MEASUREMENT ERROR

Ms Chido Dziva Chikwari; BSc, MSc, PhDc
Biomedical Research and Training Institute
March 12, 2019

Sources of Error in Epi

Chance (Random Error)

Bias (Systematic Error)
*errors in the design or conduct of the study i.e. responsibility of the researcher.

Confounding (Systematic Error)
*a problem out there i.e. a real association in the population.

Selection Bias

Information Bias
-Measurement Error
-Misclassification

Objectives

- 1) What do we mean by measurement error and misclassification?
- 2) How does it arise?
- 3) How can we quantify it?
- 4) What are the consequences for epidemiological studies?

Objectives

- 1) What do we mean by measurement error and misclassification?

Terminology: instrument, measurement error/misclassification, validity, reliability

- 2) How does it arise?

Poor design, inadequate protocol, poor execution
Data entry/analysis

- 3) How can we quantify it?

Validity: Plots, sensitivity, specificity; Reliability: Kappa statistic

- 4) What are the consequences for epidemiological studies?

Information bias. Non-differential vs differential misclassification

Objectives

- 1) <u>What do we mean</u> by measurement error and misclassification?

What is Measurement Error?

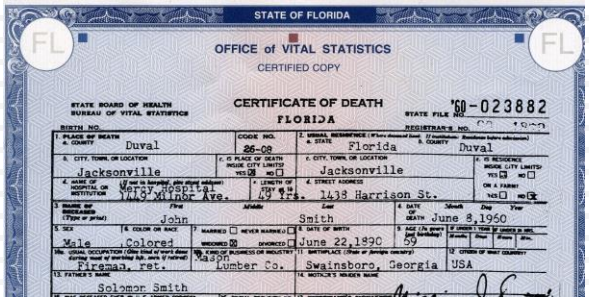
- Difference between the "measured"/recorded value and "true" value
  - Can apply to continuous variables • e.g. blood pressure, height

What is Measurement Error – Misclassification ?

...or categorical/binary data, e.g.

– Smoker misclassified as a non-smoker

– Recorded cause of death incorrect



STATE OF FLORIDA

OFFICE of VITAL STATISTICS

CERTIFIED COPY

STATE BOARD OF HEALTH
BUREAU OF VITAL STATISTICS

CERTIFICATE OF DEATH
FLORIDA

STATE FILE NO. 60-023882

BIRTH NO. REGISTRAR'S NO.

1. PLACE OF DEATH a. COUNTY: Duval — CODE NO.: 26-08
2. USUAL RESIDENCE (Where deceased lived) a. STATE: Florida b. COUNTY: Duval
3. a. CITY, TOWN, OR LOCATION: Jacksonville — c. IS PLACE OF DEATH INSIDE CITY LIMITS? YES ☒ NO ☐
c. CITY, TOWN, OR LOCATION: Jacksonville — e. IS RESIDENCE INSIDE CITY LIMITS? YES ☒ NO ☐
4. NAME OF HOSPITAL OR INSTITUTION: Mercy Hospital — f. LENGTH OF STAY IN THIS PLACE: 45 hrs. — 4. STREET ADDRESS: 1438 Harrison St. — ON A FARM? NO ☐ NO ☒
1419 Milner Ave.
3. NAME OF DECEDENT (Type or print): First: John — Middle: — Last: Smith — 4. DATE OF DEATH: June 8,1960
5. SEX: Male — 6. COLOR OR RACE: Colored — 7. MARRIED ☐ NEVER MARRIED ☐ WIDOWED ☒ DIVORCED ☐ — 8. DATE OF BIRTH: June 22,1890 — 9. AGE (In years if over 1 year, months if under 1 year, days if under 1 day): 69
10a. USUAL OCCUPATION (Give kind of work done during most of working life, even if retired): Firman, ret. — 10b. KIND OF BUSINESS OR INDUSTRY: Mason Lumber Co. — 11. BIRTHPLACE (State or foreign country): Swainsboro, Georgia — 12. CITIZEN OF WHAT COUNTRY: USA
13. FATHER'S NAME: Solomon Smith — 14. MOTHER'S MAIDEN NAME:

Terminology – Instrument

- Just a means of measuring something.

Terminology – Instrument

- Just a means of measuring something.
- A device to measure blood pressure
- A questionnaire to measure a macro or micro-nutrient
- A test to measure HIV status

