Epidemiology and practical research methods

Lecture 1

An idea or problem

A clear research question

Define objectives and hypotheses

Review of the relevant literature

Learn about End-Note

A valid methodology to address the question

Metrics of measurement

Data collection forms

Ethics proposal

Funding

Engaging others

A spread-sheet that reflects the data in the data collection form

Gather the data / conduct the study

Develop an analysis plan

Commence writing: intro / methods / dummy tables

Analysis and writing

Minor thesis / Publication

Epidemiology and operational research methods

- Basic epidemiology
- Types of studies
- Basic statistics – mean, median, incidence, prevalence, OR, RR
- How to come up with a research question
- Study design
- Choosing outcome measures that are valid
- Designing data collection tools
- Data analysis and data representation (tables, graphs)
- How to write a minor thesis / journal article

Epidemiology

- *Epi* – upon or around
- *demos* - people
- *logia* - study of

Types of epidemiology

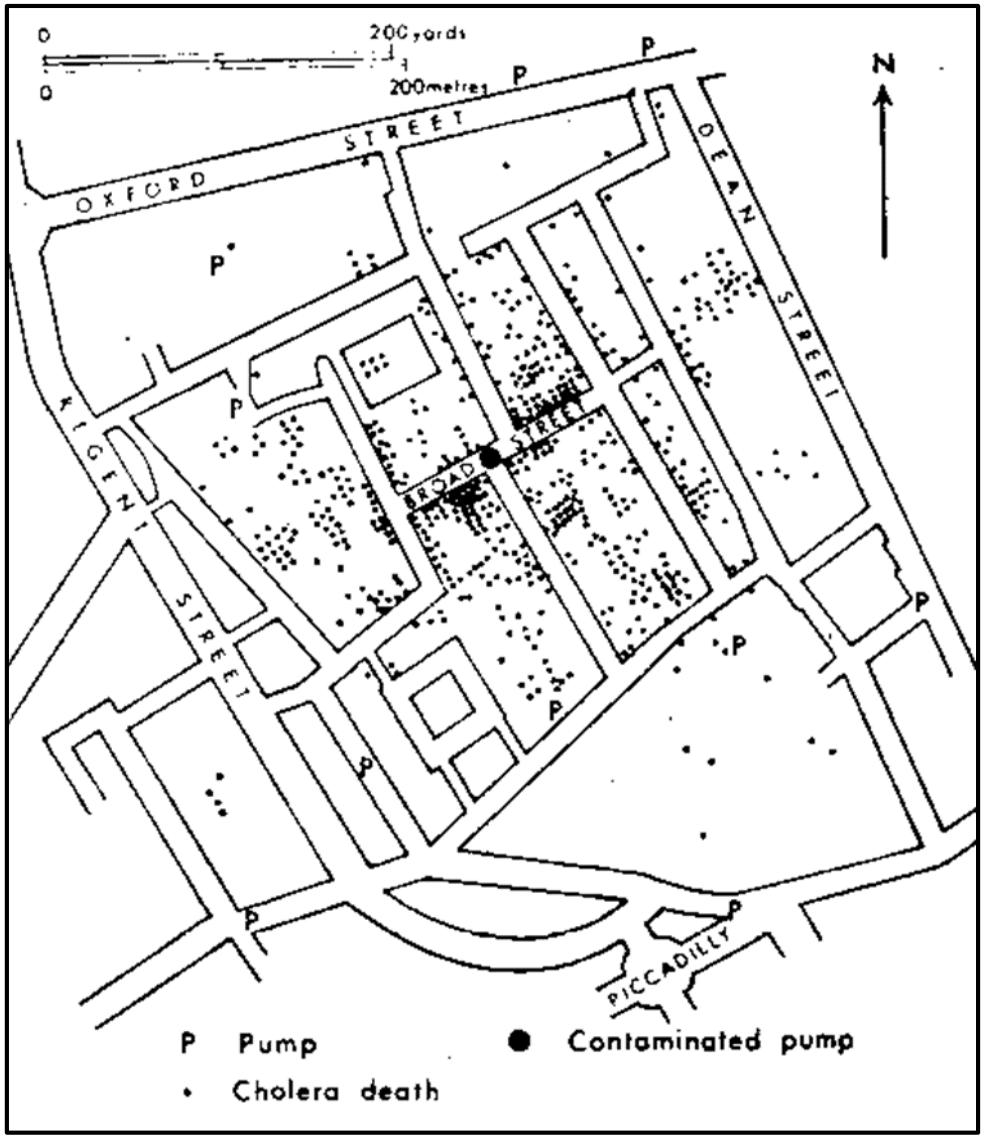
- Descriptive
  - Describing disease by time, place, person
  - Measuring the burden of disease
- Analytical
  - Looking for associations between exposures and outcomes, and between comorbidities and outcomes
- Interventional
  - Evaluating interventions
- Clinical
- Public health

19th Century England

- John Snow observed association between cholera deaths and source of water

Water supply company	Cholera death rate Per 1000 popn
Southwark	5.0
Lambeth	0.9

- Risk of death from cholera was over 5 times higher in people who used water from Southwark water supply (the Broadstreet pump)



Cholera 19th Century England

- Identified source of outbreak to be a water pump that had been contaminated by a broken sewer pipe nearby
- Removed the handle from the pump, ending the outbreak
- Thus identified cholera as a water-borne disease, even before the bacteria was isolated

Why learn epidemiology?

