004	0.287160
bias	-0.000511	0.000139	0.000002	0.000376	-0.000029	-0.000039
se	0.073055	0.041336	0.030050	0.022234	0.009782	0.002937

Observer A's ratings are generated from an Uniform (min=0,max=100).
Systematic: Observer B's rating are equal to Observer A's ratings plus 10 units.
Random: Observer B's ratings are equal to observer A's ratings plus a quantity generated by a Normal (mean=0,standard deviation=5).
Both: Observer B's ratings are equal to observer A's ratings plus a quantity generated by a Normal (mean=10, standard deviation=5).

8.5 Discussion

We can look at disagreement between observers as the distance between their ratings, so the metric properties are important. Moreover, the proposed measure of disagreement is scale-invariant, i.e., the degree of disagreement between two observers should be the same if the measurements are analyzed in kilograms or in grams, for example.

Differential weighting is another property of the proposed information-based measure of disagreement: each comparison between two ratings is divided by a normalizing factor, depending upon each pair of ratings alone, before summing. Therefore, the information-based measure of disagreement is appropriate for ratio scale measurements (with a natural zero) and it is not

appropriate for interval scale measurements (without a natural zero). For example, outside air temperature in Celsius (or Fahrenheit) scale does not have a natural zero. The 0º is arbitrary and it does not make sense to say that 20º is twice as hot as 10º. Outside air temperature in Celsius (or Fahrenheit) scale is an interval scale. On the other hand, height has a natural 0 meaning: the absence of height. Therefore, it makes sense to say that 80 inches is twice as large as 40 inches. Height is a ratio scale. Suppose the heights of a sample of subjects measured independently by two different observers. A difference between the two observers of one inch in a child subject represents a worse observers' error than a disagreement between observers of one inch in an adult subject. Due to differential weighting property of the information-based measure of disagreement, a difference between the observers of one inch in a child in fact weights less to the estimate of information-based measure of disagreement between observers than a difference between the observers of one inch in an adult.

The usual approaches used to evaluate agreement have the limitation of the comparability of populations. In fact, ICC depends upon the variance of the trait in the population; although this characteristic can be considered an advantage it does not permit one to compare the degree of agreement across different populations. Also the interpretation of the Limits of Agreement depends on what can be considered clinically relevant or not, which could be subjective and different from reader to reader. The comparison of the degree of agreement in different populations is not straightforward. Other approaches16, 17 to assess observer agreement have been proposed, however the comparability of populations is still not easy with these approaches.

The proposed information-based measure of disagreement, used as a complement to current approaches for evaluating agreement, can be useful to compare the degree of disagreement among different populations with different characteristics, namely with different variances.

Moreover, we believe that information theory can make an important contribution to the relevant problem of measuring agreement in medical research, providing not only better quantification but also better understanding of the complexity of the underlying problems related to the measurement of disagreement.

8.6 References

1 Bland M. An Introduction to Medical Statistics. 3th ed. Oxford:Oxford University Press. 2000.

2 Costa Santos C, Costa Pereira A, Bernardes J. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences. Br J Obstet Gynaecol 2005; 112 (5): 667-9.

3 Gow RM, Barrowman NJ, Lai L, Moher D. A review of five cardiology journals found that observer variability of measured variables was infrequently reported. J Clin Epidemiol. 2008;61:394-401.

4 Luiz RR, Szklo M. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. J Clin Epidemiol. 2005; 58:215-6.

5 de Vet H. Observer reliability and agreement. In: P. Armitage and T. Colton, Editors, Encyclopedia of biostatistics. 1st ed. Wiley, Chichester. 1999, pp. 3123–3127.

6 Shrout PE, Fleiss JL. Intraclass Correlation: uses in assessing rater reliability. Psycological Bulletin.1979; 88: 420-8.

7 Burdock EI, Fleiss JL and Hardesty AS. A new view of interobserver agreement. Psycological Bulletin. 1963;16:373-84.

8 McGraw KO, Wong, SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):30-46.

9 Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465-2476.

10 Steiner DL, Norman GR. Health Measurements Scales. A practical guide to their development and use. 3th ed. Oxford: Oxford University Press. 2003.

11 Lee D, Koh D, Ong CN. Statistical evaluation of the agreement between two methods for measuring a quantitative variable. Comput Biol Med 1989;19:61-70.