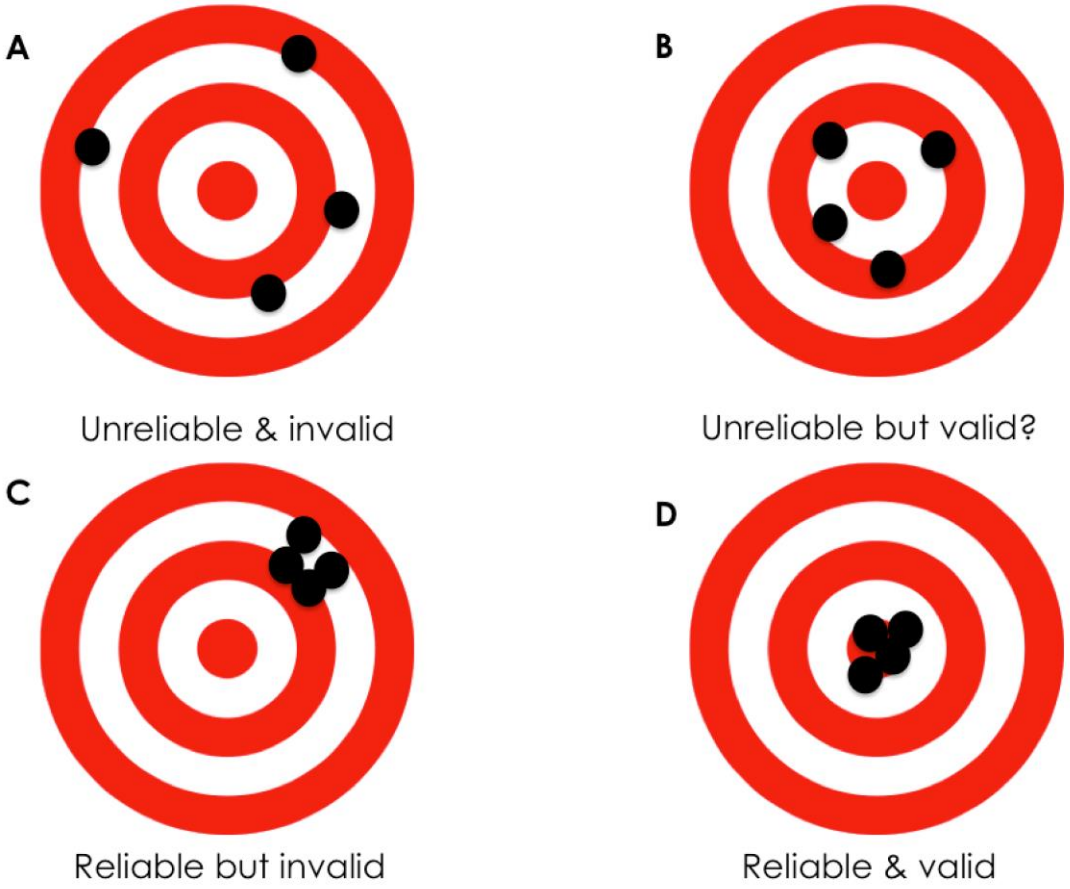
	Questions	0	1	2	3	4
1	How do you feel about the pleasure you get from food, compared with the time when you had natural teeth?	○	○	○	○	○
2	With respect to chewing, how satisfied are you with your dentures?	○	○	○	○	○
3	With respect to appearance, how satisfied are you with your dentures?	○	○	○	○	○
4	With respect to how comfortable your dentures are, how satisfied are you?	○	○	○	○	○
5	With respect to being self-assured and self-conscious, how satisfied are you with your dentures?	○	○	○	○	○
6	With respect to your social and affective relationships, how satisfied are you with your oral conditions?	○	○	○	○	○
7	With respect to your professional performance, how satisfied are you with your oral conditions?	○	○	○	○	○
8	With respect to eating, how satisfied are you with your dentures?	○	○	○	○	○
9	Are you satisfied with your smile (esthetics)?	○	○	○	○	○

Terminology – Validity and Reliability

- What is the difference between them?

Terminology – Validity and Reliability

- What is the difference between them?

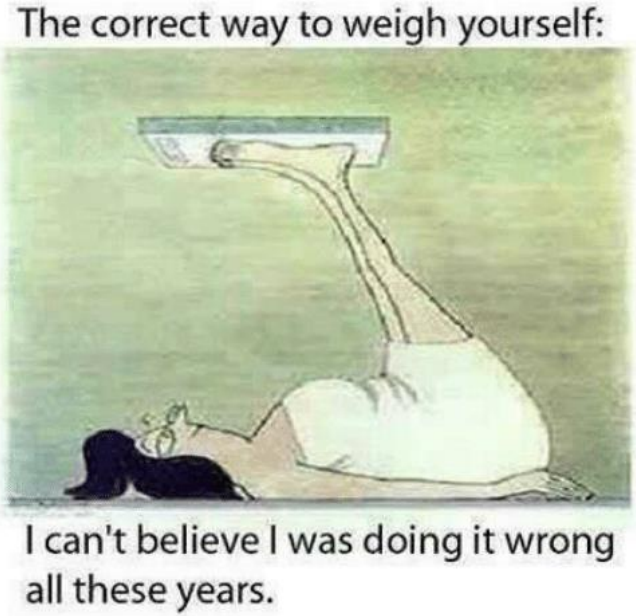


Objectives

- 2) How does it arise?

Poor Design

- E.g. Questionnaire questions
  - When did you start drinking regularly?
  - Are you (A) married; or (B) single?
- • Poorly calibrated weighing scale



Poor Design

Other examples of poorly designed questions:

- Have you ever been a smoker?
- Do you regularly wash your hands after using the toilet?
- How much do you weigh?
- Other Examples?

