Reliability and agreement

Cristina Costa Santos
MEDCIDS – Departamento Medicina da Comunidade, Informação e Decisão em Saúde
CINTESIS - Centro de Investigação em Tecnologias e Serviços de Saúde
Faculdade de Medicina da Universidade do Porto

- The quality of the measurements taken by health professionals or by measurement devices is fundamental not only for clinical care but also for research

- Measurement of variables always implies some degree of error.

- When an observer takes a measurement, the value obtained depends on several things such us:
  - the skills of the observer,
  - observer experience,
  - the measurement instrument,
  - observer’s expectations
  - ...
- Also, natural continuous variation in a biological quantity can be presente.

- When natural continuous variation in a biological quantity is present, it is outside the control of the observer.

- It is, however, possible to minimize the observer variability by:
  - training of observers,
  - use of guidelines
  - automation, ...

Reliability and agreement studies

before

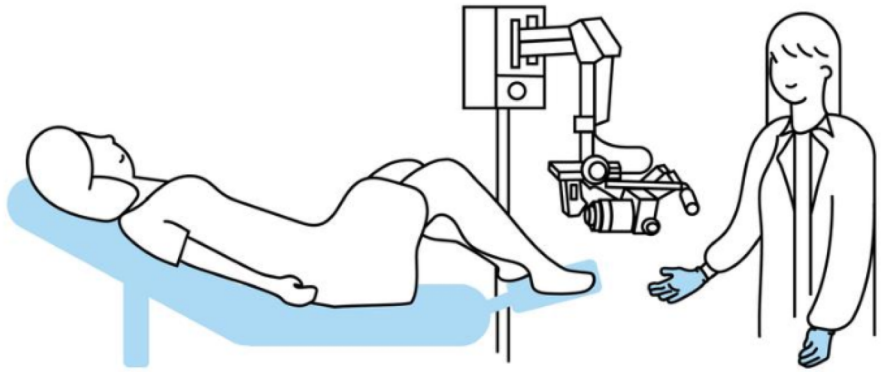
Validity studies / RCT / ...

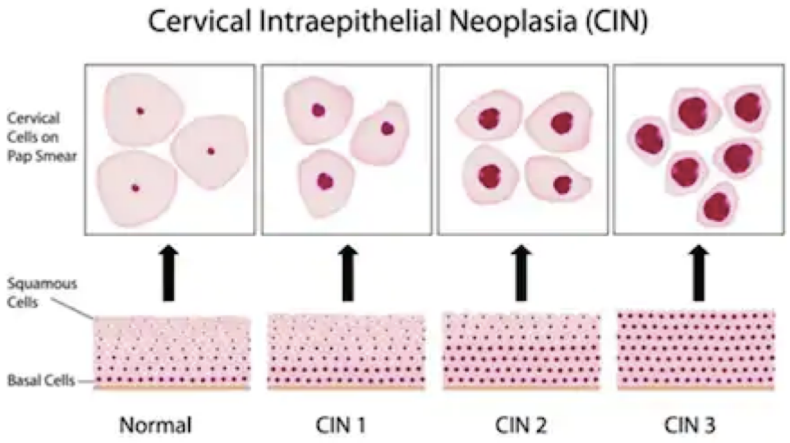
Agreement studies are very importante

however

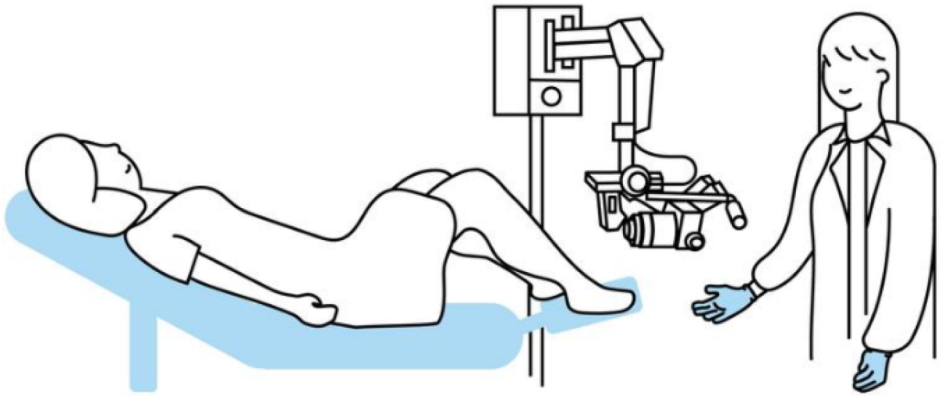
neglected in medical literature

- What is, for example, the sensitivity and specificity of colposcopy and of cervical cytology in detection of Cervical Neoplasia?