- Conduct your own research, make your own discoveries
- Use data to better understand your ward, hospital, district, province, country
- Understand as clinicians – are we doing a good job?

Basic terminology

- Proportions, rates and ratios
- Incidence and prevalence
- Means, medians, interquartile ranges, confidence intervals, z-scores

Ratios, proportions, and rates

- **Proportion** is a ratio in which the numerator **is included** in the denominator, e.g. the proportion of children with pneumonia who have severe pneumonia
  - Proportion has no unit as the unit of the numerator cancels out the unit of the denominator
- **Ratio** is one number divided by another number (numerator may or may not be included in denominator, e.g. Maternal Mortality Ratio)
- **Rate** is also a ratio
  - A rate usually has a time dimension. The unit is time or person-time to account for duration of time of follow-up (e.g. incidence rate of measles in an outbreak, infant mortality rate over a 5 year period)

Mortality measures

- Mortality
  - Population-based mortality (per 1000 live births)
  - Child mortality rate
  - Infant mortality rate
  - Neonatal mortality rate
  - Perinatal mortality rate
  - Still-birth rate
  - Maternal mortality ratio (per 100,000 live births)
- Health facility based: case fatality rate / proportion

Summary Sheet - SJNM -Kundiawa General Hospital

Search from: 1/1/2013 Search to: 31/12/2013

	Total Admissions	Total Deaths	Case Fatality Rate	
Total	1868	132	7.1	%
Pneumonia	404	24	5.9	%
Severe pneumonia	142	20	14.1	%
Diarrohoea (all cases)	373	14	3.8	%
Dysentry	54	6	11.1	%
Cholera	0	0	?	%
Typhoid	52	0	0.0	%
Malaria	26	1	3.8	%
Anaemia	155	37	23.9	%
Meningitis (all cases)	51	6	11.8	%
S. pneumonia meningitis	8	1	12.5	%
H influenzae meningitis	33	4	12.1	%
Neisseria meningitidis	4	0	0.0	%
Japanese encephalitis	0	0	?	%
Encephalitis other	1	0	0.0	%
Severe sepsis (not neonatal)	3	0	0.0	%
Dengue (all types)	0	0	?	%
Tuberculosis	191	22	11.5	%
Pulmonary TB	120	10	8.3	%
Extra pulmonary TB	71	12	16.9	%
HIV	21	6	28.6	%
Severe malnutrition	200	48	24.0	%
Iron deficiency	6	1	16.7	%
Vitamin A deficiency	0	0	?	%
Beriberi (Thiamine deficiency)	0	0	?	%
Vaccine preventable disease				
Measles	0	0	?	%
Whooping cough	4	0	0	%
Tetanus	3	1	33.3	%
Acute Flaccid Paralysis	2	1	50	%
Pigbel	15	1	6.7	%
Child protection	0	0	?	%
Cancer	9	1	11.1	%
ARF / Rheumatic heart disease	9	1	11.1	%
Neonatal	161	25	15.5	%
Neonatal sepsis	66	10	15.2	%
Birth asphyxia	71	8	11.3	%
Very Low Birth Weight 1000g - 1500g	24	15	62.5	%
Congenital heart disease	8	4	50.0	%

Total 1868 / 132 / 7.1%

Pneumonia 404 / 24 / 5.9%
Severe pneumonia 142 / 20 / 14.1%

Anaemia 155 / 37 / 23.9%

Tuberculosis 191 / 22 / 11.5%
PTB 120 / 10 / 8.3%
EPTB 71 / 12 / 16.9%

Very low birth weight 24 / 15 / 62.5%

Morbidity measures

- Prevalence (usually per 100,000 population, but can be %)
- Incidence (usually per 100,000 population *per year*)
- Hospital admissions / discharge
- Number of clinic consultations
- DALY (disability adjusted life years)
  - a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death
- QALY (Quality adjusted life years)
  - weigh each year of life by the perceived quality of that life, from one (perfect health) to zero (dead)

Other useful rates

- Treatment completion rates
- Adherence rates
- Event free rates (e.g. seizure free rate for children with epilepsy, 5-year relapse-free rates for children with leukaemia)
- Literacy rates

Disease frequency: Incidence and prevalence

- Prevalence - the number of people with the disease/outcome *at a given time*
- Incidence - the number of *new cases* of the disease/outcome over a specified time

Incidence and prevalence

- A chronic disease, such as diabetes, can have a low incidence but relatively high prevalence, because the disease is not usually fatal, but it cannot be completely cured either
  - Prevalence is the sum of new and existing cases from past years (prevalence increases as *new incident* cases are added each year)
- A short-duration, curable disease, such as the common cold, can have a high incidence but low prevalence, because many people get a cold each year, but virtually everyone is cured, so except in an outbreak season it will have a low prevalence cf incidence for the year

Incidence and prevalence

- Measuring cervical cancer in Province X, 2020
- Population at risk - the number of women living in Province X in 2020
- Prevalence - the number of existing cervical cancer cases in Province X in 2020
- Incidence - the number of *new cases* of cervical cancer diagnosed in Province X in 2020

Incidence and prevalence

- Choosing outcome metrics that are valid
- Precise description of who you consider to be a ‘case’; must be detailed and applied consistently
- Must include time, place and person
- For example, to be considered an *incident new case* of cervical cancer in Province X in 2020: A woman who resided in Province X during 2016 and was diagnosed in that year with cervical cancer
- Metrics often complicated but should be standardised – e.g. do you include *carcinoma in-situ*?