Poor Instructions

- Insufficient detail in protocol

e.g. “collect blood samples from eligible household members”

- Insufficient training of staff

Best practice: Interviewers are sent self study materials as well as have 5 days of classroom instructions –go through protocol, questionnaire, consent and other forms, answer questions, mock interviews etc.

+on the job supervision and quality control

- See the Demographic and Health Surveys websites for good examples of standard operating procedures (SOPs)

Poor Execution

- Failure to follow protocol/read instructions
- Poor supervision
- Improper handling of specimens

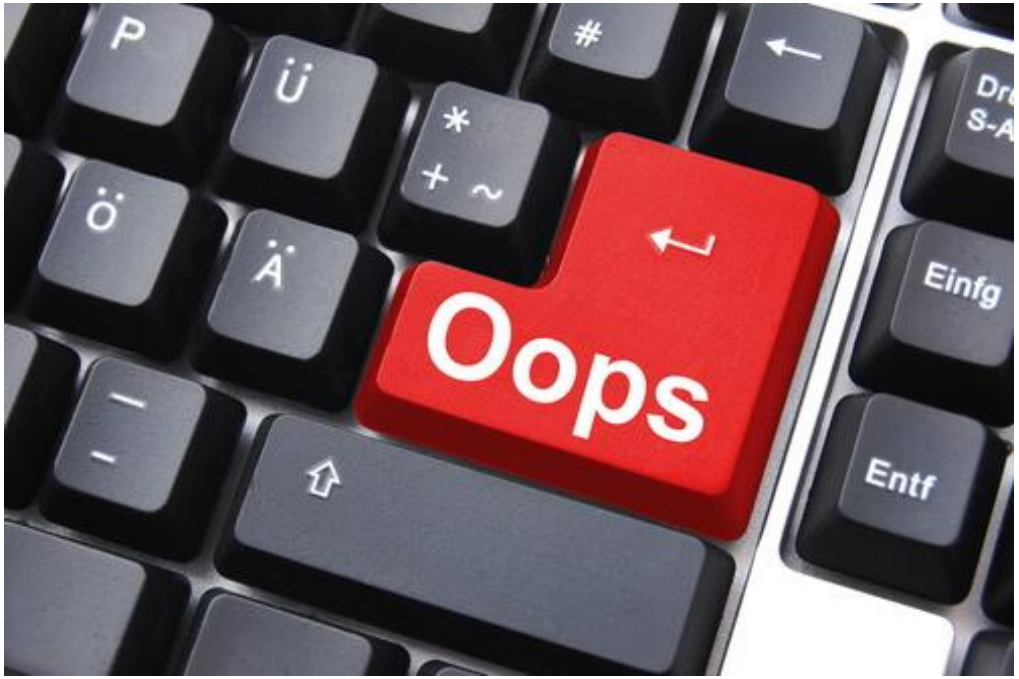
Poor Execution

Study participants

- Failure to remember
- Limited knowledge of proxy respondents

Errors in Data entry

- Data entry errors (~1 in 100 key strokes)
- Programming errors



Objectives

- 3) How can we <u>quantify</u> it?

Validity

- Extent to which the instrument measures the characteristic of interest, e.g.
  - How well does the blood pressure (BP) monitor measure TRUE BP?
  - How well does questionnaire capture TRUE fatty food intake?
  - How well does oral HIV test identify TRUE HIV status?

Validity

- Extent to which the instrument measures the characteristic of interest, e.g.
  - How well does the blood pressure (BP) monitor measure TRUE BP?
  - How well does questionnaire capture TRUE fatty food intake?
  - How well does oral HIV test identify TRUE HIV status?
- Need to know the validity of each specific measure and the standardised methods to carry out each test or tests in the correct order and/or number of times
- How do we quantify validity?

Example

DIAGNOSTIC ACCURACY OF HIV ORAL RAPID TESTS VERSUS BLOOD BASED RAPID TESTS AMONG CHILDREN

CROI 2019
Poster 0782

Chido Dziva Chikwari[1,2*], Irene N. Njuguna[3,4*], Jillian Neary[7], Crissi Rainer[5], Belinda Chihota[6], Jennifer A. Slyker[3], David Katz[3], Dalton C. Wamalwa[7], Laura Oyiengo[8], Tsitsi Bandason[2], Grace McHugh[2], Ethel Dauya[2], Kearsley A Stewart[5], Grace C. John-Stewart[3], Rashida Ferrand[1,2], Anjuli D. Wagner[3] (*co-first authors)

[1]London School of Hygiene and Tropical Medicine; [2] Biomedical Research and Training Institute; [3]University of Washington; [4]Kenyatta National Hospital; [5]Duke University; [6]Centre for Infectious Disease Research in Zambia; [7]University of Nairobi; [8]Kenya Ministry of Health;