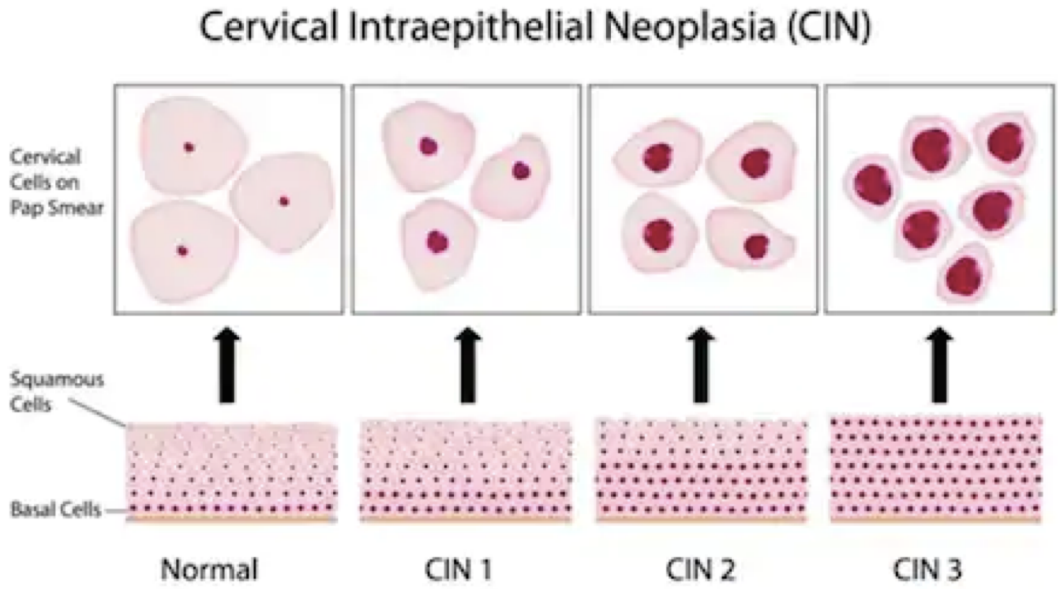


- Mitchell et al. found a variation in sensitivity and specificity of colposcopy ranging from 87% to 100% and 26 to 87%, respectively, in nine published studies using similar methods.

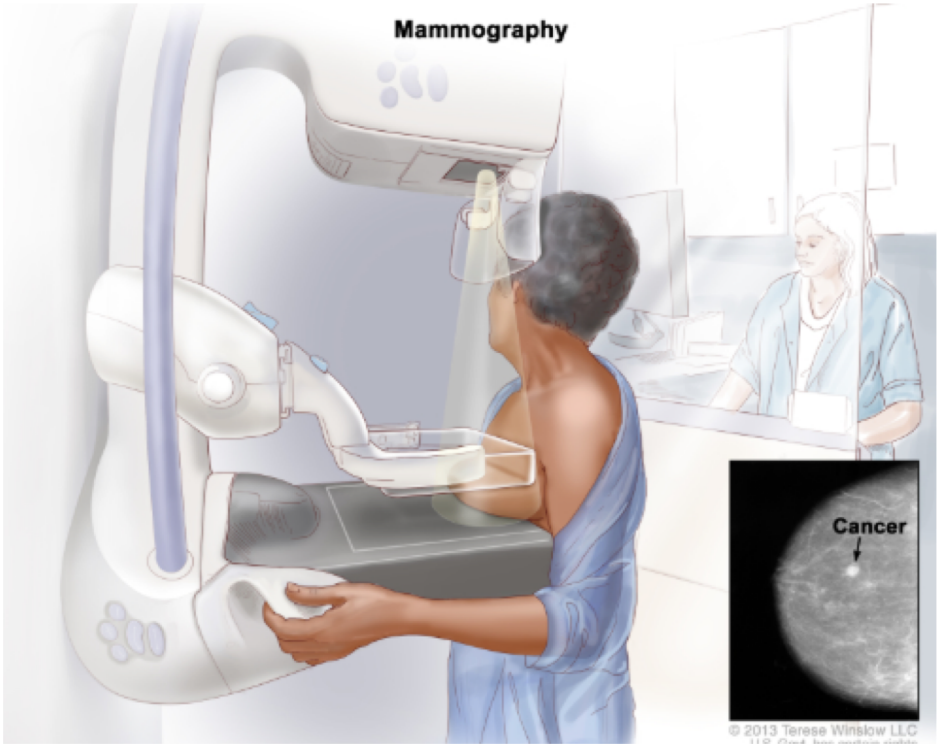


Mitchell et al.
Flourescence spectroscopy for diagnosis of squamous intraepithelial lesions of the cervix. Obstet Gynecol 1999;93:462- 470.