12 Lin LK. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45:255–68.

13 Nickerson CAE. A Note on "A Concordance Correlation Coefficient to Evaluate Reproducibility". Biometrics. 1997;53:1503–7.

14 Atkinson G, Nevill A. Comment on the Use of Concordance Correlation to Assess the Agreement between Two Variables. Biometrics. 1997;53:775-7.

15 Bland JM, Altman DG. Statistical methods for assessing two methods of clinical measurement. Lancet 1986;1:307-310.

16 Luiz RR, Costa AJ, Kale PL, Werneck GL. Assessment of agreement of a quantitative variable: a new graphical approach. J Clin Epidemiol. 2003;56:963-7.

17 Monti KL. Folded empirical distribution function curves-mountain plots. Am Stat. 1995;33:525-7.

18 Shannon ˆI. A Mathematical Theory of Communication. Bell System Technical Journal. 1948;27:379-423 and 623-656.

19 Bland JM, Altaman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8:135-60.

8.7 APPENDIX

Properties of the proposed disagreement measure

The proposed disagreement measure, d(A,B), has the flowing properties:

Property 1 (Non-negativity). *Let A and B be two non-negative, random variables of size n, d(A, B) ≥ 0.*

Proof. If the two observations are zero, then by convention its contribution is zero. On the other observations, the property follows from the fact that the factors in the argument of log are at least 1. □

Property 2 (Identity of indiscernibles). *Let A and B be two non-negative, random variables of size n. d(A, B) = 0 if and only if A = B.*

Proof. d(A, B) = 0 if and only if $\log_2\left(\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1\right) = 0$ for all $1 \le i \le n$. So, $a_i = b_i = 0$ or $\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1 = 1$ for all $1 \le i \le n$, i.e., if and only if A=B. □

Property 3 (Symmetry). *Let A and B be two non-negative, random variables of size n, d(A, B) = d(B, A).*

Proof. d(A, B) = d(B, A) because of maximum and the absolute value symmetry. □

Property 4 (Triangular inequality). *Let A, B and C be three non-negative, random variables of size n, d(A,B) ≤ d(A,C)+ d(C,B)*

Proof. In order to prove d(A,B)≤d(A,C)+d(C,B) we have to prove that for all 1 ≤ i ≤ n:

$$\log_2\left(\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1\right) \le \log_2\left(\frac{|a_i - c_i|}{\max\{a_i, c_i\}} + 1\right) + \log_2\left(\frac{|c_i - b_i|}{\max\{c_i, b_i\}} + 1\right) \qquad (I)$$

First we deal with the case where at least two of the elements a_i, b_i and c_i are 0.

If $a_i=b_i=c_i=0$ then the inequality (I) is trivially true with the convention made.

If $a_i=b_i=0$ and $c_i\neq 0$ then $\log_2\left(\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1\right) = 0$, $\log_2\left(\frac{|a_i - c_i|}{\max\{a_i, c_i\}} + 1\right) \ge 0$ and $\log_2\left(\frac{|c_i - b_i|}{\max\{c_i, b_i\}} + 1\right) \ge 0$, and then, again, the inequality (I) is true.

If $a_i=c_i=0$ and $b_i\neq 0$. In this case $\log_2\left(\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1\right) = \log_2\left(\frac{|b_i|}{b_i} + 1\right)$, $\log_2\left(\frac{|c_i - b_i|}{\max\{c_i, b_i\}} + 1\right) = \log_2\left(\frac{|b_i|}{b_i} + 1\right)$ and $\log_2\left(\frac{|a_i - c_i|}{\max\{a_i, c_i\}} + 1\right) = 0$ which verifies the inequality (I).

Now assume that at least two of the values a_i, b_i and c_i are positive. Proving the inequality (I) is equivalent, to prove:

$$\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1 \le \left(\frac{|a_i - c_i|}{\max\{a_i, c_i\}} + 1\right)\left(\frac{|c_i - b_i|}{\max\{c_i, b_i\}} + 1\right) \Leftrightarrow$$

$$\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1 \le \frac{|a_i - c_i|}{\max\{a_i, c_i\}} + \frac{|c_i - b_i|}{\max\{c_i, b_i\}} + \frac{|a_i - c_i||c_i - b_i|}{\max\{a_i, c_i\}\max\{c_i, b_i\}} + 1$$

So, is sufficient to prove that:

$$\frac{|a_i - b_i|}{\max\{a_i, b_i\}} \leq \frac{|a_i - c_i|}{\max\{a_i, c_i\}} + \frac{|c_i - b_i|}{\max\{c_i, b_i\}}$$

Suppose that $\max\{a_i, b_i\} \geq c_i$.