Introduction

- Gaps persist in HIV testing globally for children who missed testing as part of prevention of mother to child transmission (PMTCT) programs
- Saliva based tests (SBT) have high sensitivity and specificity (98.0% and 99.7%) in adults but performance has not been established in children (18 months to 12 years)
- SBT may be less traumatic, easy to perform at triage, and pose less risk to health care workers than blood-based tests (BBT)

Objective

- To validate OraQuick ADVANCE Rapid HIV-1/2 saliva based antibody test (SBT) against blood based rapid testing (BBT) in children aged 18 months to 18 years in Kenya and Zimbabwe

Methods

- Antiretroviral therapy (ART)-naïve children were tested for HIV using a series of rapid BBT and SBT
- BBT followed Kenyan and Zimbabwean national algorithms
  - Determine (3rd and 4th generation in Kenya and Zimbabwe respectively), followed by First Response if Determine was reactive
- SBT samples collected and interpreted by research staff
- BBT performed and interpreted by clinic or research staff
- Sensitivity and specificity calculated using BBT national algorithms as gold standard; secondary analysis excluded 2 cases where SBT was positive but national algorithm was initially falsely negative
- Binomial distribution used for 95% confidence intervals [95%CI]

Results

Table 1: Baseline characteristics

	BBT HIV positive n=71	BBT HIV negative n=1705
Child characteristics	n (%) or median (IQR)	n (%) or median (IQR)
Age (years)	6.8 (4.2, 11.0)	7.4 (4.7, 11.6)
18-<24 months	1 (1)	1 (0.1)
2-5 years	21 (30)	491 (29)
>5-12 years	34 (48)	811 (48)
>12-18 years	15 (21)	402 (24)
Female	46 (65)	872 (51)
Recruitment		
Zimbabwe	28 (39)	1542 (90)
Kenya	43 (61)	163 (10)

Table 2: Performance of SBT vs BBT

		BBT		
		Positive	Negative	Total
SBT	Positive	71	2	73
	Negative	0	1703	1703
	Total	71	1705	1776

Sensitivity: 100% (97.5% CI 94-100)
Specificity: 99.9% (95% CI 99.5-100)

Excluding children where BBT was incorrect

- 2 truly positive children tested SBT positive and BBT negative
  - 9 year old, mom positive, confirmed positive by ELISA 1 week after initial BBT
  - 2 year old child was confirmed positive by First Response and INSTI
- Excluding the 2 children

Sensitivity: 100% (97.5% CI 94-100)
Specificity: 99.9% (97.5% CI 99.8-100)

Stability of results (Kenyan sites)

- Among 43 children with positive SBT at 20 minutes
  - 43 (100%) had positive SBT at 40 minutes
- Among the 163 children with negative SBT at 20 minutes
  - 163 (100%) had a negative SBT at 40 minutes

Strength of test results from manufacturer reading cards (Kenya sites)

- Among 43 positive SBT results:
- Strongly positive results:
  - 26 (60%) at 20 minutes
  - 29 (67%) at 40 minutes
- Weakly positive results:
  - 3 weakly positive at 20 minutes, all strongly positive at 40 minutes

Conclusions

- **SBT tests have high sensitivity and specificity in ART-naïve children and adolescents**
- Considerations to expand use of SBT in children are warranted
- As in adults, recommendations should include a warning not to use SBT in children on ART
- The ease and safety of SBT may allow HIV testing at outpatient triage or allow task shifting from HCW to caregivers
- Future research will explore the acceptability and uptake in diverse settings (in and out of facilities) as well as by diverse users (caregivers and HCW)

ACKNOWLEDGEMENTS

Global WACh, Kizazi, Kenya Research & Training Center (KRTC)

Study team and participants

Funding provided by University of Washington Center for AIDS Research and Thrasher Pediatric Research Foundation

BGAP study team

Funding provided by: Duke Global Health Institute, the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union (MR/P011268/1)