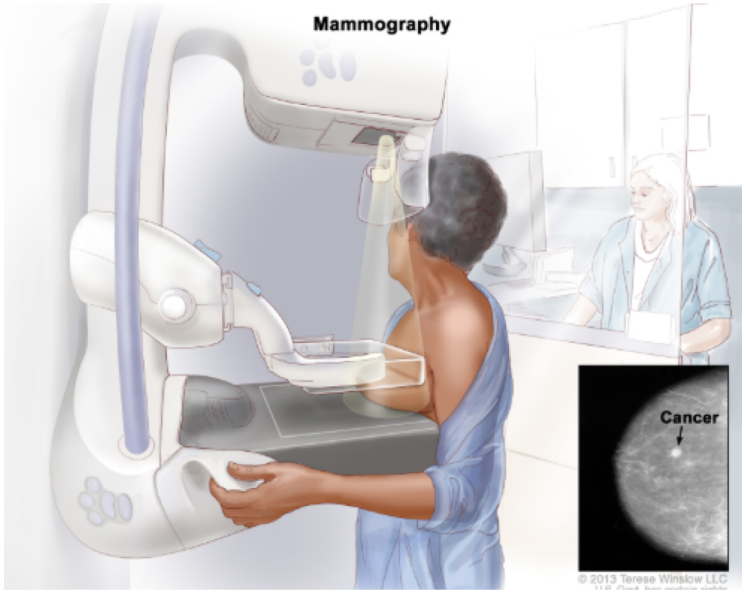
- The same authors found a variation in sensitivity and specificity of cervical cytology from 0 to 93% and 0 to 100%, respectively, in 26 published studies using similar methods.



- And, what are 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 metaanalyses or are they just the mean result of discrepant results from the different randomised clinical trial included in meta-analyses?

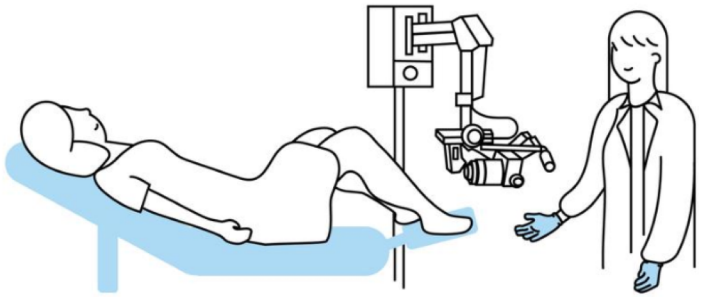


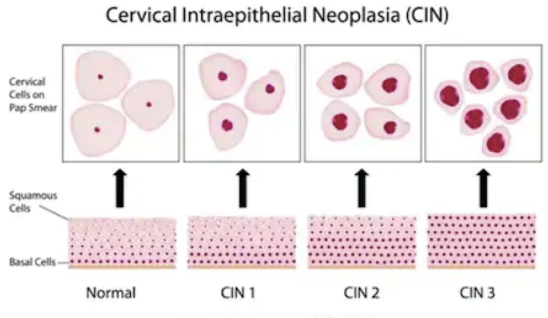
Thacker SB, Stroup DF, Peterson HB. Efficacy and safety of intrapartum electronic fetal monitoring: an update. Obstet Gynecol 1995;86:613– 620. 8. Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet 2001;358(9290):1340– 1342.

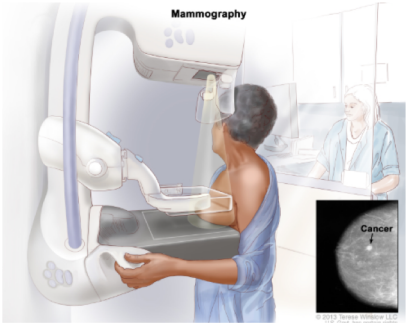
In fact, several studies reveled that all these wildly used methdos

(colposcopy, cervical cytology, cardiotocography and mammography)

have very low inter-observer agreement among among the doctors who perform them or who analyse them.