Then $\max\{a_i, b_i\} \geq \max\{b_i, c_i\}$ and $\max\{a_i, b_i\} \geq \max\{a_i, c_i\}$.

So, $\frac{1}{\max\{a_i, b_i\}} \leq \frac{1}{\max\{c_i, b_i\}}$ and $\frac{1}{\max\{a_i, b_i\}} \leq \frac{1}{\max\{c_i, a_i\}}$.

Thus $\frac{|a_i - b_i|}{\max\{a_i, b_i\}} \leq \frac{|a_i - c_i| + |c_i - b_i|}{\max\{a_i, b_i\}} \leq \frac{|a_i - c_i|}{\max\{a_i, c_i\}} + \frac{|b_i - c_i|}{\max\{c_i, b_i\}}$

Now, suppose that $\max\{a_i, b_i\} < c_i$ then $a_i, b_i \leq c_i$.

Without loss of generality we can suppose that $b < a < c$. So, prove that:

$$\frac{|a_i - b_i|}{\max\{a_i, b_i\}} \leq \frac{|a_i - c_i|}{\max\{a_i, c_i\}} + \frac{|c_i - b_i|}{\max\{c_i, b_i\}}$$

is the same to prove:

$$\frac{c_i - a_i}{c_i} + \frac{c_i - b_i}{c_i} - \frac{a_i - b_i}{a_i} \geq 0$$

and

$$\frac{c_i - a_i}{c_i} + \frac{c_i - b_i}{c_i} - \frac{a_i - b_i}{a_i} = 2 - \frac{a_i}{c_i} - \frac{b_i}{c_i} - \left(1 - \frac{b_i}{a_i}\right) = 1 - \frac{a_i}{c_i} - \frac{b_i}{c_i} + \frac{b_i}{a_i}$$

as $a < c$ then $\frac{b}{c} < \frac{b}{a}$ and $\frac{a}{c} > 0$.

Thus $1 - \frac{a_i}{c_i} - \frac{b_i}{c_i} + \frac{b_i}{a_i} > 0$ □

Property 5. *Let A and B be two non-negative, random variable of size n,* $d(A,B) \leq 1$

.

Proof. The property follows the fact that $\frac{|a_i - b_i|}{\max\{a_i, b_i\}} \leq 1$, so $\log_2\left(\frac{|a_i - b_i|}{\max\{a_i, b_i\}} + 1\right) \leq 1$ □

Property 6. (Scale invariance). *Let A and B be two non-negative, random variable of size n and k a positive, non zero constant.* $d(A,B) = d(kA,kB)$.

Proof.

$$d(kA,kB) = . \frac{\sum_{i=1}^{n} \log_2\left(\frac{|ka_i - kb_i|}{\max(ka_i, kb_i)} + 1\right)}{n}$$

$$= \frac{\sum_{i=1}^{n} \log_2\left(\frac{k|a_i - b_i|}{k\max(a_i, b_i)} + 1\right)}{n}$$

$$= d(A,B) \square$$

9. General discussion and conclusions

9 General discussion and conclusions

This closing chapter explains why the example of foetal heart rate assessment in obstetrics and gynaecology was chosen through this thesis. All the strength and weakness of each chapter were discussed. At last this chapter presents the general conclusion of this thesis.

9.1 General discussion

Agreement among measurements taken by clinicians or diagnostic devices is fundamental to proceed with validity studies and randomised controlled trials (Costa-Santos 2005) and to ensure efficient delivery of health care (Shoukri 2005). However, unfortunately, incorrect analysis of agreement studies in medical literature is not rare (Bland 2003; Gow 2008). Many researchers continue to use inappropriate and misleading statistical analysis imitatively and without taking in account the limitations of statistical methods involved. Furthermore, the statistical methodology to agreement studies has been neglected and only a few books have dedicated chapters dealing with agreement studies in a comprehensive manner and there are several controversies in literature about which are the more appropriate statistical methods to assess agreement (Luiz 2005).