Incidence and prevalence

- Rheumatic heart disease: incidence or prevalence?
  - Acute rheumatic fever
  - Rheumatic heart disease

Example: TB incidence and prevalence

- “Passive” health facility-based screening – can estimate incidence
- But many people do not present to health facilities…
  - Until it is too late
  - Until they have transmitted TB to many other people
  - Because of geographical, educational or cultural issues
  - Because of inaccessibility to health facilities (or lack of confidence / trust)
- So incidence of TB at health facilities is not a good measure of population burden of disease…

Original Article

Active Community-Based Case Finding for Tuberculosis With Limited Resources: Estimating Prevalence in a Remote Area of Papua New Guinea

Asia Pacific Journal of Public Health
1–11

DOI: 10.1177/1010539516683497
journals.sagepub.com/home/aph
SAGE

- "Active" community-based screening – can identify population prevalence
- Research questions
  1. Can a simple model of active community-based screening be carried out in remote areas in PNG (i.e. is it feasible)?
  2. What is needed to achieve this (method, logistics, human resources, skills)?
  3. What is *the yield?*
     - Number of new TB cases found
     - What is the TB prevalence in the Etep Region?
  4. Can it be done at an affordable cost?
     - Cost of each new case identified

Bindu Karki[1], Guenter Kittel, MD[2], Ignatius Bolokon Jr, MBBS[2], and Trevor Duke, MD, FRACP[3,4]