- The terms “reliability” and “agreement” are often used interchangeably.

However, the two concepts are conceptually distinct

- **Reliability** can be defined as the ability of a measurement to differentiate between subjects.

- **Agreement** is the degree to which scores or ratings are identical.

- Both concepts are important, because they provide information about the quality of measurements.
- The study designs for examining the two concepts are similar.

ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 96–106

Journal of Clinical Epidemiology

Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed

Jan Kottner[a,*], Laurent Audigé[b], Stig Brorson[c], Allan Donner[d], Byron J. Gajewski[e], Asbjørn Hróbjartsson[f], Chris Roberts[g], Mohamed Shoukri[h], David L. Streiner[i]

[a]*Department of Nursing Science, Centre for Humanities and Health Sciences, Charité-Universitätsmedizin Berlin, Berlin, Germany*
[b]*AO Clinical Investigation and Documentation, Dübendorf, Switzerland*
[c]*Department of Orthopaedic Surgery, Herlev University Hospital, Herlev, Denmark*
[d]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario, Canada*
[e]*Department of Biostatistics, University of Kansas Schools of Medicine & Nursing, Kansas City, KS, USA*
[f]*The Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark*
[g]*School of Community Based Medicine, The University of Manchester, Manchester, UK*
[h]*Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, The University of Western Ontario, London, Ontario, Canada*
[i]*Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada*

Accepted 2 March 2010

Measures of agreement

Statistical methods for analyzing interrater/intrarater reliability and agreement studies

Level of measurement	Reliability measures	Agreement measures
Nominal	Kappa statistics	Proportions of agreement
		Proportions of specific agreement
Ordinal	Ranked intraclass correlation	Proportions of agreement
	Matrix of kappa coefficients	Proportions of specific agreement
	Weighted kappa	
Continuous	Intraclass correlation coefficients	Proportions of agreement (ranges)
		Proportions of specific agreement (ranges)
		Standard errors of measurement
		Coefficients of variation
		Bland–Altman plots and limits of agreement

Cohen's Kappa coefficient

Two raters are asked to classify objects into categories 1 and 2.
The table below contains cell probabilities for a 2 by 2 table.

		Rater #1		
		1	2	Total
Rater #2	1	p_{11}	p_{12}	$p_{1.}$
	2	p_{21}	p_{22}	$p_{2.}$
	Total	$p_{.1}$	$p_{.2}$	1

Proportions of agreement (Po) = $P_{11}+P_{22}$

Proportions of espected by chance (Pe) = $P_{.1}\,P_{1.} + P_{.2}\,P_{2.}$

$$K = \frac{Po - Pe}{1 - Pe}$$

Chance-Corrected Agreement? Or measure of relibility?

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 196037-46, 1960.

High agreement but low kappa

		Observer 1			
		yes	no	total	
Observer 2	yes	40	5	45	
	no	3	2	5	
	total	43	7	50	

PA=42/50=0.84

Pe=(43/50)(45/50)+(7/50)(5/50)=0.79

K=(0.84-0.79)/(1-0.79)=0.24

Specific agreement

Summary of binary ratings by two raters

		Rater 2		
Rater 1		**+**	**-**	*total*
	+	**a**	**b**	a + b
	-	**c**	**d**	c + d
total		a + c	b + d	N

The values a, b, c and d here denote the observed frequencies for each possible combination of ratings by Rater 1 and Rater 2.

Proportion of overall agreement = (a+d)/N

$$PA = \frac{2a}{2a + b + c}; \quad NA = \frac{2d}{2d + b + c}.$$

PA estimates the conditional probability, given that one of the raters, randomly selected, makes a positive rating, the other rater will also do so

NA estimates the conditional probability, given that one of the raters, randomly selected, makes a negative rating, the other rater will also do so

Cicchetti DV. Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J ournal of Clinical Epidemiology, 1990, 43, 551-558.

Spitzer R, Fleiss J. A re-analysis of the reliability of psychiatric diagnosis. British Journal on Psychiatry, 1974, 341-47.

High agreement but low kappa

		Observer 1		
		yes	no	total
Observer 2	yes	40	5	45
	no	3	2	5
	total	43	7	50

PA=42/50=0.84

K=(0.84-0.79)/(1-0.79)=0.24

PA yes = 40x2/(40x2+3+5)=0.92

PA no = 2x2/(2x2+3+5)=0.33

Intraclass Correlation Coefficient

- The Intraclass Correlation (ICC) assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects.