In spite of the importance of agreement studies and the limitations of current statistical methods for agreement assessment are universal to all medical areas, almost all the examples of this thesis are on obstetrics and gynaecology, in particular on prediction of newborn outcomes (umbilical artery blood pH and Apgar scores) based in the foetal heart rate analysis during the labour. The umbilical artery blood pH is the most objective, currently available measure of fetal oxygenation. A low UAB pH establishes the presence of respiratory and/or metabolic acidemia occurring during labour and delivery, representing a higher risk of perinatal death or neurological injuries from hypoxia. The Apgar score is widely used in developed countries for assessing the clinical status of the newborn determining whether or not there is a need for immediate medical care. It ranges between 0 and 10 and is based on the analysis of the newborn's skin color, response to stimulus, breathing, muscle tonus, and heart rate. In fact, the focus on a single medical area, obstetrics and gynaecology, facilitates the systematic review of which statistical measures that have been used and the exemplification of the limitations of current statistical methods to assess agreement. It is important to focus in a single medical area also for searching examples of studies published in literature with misleading results. In fact, only with some clinical knowledge we can fully understand a misinterpreted result of an agreement study due to misuse of statistical methodology. To get some clinical skills it is better to focus in a specific medical area. Moreover, obstetrics and in particularly foetal heart rate monitoring seems to be a good example to

illustrate all the chapters of this thesis. Actually, to illustrate the behaviour of the agreement measures we need a procedure with low agreement. As several studies have evidenced a low observer agreement over foetal heart rate analysis and over clinical decisions based on its interpretation (Ayres-de-Campos 1999; Bernardes 1997) we used, in chapters 5, 6, 7 and 8 of this thesis, examples with foetal heart rate analysis. Moreover, some statistical methods to assess agreement in continuous variables, as ICC, have a worst behaviour in measurements with low variance (de Vet 2005; Stratford 1989). Consequently, the prediction of umbilical artery blood pH, a variable with low variance, based in the foetal heart rate analysis, a procedure with low observer agreement, seems to be an ideal example to illustrate the chapters of this thesis considering continuous variables. Furthermore, both Apgar scores and umbilical artery blood pH were frequently categorized in three categories with clinical meaning (normal, suspicious and pathological) with a very low prevalent category (pathologic). Consequently, as some commonly used statistical methods to assess agreement in categorical variables, like kappa, have a worst behaviour when there is a low prevalent category, the prediction of umbilical artery blood pH and Apgar score based in the foetal heart rate analysis is also a good example to illustrate the chapters of this thesis considering categorical variables.

In order to achieve the **first objective** of this thesis: "to assess how agreement has been measured in obstetrics and gynaecology literature" a literature review was performed in three world-leading obstetrics and gynaecology journals and was presented in chapter 4 (Costa-Santos 2005). Although there are some controversies in literature about which are the best statistical methods for agreement assessments there are however some methods clearly inadequate. This is the case of Pearson correlation coefficient; there is general consensus in literature that Pearson correlation coefficient is inadequate to measure agreement. It is, therefore, surprising that this coefficient continue to be used to assess agreement, as our study, presented in chapter 4, revelled. In other medical area, another review performed latter, in five world-leading cardiology journals found similar results, correlation coefficient was the single most reported index for assessing observer agreement (Gow 2008). The major cause for concern is that both studies were performed in the world-leading journals in each area, so, the large proportion of agreement studies using inappropriate method to assess agreement can be underestimated. Our review also revealed that the presentation of one single statistical measure was the most common way used to assess agreement (Costa-Santos 2005) and due to the limitations of the current statistical methods, the assessment of agreement

based in one single measure can be misleading. One limitation of study presented in chapter 4 is that it was focused only in statistical methods and sometimes the problems in agreement studies start in their design. Another limitation is that our study did not assess how the results of agreement studies are interpreted, in fact, despite the use of the most recommended measures of agreement, their interpretation without taking in account their limitation can lead to erroneous conclusions.

Chapter 4 also describe some consequences of lack of agreement in medicine trying to accomplish the **second objective** of this thesis: "to describe the consequences of lack of agreement in medicine". In fact, our article also illustrates how poor reproducibility can dramatically influence, not only obviously the clinical practice, but also the results of clinical research (Costa-Santos 2005). In fact, in validity studies, a commonly observed observer variation in a diagnostic test can led to variations in the test sensitivity as wide as 0 to 100%. Considering randomised controlled trials, poor reproducibility in a diagnostic test A can lead to a variation in the relative risk in three trials comparing test A plus an 100% effective intervention with test B plus a 100% effective intervention as wide as 0.1 to 2. All theses discrepant results due to observer variation may be the cause of high heterogeneity in meta-analyses (Costa-Santos 2005). In fact, as claimed by Fleiss, even the most correct design of a validity study, a randomized controlled trial or other clinical study will not overcome the damage caused by unreliable or imprecise measurement (Fleiss 1986).