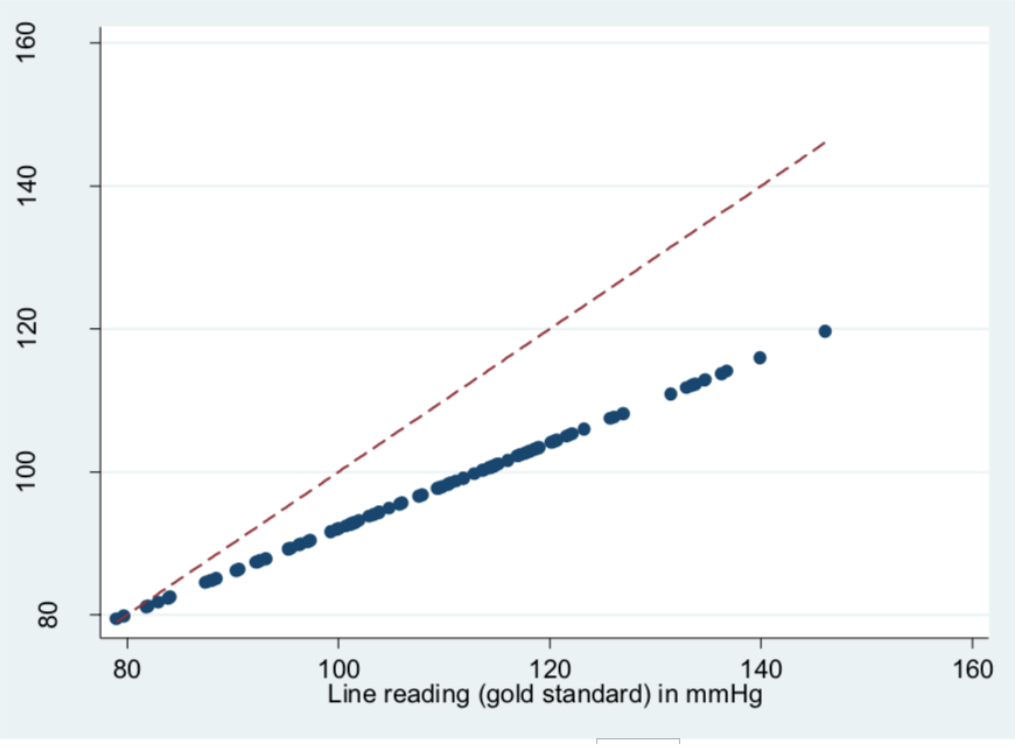
Quantifying validity – continuous variable

- Need some data on instrument measure vs "true" measure

Patient	Cuff SBP (mmHg)	Arterial line SBP (mmHg)
1	103	105
2	125	125
3	91	96
4	136	132
5	111	110
6	110	112
…	130	125

Quantifying validity – continuous variable

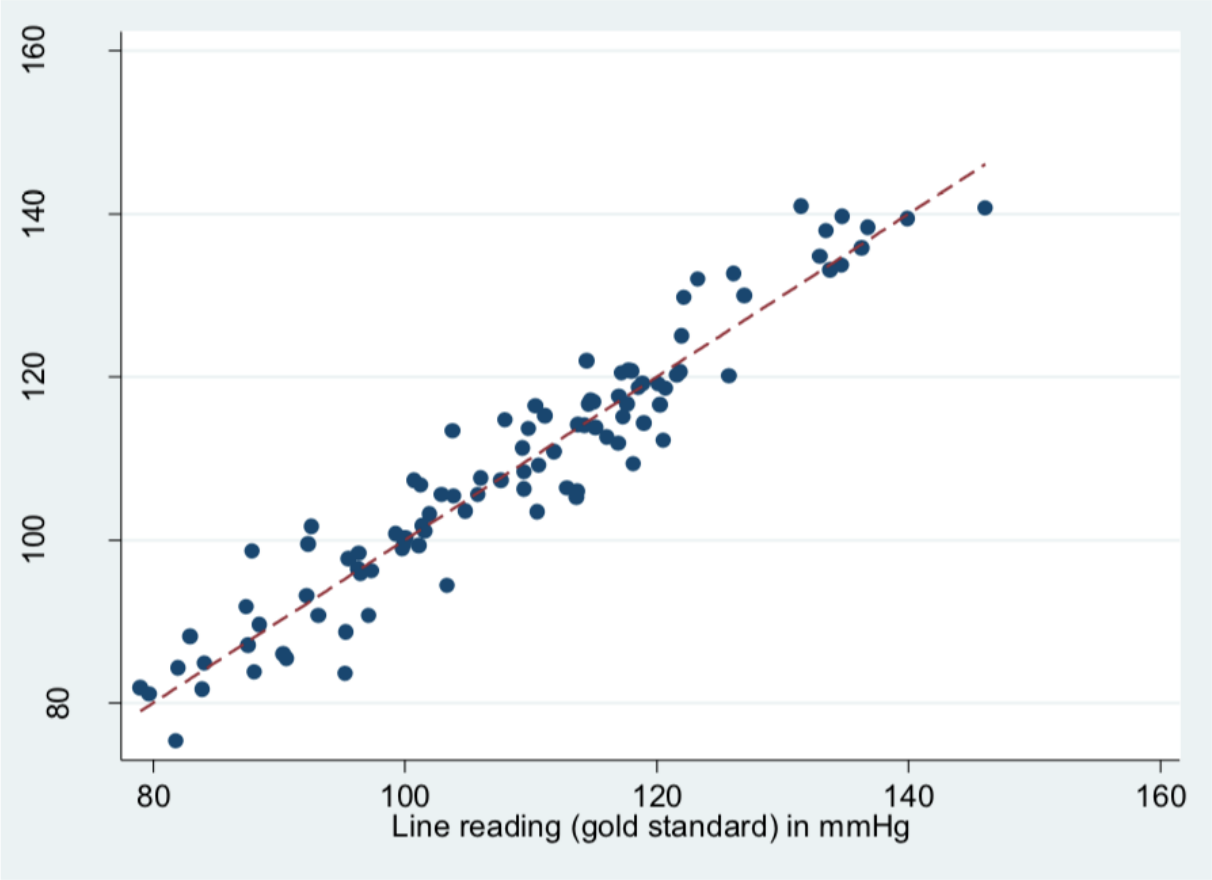
- Pearson correlation?
  - Often used
  - …but measures association NOT agreement