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bulletin 1979;86:420-427.

Information Based Measure of Disagreement

- 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 Y and Observer X .
  The disagreement between Y and X is related to the differences between them. So, we consider

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

  the amount of information contained in the differences between observers.

Costa-Santos C. et al. Assessment of Disagreement: A New Information-Based Approach. ANNALS OF EPIDEMIOLOGY. 20(7):555-561

By adding 1 to the differences, we avoid the behavior of the logarithmic function between 0 and 1.

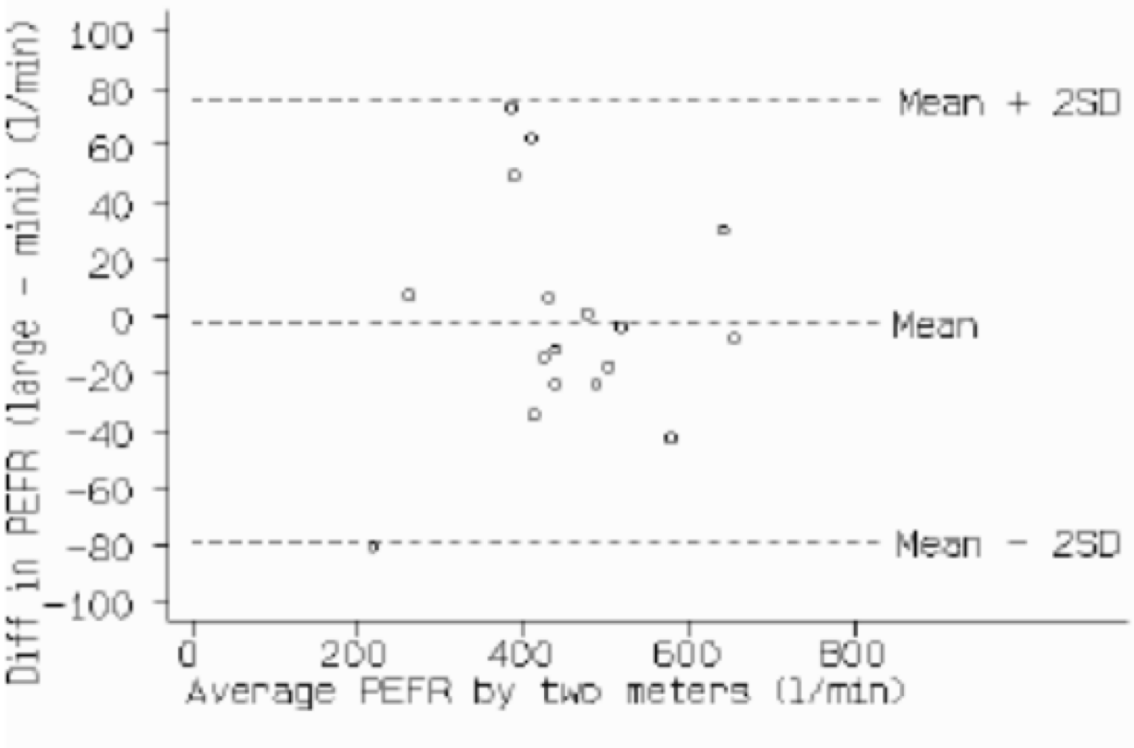
To get a value between 0 and 1 we normalize the amount of information contained in the differences to obtain the following measure of information-based measure of disagreement (IBMD):

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

0 – no disagreement
Tends to 1 (total disagreement)