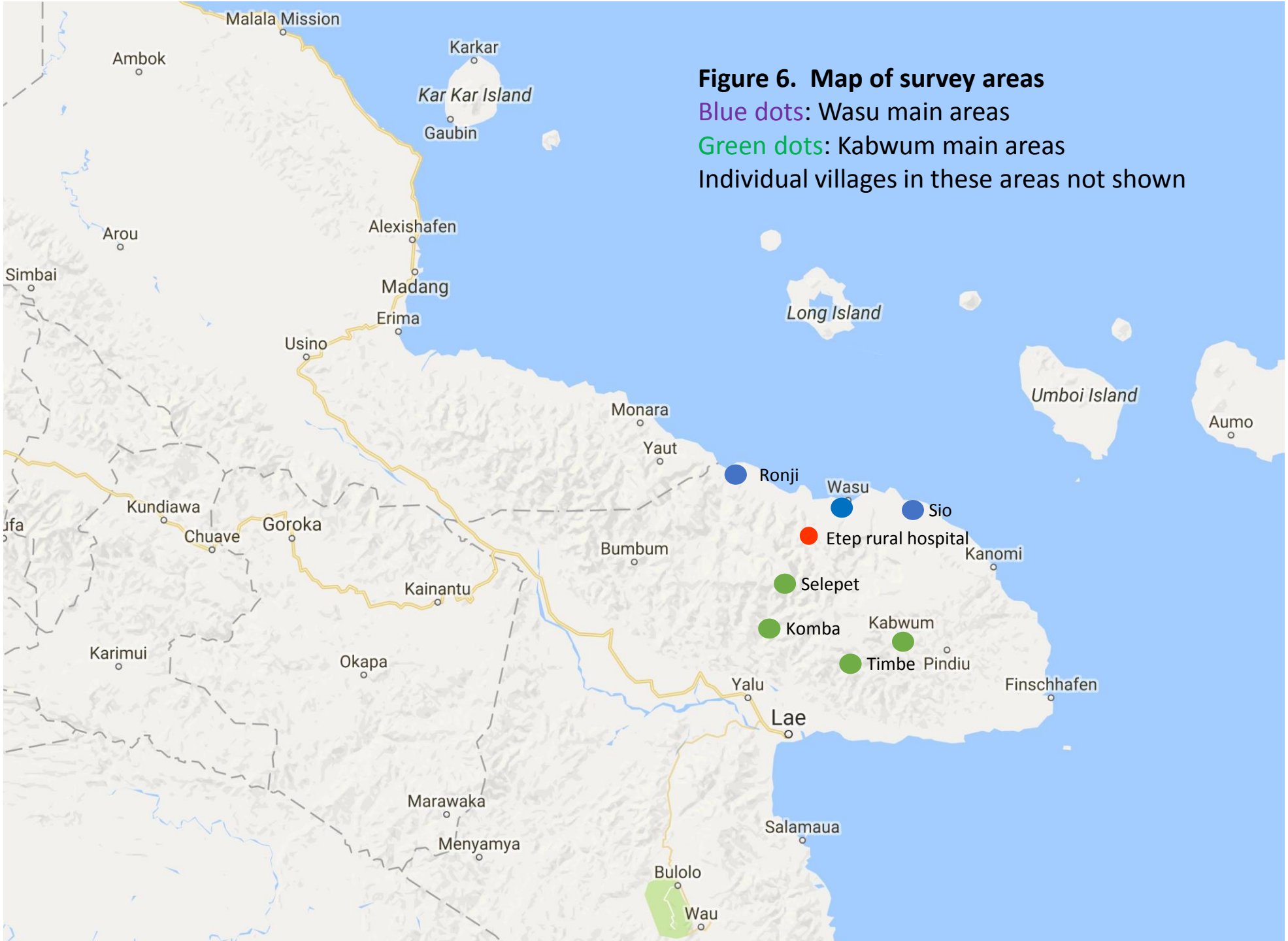
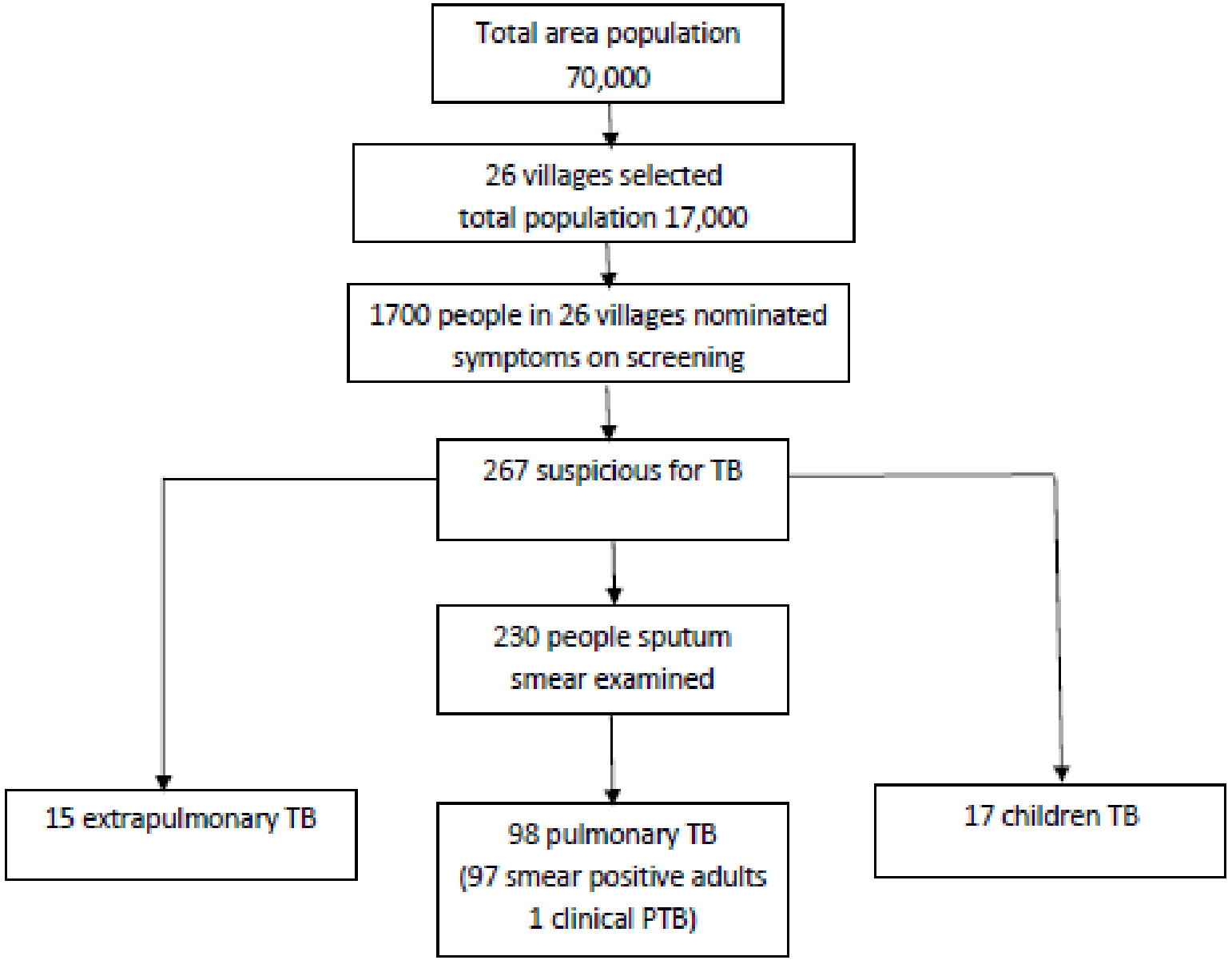