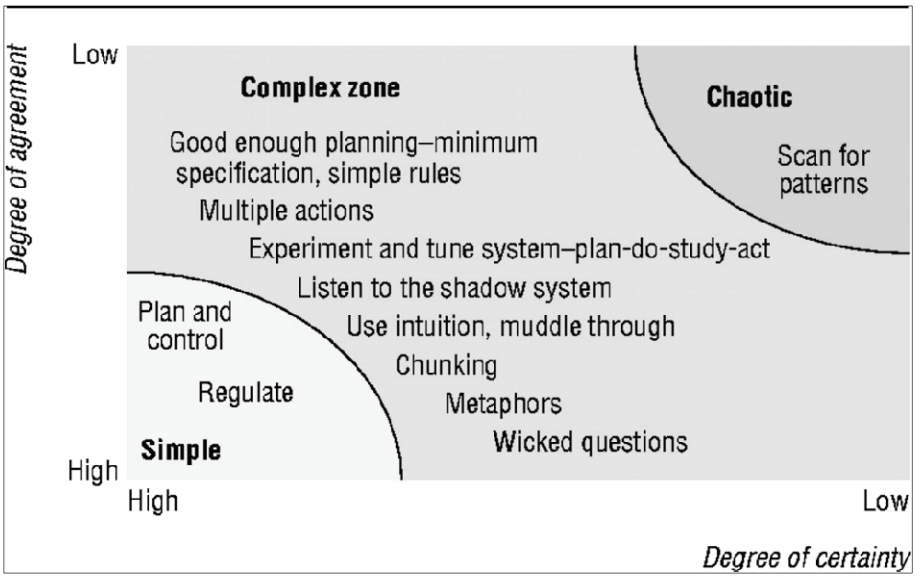
To answer to the **third objective**: "to describe the consequences of misuses and misinterpretation of the current statistical strategies to assess agreement" chapter 5, based in 3 letters to the editor, shows with real examples how the use of inappropriate methods for agreement assessment considering categorical variables, the inappropriate design of the agreement study and other common problems can lead to inconsistent results (Costa-Santos 2008-a; Costa-Santos 2008-b; Costa-Santos 2009). This chapter, based only in 3 letters to the Editor, is not exhaustive; we believe that there are in medical literature much more different examples of consequences of misuses and misinterpretation of the current statistical strategies to assess agreement. However, this kind of letter to the Editor is important, because it is essential that all the readers of medical literature be aware of the limitations of statistical methods to assess agreement and the consequences of misuses and misinterpretation of the statistical strategies currently used in agreement studies. Also, to answer to the third objective, chapter 6 describes how the most recommended strategies to assess

agreement considering continuous variables can be interpreted providing inconsistent and misleading conclusions. In this chapter the inconsistencies between ICC and limits of agreement are demonstrated with clinical neonatal outcomes (Costa-Santos 2010-a). The inconsistencies observed in this study may largely be explained by the intrinsic dependence of the ICC on variance, and the subjectivity that occurs in interpretation of the limits of agreement. These limitations have been identified before, but their consequences on the interpretation of results had not been demonstrated in this fashion. We believe it is very useful to explain methodological and statistical issues with clinical data and the chapter 6 contributes to the discussion of the application of agreement results in clinical settings.

Until better ways of assessing agreement are developed, in Chapter 4 we recommend, for assessment of observer agreement considering categorical variables, the use of both kappa and proportions of agreement for each category and overall, and for continuous variables, the use of both ICC and limits of agreement (Costa-Santos 2005). Simultaneously, Luiz et al. also recommend the use of both ICC and limits of agreement for agreement assessment considering continuous variables (Luiz 2005). However this in not an ideal solution, in fact, as we show in chapter 6 the limits of agreement and the intra-class correlation coefficient may be inconsistent in assessment of agreement. And, how should we interpret discrepant results in agreement studies? As we suggest in chapter 6 the use of more than one strategy and a interpretation of results with their limitations in mind can help, however agreement remains a difficult concept to represent mathematically, and further development of statistical methods needs to be considered.