Bland and Altman limits of agreement

Difference against mean



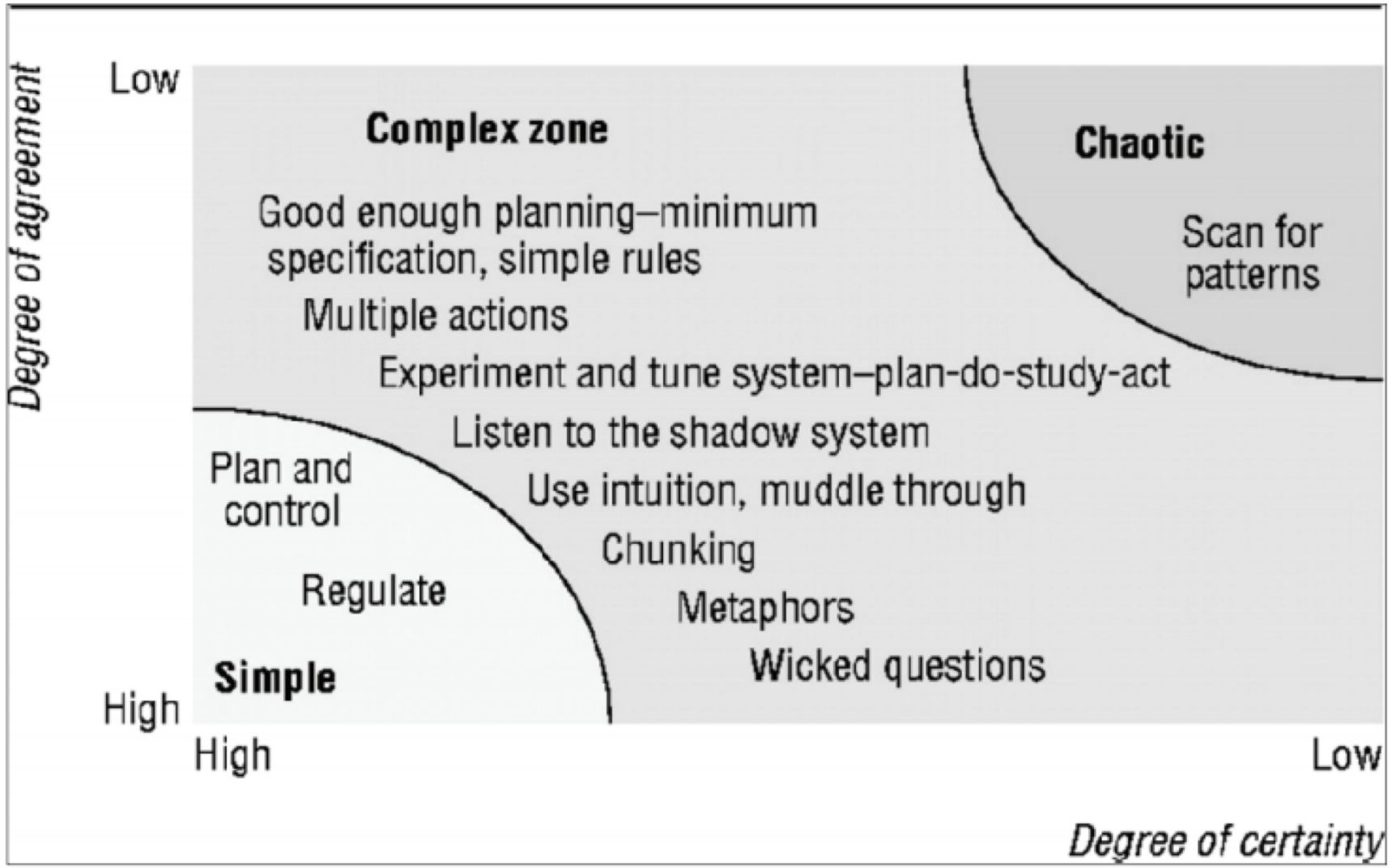
Provided differences within *mean differences +/- 2SD* would not be clinically important, we could use the two measurement methods interchangeably.

IS A CLINICAL (NOT STATISTICAL) INTERPRETATION

*Bland JM, Altman DG. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, **i**, 307-310.*

Cardiotocography CTG guidelines:

- FIGO, International Federation of Gynecology and Obstetrics;
- ACOG, the American College of Obstetricians and Gynecologists,
- NICE, National Institute for Health and Care Excellence.