Figure 6. Map of survey areas

Blue dots: Wasu main areas

Green dots: Kabwum main areas

Individual villages in these areas not shown


Total area population
70,000
26 villages selected
total population 17,000
1700 people in 26 villages nominated
symptoms on screening
267 suspicious for TB
230 people sputum
smear examined
15 extrapulmonary TB
98 pulmonary TB
(97 smear positive adults
1 clinical PTB)
17 children TB

Results

- 98+15+17 = 130 people with TB (*yield* - numerical)
- Source population 17,000
- What is the prevalence?
  - population percentage
  - prevalence / 100,000 population

- Total cost K56,900
- Cost per case identified

Results

- 98+15+17 = 130 people with TB (yield - numerical)
- Source population 17,000
- What is the prevalence?
  - 130 / 17,000 x 100 = population % = 0.76%
  - 130 / 17,000 x 100,000 = prevalence / 100,000 population = 765 / 100,000

- Total cost K56,900
- Cost per case identified = 56900 / 130 = K438

Several types of prevalence

“Do you currently have asthma?”	Life-time cumulative prevalence?
“Have you had asthma during the last 2 years?”	Point prevalence?
“Have you ever had asthma?”	Period prevalence?

Several types of prevalence

“Do you currently have asthma?”	Point prevalence
“Have you had asthma during the last 2 years?”	Period prevalence
“Have you ever had asthma?”	Life-time cumulative prevalence

Spreadsheets – No!

Number	Name	Sex	Hospital number	Age	neonate	Diagnosis	Blood pressure	Weight	Cough duration	Outcome
1	b/georgina gauma	f		30 days	1	Sepsis, malnutrition	90/30	2.8kg	20	Survived
2	moses otto	m		2 months	no	Infection	85/42	2.9 kg	7 days	Discharged
3	davai kwalu	m	readmitted	123 months	no	SAM	95/45	21	1 week	Died
4	onnea leka	m	407379	22 days	1	Neonatal sepsis		3500 g	5days	DC
5	grace avae	f	readmitted	156months	no	Pneumonia, malnutrition		19	28 days	DC
6	b/o doreen frank	male		5 days	1	Sev Malnutrition, HIV		3	?	Survived
7	paul masiaresi	m	405922	4 months	no	LRTI		6.1	5 days	Absconded
8	jennifer john	f		24 months	no	Pneumonia	110/54	6.5kg	1 day	DC
9	joshua vaki	m	403745	2 months	no	Pneumonia – mod		4	6 days	Discharged
10	catherine george	f		7months	no	Malaria		6kg	4 days	Died
11	gabie vetali	m	404904	2 months	no	Pf positive		4.6	3 weeks	Died
12	B/O eunice morea	m		1 wk	1	HIV		2	?	Survived
13	b/o sharry yagena	female	404369	4 months	no	Pneumo – sev		4.8	1 mth	Survived
14	junior rex	m	readmitted	20 days	1	NNS		1500g	?	Died

Spreadsheets – Yes!

Number	Name	Sex	Hospital number	Age (months)	Neonate	Pneumonia	Malaria	HIV	Malnutrition	Sepsis	Systolic BP	Diastolic BP	Weight (kg)	Cough duration (days)	Outcome
1	b/georgina gauma	0	405643	1	1	0	0	0	1	1	90	30	2.8	20	1
2	moses otto	1	407643	2	0	0	0	0	0	1	85	42	2.9	7	1
3	davai kwalu	0	409876	123	0	0	0	0	1	0	95	45	21	7	0
4	onnea leka	1	407374	0.6	1	0	0	0	0	1			3.5	5	1
5	grace avae	0	405187	156	0	1	0	1	1	0			19	28	1
6	b/o doreen frank	1	407892	0.17	1	0	0	0	1	0			3		1
7	paul masiaresi	1	405922	4	0	1	0	0	0	0			6.1	5	
8	jennifer john	0	403456	24	0	1	0	0	0	0	110	54	6.5	1	1
9	joshua vaki	1	403745	2	0	1	0	0	0	0			4	6	1
10	catherine george	0	407685	7	0	0	1	0	0	0			6	4	0
11	gabie vetali	1	404904	2	0	0	1	0	0	0			4.6	21	0
12	B/O eunice morea	1	407623	0.25	1	0	0	1	0	0			2		1
13	b/o sharry yagena	0	404369	4	0	1	0	0	0	0			4.8	30	1
14	junior rex	1	401239	0.6	1	0	0	0	0	1			1.5		0

Mean, median
Confidence intervals
Case control studies
Odds ratios

Lecture 2

Mean and median

- **Mean** - used for symmetric numerical data (“normally distributed”).
  - Add all the values in a sample and divide by the number of values that are added.
  - The mean is affected by the extreme values in the dataset because it considers information from all patients and is appropriate for symmetric data.

  - Calculate the mean: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14

Mean and median

- **Mean** - used for symmetric (or approximately symmetric) numerical data (“normally distributed”).
  - Add all the values in a sample and divide by the number of values that are added.
  - The mean is affected by the extreme values in the dataset because it considers information from all patients and is appropriate for symmetric data.