ρ = 1 i.e. perfectly correlated but cuff underestimates BP

Quantifying validity – continuous variable

- Do an initial raw scatter plot



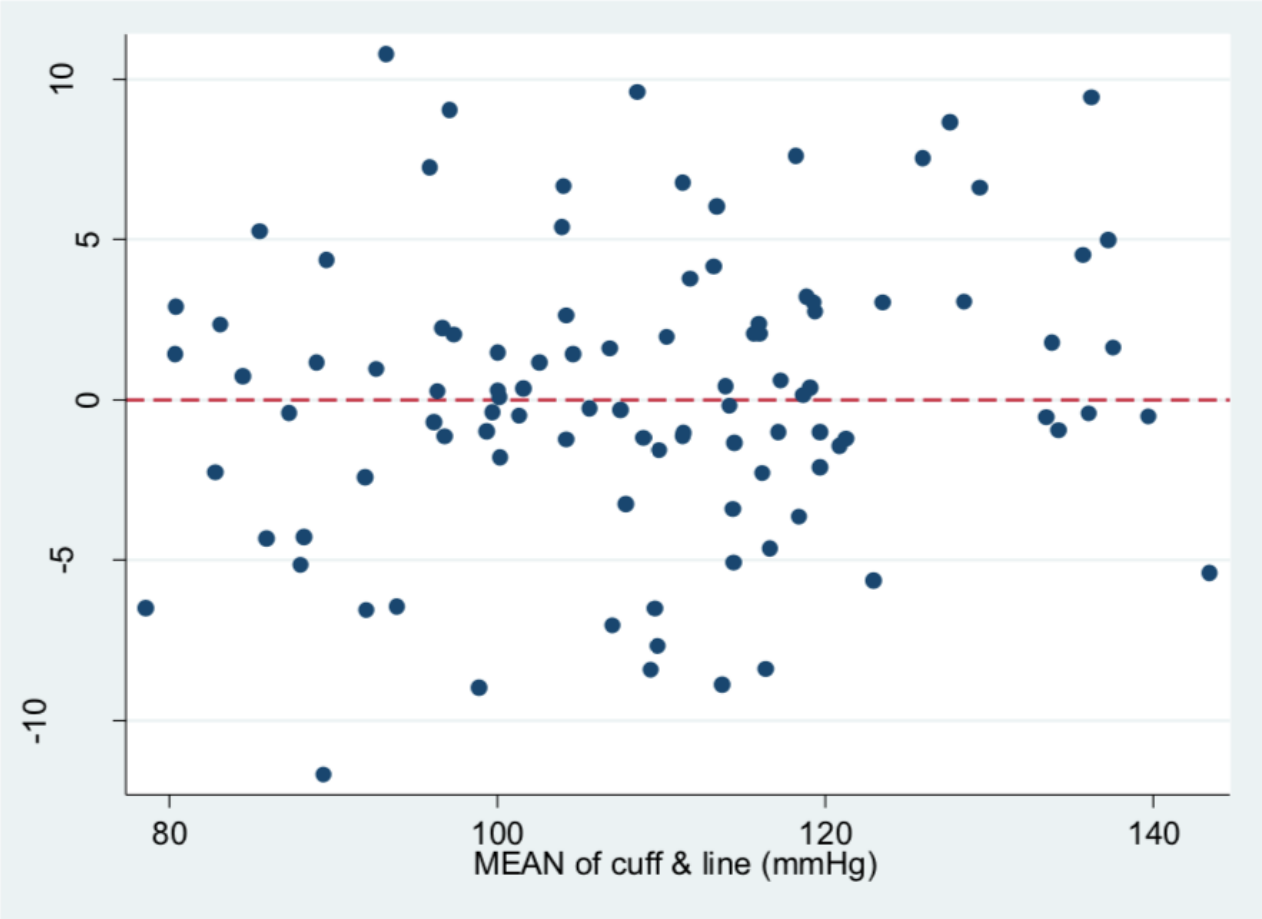
Quantifying validity – continuous variable

- Should look at differences vs mean

Patient	Cuff SBP (mmHg)	Arterial line SBP (mmHg)	DIFFERENCE (instrument – gold standard)	MEAN
1	103	105	-2	104
2	125	125	0	125
3	91	96	-5	93.5
4	136	132	+4	134
5	111	110	+1	110.5
6	110	112	-2	111
…	130	125	+5	127.5

Quantifying validity – continuous variable

- Plot the differences vs mean



Mean (sd) of differences = 0.17 (4.51) mmHg

See also Bland & Altman, Lancet 1986; 307-310

Quantifying validity – binary variable

Example: oral test for HIV

– how can we quantify its validity?