Motivated by the need of further development of statistical methodology we explore, in chapters 7 and 8, how complexity science can contribute to understand or assess the observer agreement in Medicine, trying to accomplish the **fourth objective**: "to explore whether complexity science and concepts such as entropy could contribute to our understanding of assessment of observer agreement in medicine". In fact, a complex system is a collection of individual agents which act in ways that are not totally predictable and whose actions are interconnected, so that the action of one part changes the context for other agents (Plsek 2001). A complex system behaves in a non-linear fashion, i.e. small changes in input may lead to large changes in outcome (Shiell 2008). Consequently, in complex medical systems, it is difficult to establish predictions or to develop simple cause-effect models (Wilson 2001). In such cases, clinical judgment is known to involve a high degree of uncertainty,

frequently leading to a poor agreement among clinicians over the diagnosis and management, even when hard observable criteria are present. The certainty-agreement diagram (Figure 9.1) related high certainty with high agreement and low certainty (chaos or complexity) with low agreement (Wilson 2001).



Reprinted from BMJ, 323, Suneet Wilson T, Holt T, Greenhalgh T, Complexity science: complexity and clinical care, 685-6888, Copyright (2001), with permission from BMJ Publishing Group Ltd.

Figure 9.1. Adapted certainty-agreement diagram.

As there is a strong relation between complexity and agreement, we think that the complexity analysis may improve the interpretation of agreement using continuous variables. The chapter 7 shows how analysis of entropy, a measure of complexity or uncertainty, and the categorization of continuous variables helps to shed some light into the results of observer agreement and the discrepancies between limits of agreement and ICC found in the previous study presented in chapter 6. The chapter 7 only introduce the idea that complexity science can contribute to understand the observer agreement in Medicine, an idea to develop in the future. Another idea presented in this thesis is the use of complexity science and information theory, not only to understand the lack of agreement in Medicine and to help in the interpretation of discrepant agreement results, but also to develop new measures of agreement. As the usual

approaches used to evaluate agreement (or disagreement) have the limitation of the comparability of populations, in chapter 8, we propose a new information-based approach to assess disagreement which allows the comparability of different trait populations. This new approach is based on the concept of entropy, a measure of complexity or uncertainty (Costa-Santos 2010-b). The proposed information-based measure of disagreement, used as a complement to current approaches for evaluating agreement, can be useful to compare the degree of disagreement among different populations with different characteristics, namely with different variances(Costa-Santos 2010-b). Actually, the major advantage of the proposed information-based measure of disagreement is its comparability across populations. In fact, all other methods used to assess agreement do not allow the comparability across populations. For example ICC is strongly influenced by the variance of the trait in the sample/population in which it is assessed. ICCs measured for different populations might not be comparable (Muller 1994). The proposed measure have however some limitations. It is appropriate for ratio scale measurements (with a natural zero) but it is not appropriate for interval scale measurements (without a natural zero). Yet, the ratio scales are the most common in biomedical research. Another limitation is that it can only measure disagreement between two observers or methods, an extension for more than two observers or methods can be explored in a future work.

9.2 Conclusions

This thesis shows the importance of agreement studies and the frequent misuse of the current statistical strategies to assess agreement. As all method for agreement measurement has important limitations, it should be the overall impression from several statistical approaches that gives one the feeling that a measure is reliable and to what degree. It is important to discuss methods specific drawbacks when interpreting the results of individual agreement study.

Many researchers continue use inappropriate methods to assess agreement imitatively and without thinking about the issues involved. Frequently the most appropriate methods are used but their limitations are not taking in account when the results are interpreted, leading to erroneous conclusion. Inappropriate methods and misinterpreted results of agreement studies can lead to important clinical, research and medico-legal consequences.

This thesis has also shown that information theory and complexity science, the study of complex systems, can make a contribution to the agreement assessment area. In fact, complexity science have been more popular in medical literature in the last years (Campbell 2007; Notcutt 2002; Patel 2008; Plsek 2001-a; Plsek 2001-b; Rambihar 2009; Reid 2002; Shiell. 2008; Wilson 2001). And the strong relation between agreement and uncertainty and complexity lead us to the idea that complexity can provide not only better quantification but also better understanding of the difficulty of the underlying problems related to the measurement of agreement. The development of measures based in uncertainty instead of variance is an open area that we think that is important to explore as future work.