  - Calculate the mean: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
  - 75/11 = 6.8

Mean and median

- **Mean** - used for symmetric (or approximately symmetric) numerical data ("normally distributed").
  - Add all the values in a sample and divide by the number of values that are added.
  - The mean is affected by the extreme values in the dataset because it considers information from all patients and is appropriate for symmetric data.

  - Calculate the mean: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
  - 75/11 = 6.8
  - Calculate the mean if one number extreme: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 44

Mean and median

- **Mean** - used for symmetric (or approximately symmetric) numerical data ("normally distributed").
  - Add all the values in a sample and divide by the number of values that are added.
  - The mean is affected by the extreme values in the dataset because it considers information from all patients and is appropriate for symmetric data.

  - Calculate the mean: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
  - 75/11 = 6.8
  - Calculate the mean if one number extreme: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 44
  - 105/11 = 9.5

Mean and median

- **The median** is for asymmetric ("non-normally distributed") numerical data.
- For symmetric data, mean and the median similar.
- If comparing summary statistics (averages) for multiple groups of subjects where some of the groups are *asymmetric*, median should be reported for each group.
- The median is that value which divides the data set into two equal parts.
- If the number of values is odd = median will be the middle value
- If the number of values is even= there is no single middle value. Instead there are two middle values – take the average of them.
    - Calculate the median: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14

Mean and median

- **The median** is for asymmetric ("non-normally distributed") numerical data.
- For symmetric data, mean and the median similar.
- If comparing summary statistics (averages) for multiple groups of subjects where some of the groups are *asymmetric*, median should be reported for each group.
- The median is that value which divides the data set into two equal parts.
- If the number of values is odd = median will be the middle value
- If the number of values is even= there is no single middle value. Instead there are two middle values – take the average of them.

  - Calculate the median: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
  - Median = 5

Mean and median

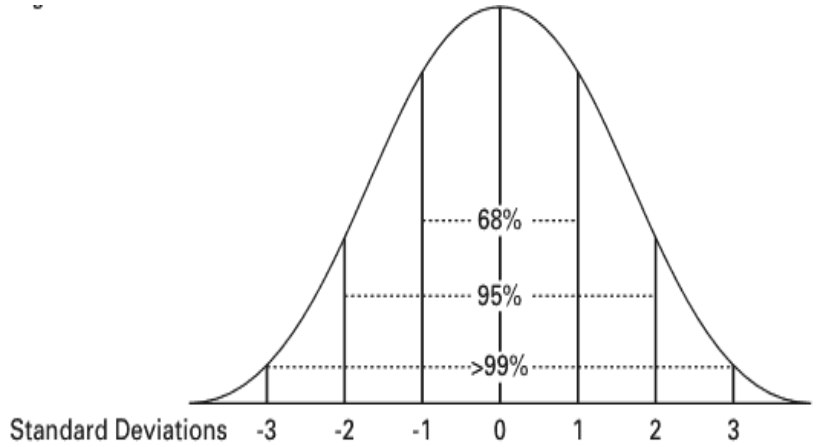
- **The median** is for asymmetric ("non-normally distributed") numerical data.
- For symmetric data, mean and the median similar.
- If comparing summary statistics (averages) for multiple groups of subjects where some of the groups are *asymmetric*, median should be reported for each group.
- The median is that value which divides the data set into two equal parts.
- If the number of values is odd = median will be the middle value
- If the number of values is even= there is no single middle value. Instead there are two middle values – take the average of them.
    - Calculate the median: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
    - Median = 5
    - Calculate the median if one number extreme: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 44

Mean and median

- **The median** is for asymmetric ("non-normally distributed") numerical data.
- For symmetric data, mean and the median similar.
- If comparing summary statistics (averages) for multiple groups of subjects where some of the groups are *asymmetric*, median should be reported for each group.
- The median is that value which divides the data set into two equal parts.
- If the number of values is odd = median will be the middle value
- If the number of values is even= there is no single middle value. Instead there are two middle values – take the average of them.
    - Calculate the median: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
    - Median = 5
    - Calculate the median if one number extreme: 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 44
    - Median = 5

Mean, median, range, interquartile range, confidence intervals

- 5, 8, 2, 12, 11, 14, 1, 4, 2, 2, 14
- Mean 6.8
- Median 5
- Need a *measure of spread or precision*
  - Mean - standard deviation
    - 68% of observations fall within the range (mean +- 1SD)
    - 95% of observations fall between mean +- 2SD
    - 99.7% of observations fall between mean +- 3SD