– i.e. how well it measures true HIV status

Quantifying validity – binary variable

Oral HIV Test

Western Blot (Gold Standard)

Measuring HIV status OraQuick vs Western blot

- Since we can measure the true HIV status, we can evaluate validity of oral test.

How?

- Test sample of people with gold standard test *and* instrument – cross tabulate

(Gold standard test)

Western Blot (Gold standard test)

	TRUE +	TRUE -	
Total	285	5472	

Measuring HIV status OraQuick vs Western blot

- Since we can measure the true HIV status, we can evaluate validity of oral test.

How?

- Test sample of people with gold standard test *and* instrument – cross tabulate

(Gold standard test)

Western Blot (Gold standard test)

OraQuick test

Instrument

	TRUE +	TRUE -	[redacted]
TEST +			[redacted]
TEST -			[redacted]
Total	285	5472	[redacted]

Measuring HIV status OraQuick vs Western blot

- Since we can measure the true HIV status, we can evaluate validity of oral test.

How?

- Test sample of people with gold standard test *and* instrument – cross tabulate

(Gold standard test)

Western Blot (Gold standard test)

OraQuick test

Instrument

	TRUE +	TRUE -	Total
TEST +	280	2	282
TEST -	5	5470	5475
Total	285	5472	5757

Measuring HIV status OraQuick vs Western blot

- **What percentage of genuinely HIV+ people are correctly identified by the test?**

(Gold standard test)

Western Blot (Gold standard test)

OraQuick test

Instrument

	TRUE +	TRUE -	Total
TEST +	280	2	282
TEST -	5	5470	5475
Total	285	5472	5757

Measuring HIV status OraQuick vs Western blot

- **What percentage of genuinely HIV+ people are correctly identified by the test?**

		Western Blot (Gold standard test)	
		TRUE +	TRUE -
OraQuick test	TEST +	280	2
	TEST -	5	5470
	Total	285	5472

Measuring HIV status OraQuick vs Western blot

- **What percentage of genuinely HIV+ people are correctly identified by the test?**

280/285 x 100 = 98.21% [*SENSITIVITY]*

		Western Blot (Gold standard test)	
		TRUE +	TRUE -
OraQuick test	TEST +	280	2
	TEST -	5	5470
	Total	285	5472

Measuring HIV status OraQuick vs Western blot

- **BUT what percentage of genuinely HIV- people are correctly identified by the test?**

		Western Blot (Gold standard test)	
		TRUE +	TRUE -
OraQuick test	TEST +	280	2
	TEST -	5	5470
	Total	285	5472

Measuring HIV status OraQuick vs Western blot

- **BUT what percentage of genuinely HIV- people are correctly identified by the test?**

5470/5472 x 100 = 99.96% [*SPECIFICITY*]

		Western Blot (Gold standard test)	
		TRUE +	**TRUE -**
OraQuick test	TEST +	280	2
	TEST -	5	5470
	Total	285	5472

Sensitivity and Specificity

- Quantify validity of an instrument measuring a binary quantity

Sensitivity

% *of those truly with the condition* that are identified correctly (e.g. test +ve) by the instrument

Specificity

% *of those truly free of the condition* that are identified correctly (e.g. test –ve) by the instrument

OraQuick HIV test

	TRUE +	TRUE -	[redacted]
TEST +	280	2	[redacted]
TEST -	5	5470	[redacted]
Total	285	5472	[redacted]

Sensitivity = 98.21%; Specificity = 99.96%

A particular individual receives a +ve test result. What is the probability they really have HIV?

POSITIVE PREDICTIVE VALUE (PPV)

= 280/282 x 100 = 99.3%

Positive Predictive Value

- Quantifies probability that an individual with a +ve test result really has the condition...
- Hence of interest in interpreting individual test results

Positive Predictive Value

- Quantifies probability that an individual with a +ve test result really has the condition...
- Hence of interest in interpreting individual test results
- Not in itself a useful measure of validity – depends on validity (sensitivity & specificity) AND
- underlying prevalence of condition
- same test can have very different PPV in different populations

Positive Predictive Value

	TRUE +	TRUE -	
TEST +	280	2	
TEST -	5	5470	
Total	285	5472	

Sensitivity = 98.21%; Specificity = 99.96%

Prevalence = 5%

PPV = 280/282 x 100 = 99.3%

Positive Predictive Value

	TRUE +	TRUE -	[redacted]
TEST +	~~280~~ 56	2	[redacted]
TEST -	~~5~~ 1	5470	[redacted]
Total	~~285~~ 57	5472	[redacted]

Sensitivity = 98.21%; Specificity = 99.96%

Prevalence = 5%

PPV = 280/282 x 100 = 99.3%

Prevalence = 1%

PPV = 56/58 = 96.5%

Reliability

- How **consistent** is the instrument?

a) When different observers make the same measurement (Int**er** – observer reliability)

b) When the same observer makes repeated measurements (Int**ra** – observer reliability)

- Sometimes called repeatability or reproducibility

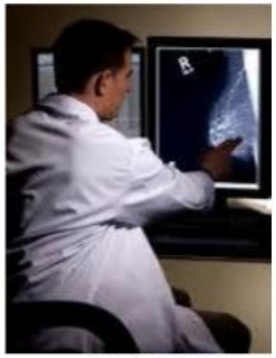
Measuring reliability

- We will focus on measures for *categorical* data

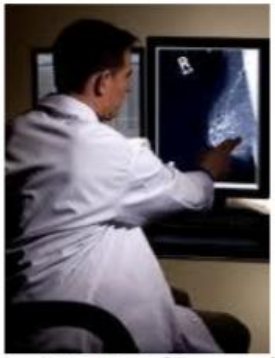
E.g. How *reliable* is mammogram for identifying breast cancer?