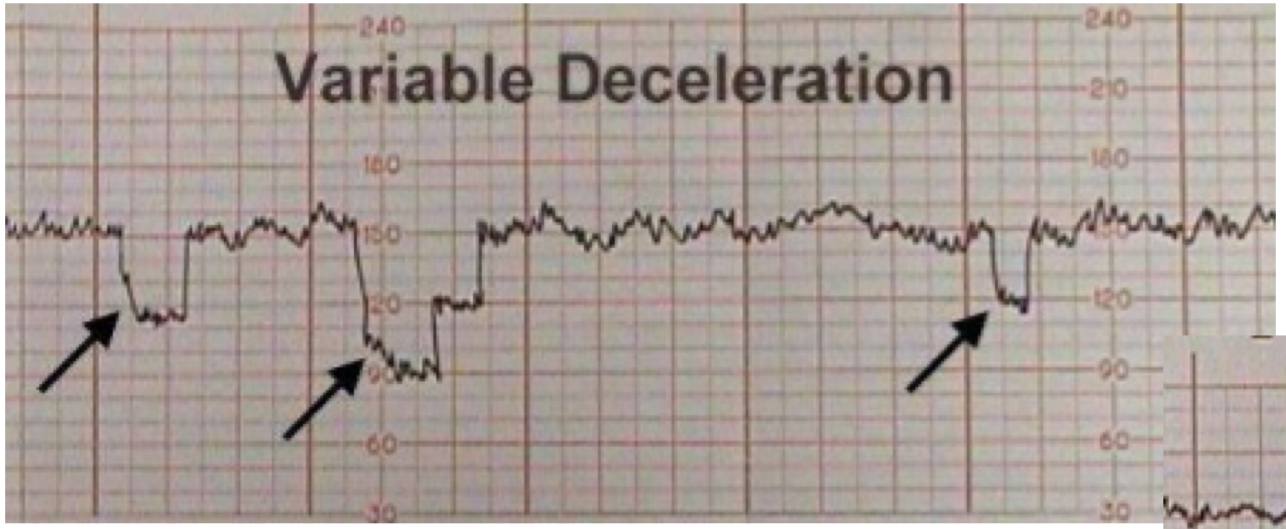
Reprinted from BMJ, 323, Suneet Wilson T, Holt T, Greenhalgh T, Complexity science: complexity and clinical care, 685-6888, Copyright (2001), with permission from BMJ Publishing Group Ltd.

AOGS ORIGINAL RESEARCH ARTICLE

Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines

SUSANA SANTO[1], DIOGO AYRES-DE-CAMPOS[2], CRISTINA COSTA-SANTOS[3], WILLIAM SCHNETTLER[4], AUSTIN UGWUMADU[5] & LUÍS M. DA GRAÇA[1] FOR THE FM-COMPARE COLLABORATION*

[1]Department of Obstetrics and Gynecology, Santa Maria Hospital, Faculty of Medicine of Lisbon University, Lisbon, [2]Department of Obstetrics and Gynecology, Medical School, University of Porto, S. João Hospital, Institute of Biomedical Engineering, Porto, [3]Department of Medical Informatics, Medical School, University of Porto, Porto, Portugal, [4]Center for Maternal Cardiac Care, TriHealth, Good Samaritan Hospital, Cincinnati, OH, USA, and [5]Department of Obstetrics & Gynecology, St George's Hospital, University of London, London, UK


Variable Deceleration
240
210
180
150
120
90
60
30

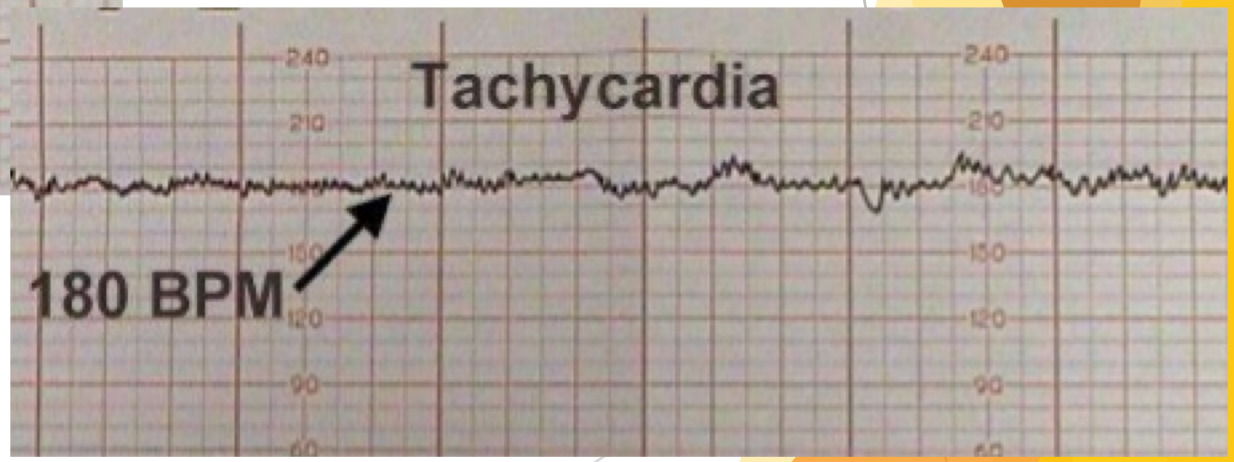

Tachycardia
180 BPM
240
210
150
120
90

Table 2. Comparison of cardiotocography (CTG) classification criteria in the FIGO, NICE and ACOG guidelines.