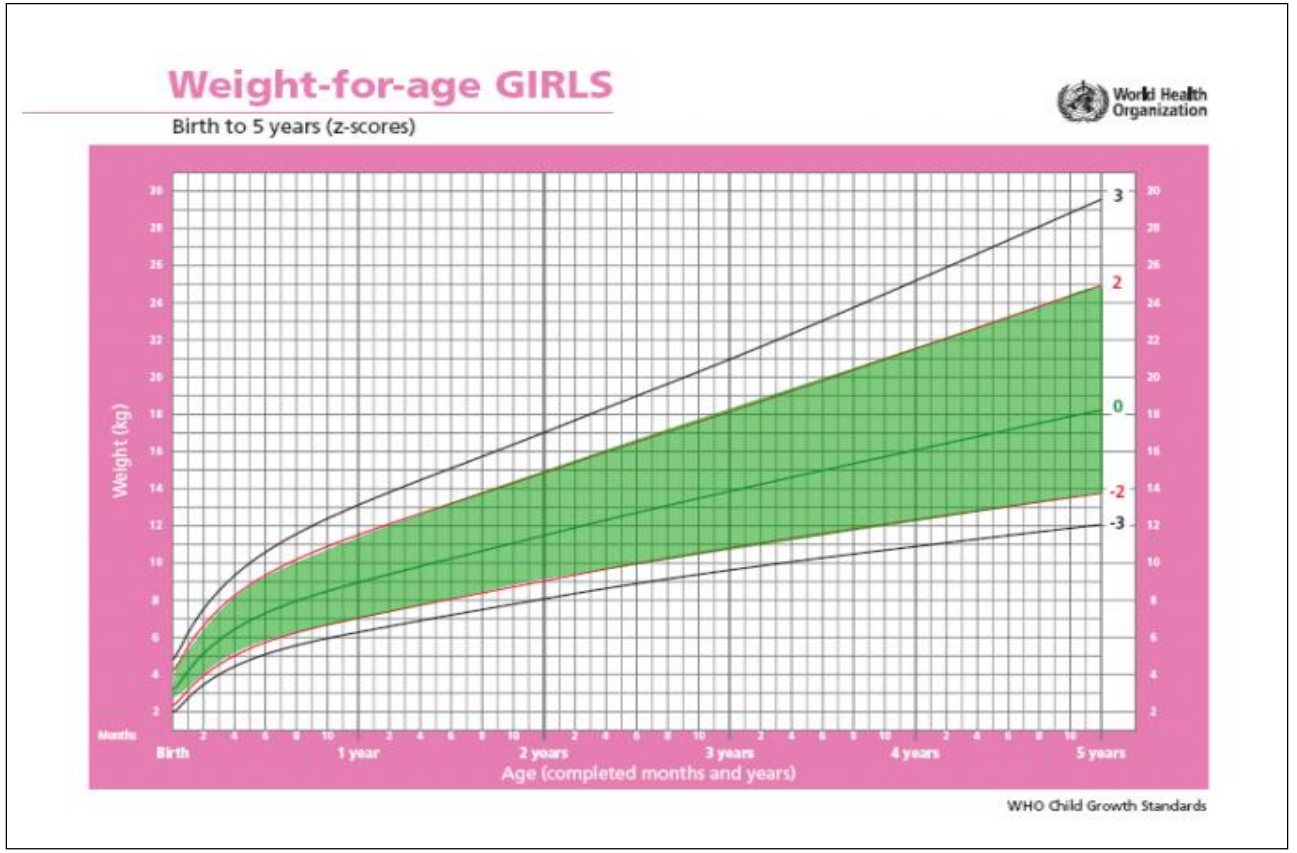


  - Median - “interquartile” range (middle 50% of the values; difference between the 25th percentile and the 75th percentile). Not affected by extreme values, so used in skewed / non-normally distributed data.

Summary

- If it is symmetric report the mean and SD
- If it is asymmetric report the median and IQR