While in simple problems with low uncertainty it is reasonable to use linear models and protocols to help guide clinical diagnosis and decisions, complex situations pose a higher level of uncertainty for clinical judgement, which frequently has to rely on intuition. In these scenarios the use of non-linear models and mining of patterns may be a helpful way to deal with higher levels of complexity and consequently reduce the observer disagreement frequently found in complex scenarios (Figure 9.1) (Wilson 2001).

9.3 References

Ayres-de-Campos, D, Bernardes, J, Costa-Pereira, A and Pereira-Leite, L (1999). "Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision." Br J Obstet Gynaecol 106(12): 1307-1310.

Bernardes, J, Costa-Pereira, A, Ayres-de-Campos, D, van Geijn, HP and Pereira-Leite, L (1997). "Evaluation of interobserver agreement of cardiotocograms." Int J Gynaecol Obstet 57(1): 33-37.

Bland, JM and Altman, DG (2003). "Applying the right statistics: analyses of measurement studies." Ultrasound Obstet Gynecol 22(1): 85-93.

Campbell, NC, Murray, E, Darbyshire, J, Emery, J, Farmer, A, Griffiths, F, Guthrie, B, Lester, H, Wilson, P and Kinmonth, AL (2007). "Designing and evaluating complex interventions to improve health care." BMJ 334(7591): 455-459.

Costa-Santos, C, Costa-Pereira, A and Bernardes, J (2005). "Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences." BJOG 112(5): 667-669.

Costa-Santos, C, Bernardes, J and Ayres-De-Campos, D (2008-a). "Assessment and reporting of the reproducibility of fetal heart rate monitoring." Am J Obstet Gynecol 198(3): 343; author reply 343-344.

Costa-Santos, C and Bernardes, J (2008-b). "Interobserver agreement according to STAN guidelines." Acta Obstet Gynecol Scand 87(11): 1260.

Costa-Santos, C and Bernardes, J (2009). "Differences and consequences in interobserver variation in intrapartum fetal heart rate assessment." Am J Obstet Gynecol 200(6): e8.

Costa-Santos, C, Bernardes, J, Ayres-de-Campos, D, Costa, A and Costa C. (2010-a). "The limits of agreement and the intra-class correlation coefficient may be inconsistent in assessing agreement." J Clin Epidemiol (in press).

Costa-Santos, C, Antunes, L, Souto, A and Bernardes, J (2010-b). "Assessment of disagreement: a new information based approach." Ann Epidemiol (in press).

de Vet, HC (2005). Observe Reliability and Agreement. Encyclopedia of biostatistics. Armitage, Pc, T., John Wiley.

Fleiss, J (1986). The design and Analysis of Clinical Experiments. New York, John Wiley & Sons.

Gow, RM, Barrowman, NJ, Lai, L and Moher, D (2008). "A review of five cardiology journals found that observer variability of measured variables was infrequently reported." J Clin Epidemiol 61(4): 394-401.

Luiz, RR and Szklo, M (2005). "More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported." J Clin Epidemiol 58(3): 215-216.

Muller, R and Buttner, P (1994). "A critical discussion of intraclass correlation coefficients." Stat Med 13(23-24): 2465-2476.

Notcutt, W (2002). "Complexity science. New approaches to evaluation of treatment are needed." BMJ 324(7330): 171-172.

Patel, AM, Sundt, TM, 3rd and Varkey, P (2008). "Complexity science: core concepts and applications for medical practice." Minn Med 91(2): 40-42.

Plsek, PE and Greenhalgh, T (2001-a). "Complexity science: The challenge of complexity in health care." BMJ 323(7313): 625-628.

Plsek, PE and Wilson, T (2001-b). "Complexity, leadership, and management in healthcare organisations." BMJ 323(7315): 746-749.

Rambihar, VS and Rambihar, V (2009). "Complexity science may help in defining health." BMJ 338: b32.

Reid, I (2002). "Complexity science. Let them eat complexity: the emperor's new toolkit." BMJ 324(7330): 171.

Shiell, A, Hawe, P and Gold, L (2008). "Complex interventions or complex systems? Implications for health economic evaluation." BMJ 336(7656): 1281-1283.

Shoukri, M (2005). Agreement, Measurement of. E. Armitage, Pc, T., John Wiley.

Stratford, P (1989). "Reliability: consistency or differentiating among subjects?" Phys Ther 69(4): 299-300.

Wilson, T, Holt, T and Greenhalgh, T (2001). "Complexity science: complexity and clinical care." BMJ 323(7314): 685-688.