FIGO	NICE	ACOG
Normal pattern • Baseline heart rate between 110 and 150 bpm • Amplitude of heart rate variability between 5 and 25 bpm	Normal (a CTG where all of the following four reassuring features are present) • Baseline rate: 110–160 bpm • Variability: ≥5 bpm • No decelerations • Accelerations: present	Category I (category I FHR tracings include all of the following) • Baseline rate: 110–160 bpm • Baseline variability: 6–25 bpm • Late or variable decelerations: absent • Early decelerations: present or absent • Accelerations: present or absent
Suspicious pattern • Baseline heart rate between 150 and 170 bpm or between 100 and 110 bpm • Amplitude of variability between 5 and 10 bpm for more than 40 min • Increased variability above 25 bpm • Variable decelerations	Suspicious (a CTG where one of the following features is present and all others fall into the reassuring category) • Baseline rate - 100–109 bpm - 161–180 bpm • Baseline variability - <5 bpm for 40–90 min • Decelerations - Typical variable decelerations with >50% of contractions occurring for >90 min - Single prolonged deceleration for up to 3 min	Category II (Category II FHR tracings include all FHR tracings not categorised as Category I or Category III. Examples of Category II FHR tracings include any of the following) • Baseline rate - Bradycardia not accompanied by absent baseline variability - Tachycardia • Baseline variability - Minimal variability - Absent variability with no recurrent decelerations - Marked variability • Accelerations

	FIGO 1987		ACOG 2010		NICE 2007	
	PA (95% CI)	κ (95% CI)	PA (95% CI)	κ (95% CI)	PA (95% CI)	κ (95% CI)
FHR baseline	0.81 (0.78–0.85)	0.63 (0.57–0.70)	0.88 (0.84–0.91)	0.59 (0.49–0.69)	0.88 (0.85–0.91)	0.65 (0.58–0.72)
Normal	0.86 (0.83–0.89)		0.93 (0.90–0.95)		0.93 (0.91–0.96)	
Tachycardia	0.80 (0.73–0.85)		0.67 (0.57–0.77)		0.73 (0.66–0.80)	
Bradycardia	0.40 (0.25–0.54)		0.49 (0.07–0.71)		0.42 (0.00–1.00)	
Variability	0.83 (0.80–0.86)	0.51 (0.42–0.61)	0.85 (0.82–0.88)	0.49 (0.39–0.59)	0.83 (0.80–0.86)	0.38 (0.29–0.50)
Normal	0.89 (0.87–0.92)		0.91 (0.89–0.93)		0.90 (0.88–0.92)	
Abnormal	0.61 (0.51–0.69)		0.57 (0.45–0.66)		0.44 (0.32–0.53)	
Accelerations	0.67 (0.64–0.71)	0.34 (0.29–0.41)	0.67 (0.63–0.70)	0.34 (0.28–0.40)	0.71 (0.68–0.75)	0.41 (0.35–0.48)
Yes	0.73 (0.68–0.77)		0.72 (0.67–0.76)		0.61 (0.57–0.68)	
No	0.59 (0.54–0.64)		0.59 (0.53–0.64)		0.76 (0.72–0.80)	
Decelerations	0.92 (0.89–0.95)	0.53 (0.43–0.66)	0.85 (0.82–0.88)	0.28 (0.18–0.46)	0.89 (0.85–0.91)	0.47 (0.35–0.59)
Yes	0.96 (0.94–0.97)		0.92 (0.89–0.93)		0.94 (0.92–0.95)	
No	0.59 (0.48–0.69)		0.35 (0.23–0.45)		0.49 (0.36–0.59)	
Classification	0.64 (0.61–0.67)	0.37 (0.31–0.43)	0.73 (0.70–0.76)	0.15 (0.10–0.21)	0.55 (0.51–0.58)	0.33 (0.28–0.39)
Cat. I/Normal	0.54 (0.39–0.64)		0.26 (0.18–0.33)		0.55 (0.48–0.62)	
Cat. II/Suspicious	0.67 (0.62–0.70)		0.83 (0.81–0.86)		0.42 (0.38–0.47)	
Cat. III/Pathological	0.63 (0.57–0.68)		0.26 (0.18–0.34)		0.66 (0.59–0.71)	

"Fine, tell him he's going to make it.
We'll just have to agree to disagree."