


PHC 6010
EPIDEMIOLOGY METHODS I


**Measurement and Classification
of Exposure**

Hamisu Salihu, MD, PhD




**Misclassification of
Exposure**

- Definition: the erroneous measurement of one or several categorical variables.
- Measurement of any subject characteristic is subject to potential error.
- In health sciences, measurement error is common because gold standards usually much more expensive than error-prone measurements.
- A major concern in health and behavioral sciences because of public health consequences




**Misclassification
bias**

- Errors in measurement are classified broadly into “random” and “systematic error”.
- **Random error**
 - It is the error that occurs due to natural variation in the measurement process.
 - It occurs by chance alone and is therefore unpredictable
 - Normal distribution with zero mean and constant variance
 - Example of the basketball throw^f
 - All things being equal there is equal probability that sometimes my shots deviate to the right and other times to the left. Whichever happens is a matter of chance



**Misclassification
bias**


- Hence, the presence of only random error will average the correct “true” value (i.e., deviations to the right and left will be equal and cancel out).
- Net effect: the average of the positive (right) and negative (left) deviations will equal zero. This is why random error is said to have a normal distribution with a mean of zero.



**Misclassification
bias**


Systematic errors:

- They represent biased circumstances that cause a series of measurements to be consistently either too high or too low
- Are reproducible inaccuracies that are consistently in the same direction
- E.g., the basketball experiment^f
- The error or deviation is now systematic and predictable.
- The probability of my shot going left or right is not equal. It will overwhelmingly be to the dominant right hand.




**Misclassification
bias**

- Another example is that of repeated measurement (e.g., finger measurement)
- Attempt 3 measurements by three different persons who are blinded.
- The measurement scores could be entirely different by chance (hence, the word random); e.g.
 - 6.3cm; 6.1cm; 6.2cm
 - The measurements tend to centralize to a certain value, in this case 6.2 cm. This is usually referred to as the average.
- ^fTo denote the degree of uncertainty (error or lack of precision) in our measurement, we usually write it thus 6.2± 0.1 cm. The estimate of our error in this case is ± 0.1.



Misclassification bias

- **Bias in information** leads to misclassification of exposure and/outcome status (maybe the interviewer or the responder that is the source of the bias). This is what leads to misclassification of study subjects.
- E.g., recall bias in case-control study (e.g., smoking and stillbirth)
- E.g., A positive outcome may be missed (e.g., association between exercise and silent MI).
- Or a pseudoevent may be mistakenly classified as a positive outcome (e.g., interpreting any history of chest pain in an elderly patient as CHD)
- Erroneous measurements usually yield biased estimates for both the estimation of marginal parameters (e.g., prevalence or incidence rates) and estimation of association.




Misclassification bias

	E	\hat{E}
D	7	18
ND	5	20


	E	\hat{E}
D	13	12
ND	15	10

- Consider the cohort study in the two tables.
- Table 1 = correct classification with incidence rate (a marginal parameter) of 0.58 and 0.47 for exposed and unexposed groups respectively.
- The relative risk (measure of association) = 1.23
- Table 2 (below) shows a situation in which misclassification of exposure status occurs, with the incidence rate (a marginal parameter) of 0.46 and 0.55 for exposed and unexposed groups respectively.
- The relative risk (measure of association) is now 0.83




Theoretical framework

- Assuming we want to estimate the degree of the association between two binary variables, S and Y.
- Each is coded as 0 (i.e., condition = absent) or 1 if the condition is present.
- S and Y denote the variable for the “true value” and \hat{S} and \hat{Y} represent the corresponding variables for the observed values or surrogates of S and Y that are error-prone.




Theoretical Framework

- Let S be an exposure variable or a potential risk factor that precedes Y.
- Let's use a prospective cohort design to establish a temporal ordering (namely, that S must necessarily precede Y).
- For example: S and Y may denote the presence of MDD (major depressive disorder) in the parents and offspring respectively.
- In other words, the hypothesis could be MDD in parents (exposure or S) is a risk factor for MDD in offspring (Y)




Theoretical framework

- The presence of the disorder in the relatives of affected probands could be analyzed with respect to the presence of the disorder in relatives of unaffected probands.
- Note: A proband is a person forming starting point for study of family or pedigree
- The amount of misclassification in \hat{S} , is expressed by the size of misclassification probabilities



Theoretical framework


- Imagine that \hat{S} represents an error-prone method (e.g., orally obtained family history) of deriving a diagnosis for MDD in the parents and that S represents an appropriate gold standard.
- Given the above, there are two misclassification probabilities in each variable:
 1. $P(\hat{S}=1|S=0)$: denotes the conditional probability that MDD in the parents is observed although there is actually no MDD (this is synonymous with a false positive rate)
 2. $P(\hat{S}=0|S=1)$: is the probability that there is apparently no MDD although the diagnostic criteria are actually met (this is synonymous with a false negative rate)



Theoretical framework


- This therefore means that the corresponding probabilities of measuring without error, namely:
 - $P(\hat{S}=1|S=1)$
 - $P(\hat{S}=0|S=0)$:

are the sensitivity and specificity, respectively, of the erroneous instrument as measures or surrogate of the “true” value or gold standard



Theoretical framework

- The question then arises: what are the various patterns of misclassification based on imperfect measurement of a categorical exposure and a categorical outcome.
- Using our earlier example, 3 patterns are discernable:
 - 1. The diagnosis of MDD in parents (S) could be subject to misclassification but not the diagnosis of the offspring (Y) or vice versa or both variables could be subject to misclassification




Theoretical framework

2. The misclassification (for example, in S) could depend on the value assumed by Y (namely, misclassification of S varies across Y=0 and Y=1). For instance the probability of a false-positive diagnosis among the parents might depend on the presence of a true diagnosis of MDD among the offspring, namely

- $P(\hat{S}=1|S=0, Y=0)$ might differ from
- $P(\hat{S}=1|S=0, Y=1)$


- This is classically called “differential misclassification” because erroneous assignment of exposure depends on disease status (cf with non-differential misclassification)



Theoretical Framework

3. The third type of misclassification: both variables (S and Y) are subject to misclassification, and the misclassification probabilities could be correlated or uncorrelated.

However, in this class, we will be concerned mainly with differential and non-differential misclassifications only.



Theoretical framework

Non-differential misclassification
 Is said to occur when the degree of misclassification of exposure is independent of case-control (disease) status. E.g., nuchal translucency test, BMI and subsequent stillbirth.

In this case the magnitude of association in terms of common measures like the risk ratio

$$RR = \frac{P(Y=1|S=1)}{P(Y=1|S=0)}$$

or the odds ratio is always biased toward the null value (1).

The degree of error in estimation is proportional to the magnitude of misclassification probabilities (and what are these?)⁷

Flegal, Browne and Haas (1986) have demonstrated that if the sum of sensitivity and specificity is less than 1 the direction of the risk ratio is reversed, thus, a risk ratio that is truly > 1 appears to be <1 and vice versa. We will examine this concept in subsequent lectures.