Measuring reliability

Intra-observer reliability:



BART

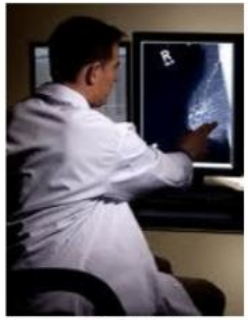


BART (later)

Patient	Cancer? (Attempt 1)	Cancer? (Attempt 2)
1	Yes	Yes
2	No	No
3	No	No
4	No	No
5	Yes	No
6	No	No
..	..	..

Measuring reliability

Inter-observer reliability:



BART

LISA

Patient	Cancer? (Bart)	Cancer? (Lisa)
1	Yes	Yes
2	No	No
3	No	No
4	No	Yes
5	Yes	Yes
6	No	No
..	..	..

Measuring reliability

- We can summarise the results in a frequency table

BART

LISA	Yes	No	Row total
Yes	7	2	9
No	3	84	87
Column total	10	86	96

Mean pair agreement /observed agreement (A) = (7+84)/96 = 0.95

Measuring reliability

BART

LISA	Yes	No	Row total
Yes	7	2	9
No	3	84	87
Column total	10	86	96

Mean pair agreement /observed agreement (A) = (7+84)/96 = 0.95

- BUT a proportion would agree by chance alone ..and this depends on the prevalence.
- The Kappa statistic gives a measure of agreement that takes into account level of agreement due to just chance.

Measuring reliability: Kappa

BART

LISA	Yes	No	Row total
Yes	7	2	9
No	3	84	87
Column total	10	86	96

Mean pair agreement (observed (A) = (7+84)/96 = 0.95

Expected in "yes/yes" cell by chance alone? (10 x 9)/ 96 = 1
Expected in "no/no" cell by chance alone? (86 x 87)/ 96 = 78
Expected pair agreements by chance alone (B) = 79/96 = 0.82

$$\text{Kappa } (\kappa) = \frac{\text{agreement} > \text{chance}}{\textit{max possible}\text{ agreement} > \text{chance}} = \frac{(\text{observed } (A) - \text{expected } (B))}{(1 - \text{expected } (B))} = \frac{(0.95 - 0.82)}{(1 - 0.82)} = 0.72$$

Interpreting Kappa

0 = no agreement better than chance

1 = perfect agreement

- No universal rules, but as a guideline, it is generally considered that:

<0.4 Reflects fairly poor agreement

0.4-0.75 Moderate to good

>0.75 Very good/excellent

Objectives

- 4) What are the consequences for epidemiological studies?

Non-differential misclassification

- For an <u>exposure</u> variable, this means misclassification is independent of outcome

E.g. Case-control study - same chance of misclassification for cases and controls

- In general, dilutes effects
  - Shifts odds ratio/rate ratio *towards the null*
- Can lead to incorrect conclusion of *lack of effect*

Non-differential misclassification

- For an <u>outcome</u> variable, this means misclassification is independent of exposure

E.g. Cohort study - same chance of misclassification for exposed and unexposed

- The impact depends on study type and direction of misclassification:

Case control study: dilutes odds ratio towards 1

Cohort study:

- If outcome is under-ascertained –rate ratio unbiased.
- If outcome is over-ascertained – rate ratio diluted towards 1 – Rate *difference* will be biased

Differential misclassification

- Misclassification of exposure is systematically different for those with vs those without outcome.

e.g. mesothelioma cancer cases more likely to recall exposure to asbestos than controls

- Misclassification of outcome is systematically different for exposed vs unexposed

e.g. High BP more likely in women on oral contraceptives than those not attending family planning clinics.

- Can go the other way as well. Impact is harder to predict.

Summary 1

- Measurement error occurs when our measured value differs from the true value (Categorical data: “misclassification”)

– Poor design/instructions/execution; limitations of participants; data/programming errors

- Important to try and quantify measurement errors and their impact.
- We discussed various tools and concepts that can help.

Summary 2

- **Validity:** how well does instrument measure what we are trying to measure; quantify (binary data) with sensitivity/specificity
- **Reliability:** how consistent is instrument in measuring same thing; quantify (categorical data) with Kappa statistic
- **Impacts** of measurement error can depend on whether **non- differential or differential**, as well as on study design, and type of variable (exposure/outcome)

Questions?