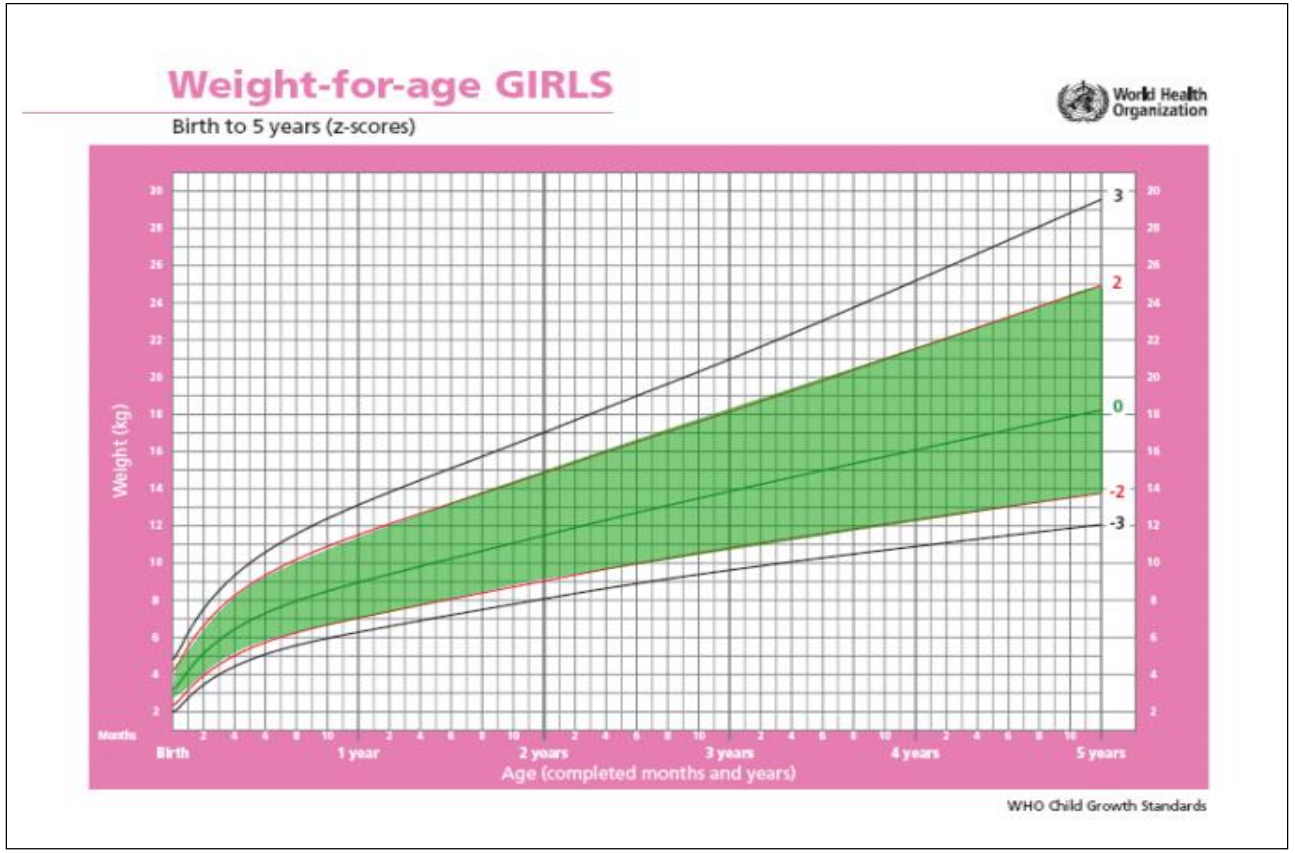
- Z-score


Weight-for-age GIRLS
Birth to 5 years (z-scores)
World Health Organization
Weight (kg)
3
2
0
-2
-3
Months
Birth
1 year
2 years
3 years
4 years
5 years
Age (completed months and years)
WHO Child Growth Standards

- Z-score = observed value – true mean

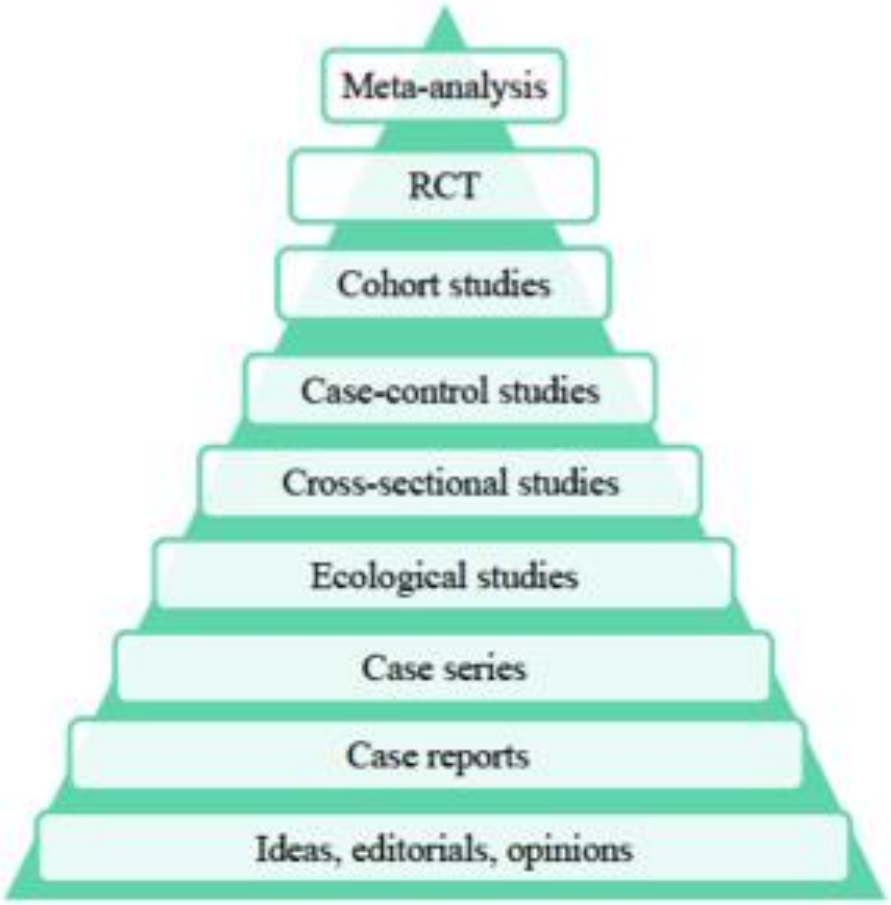
  ———————————————

  true standard deviation



Types of studies

- Observational
  - Case report / ecological observation
  - Case series / audit
  - Case-control
  - Cohort
- Experimental / Interventional
  - Controlled trial
    - Randomised controlled trial
    - Before-and-after design
    - Stepped wedge design
  - Field or community effectiveness trial
  - Operational research
- Meta-analysis

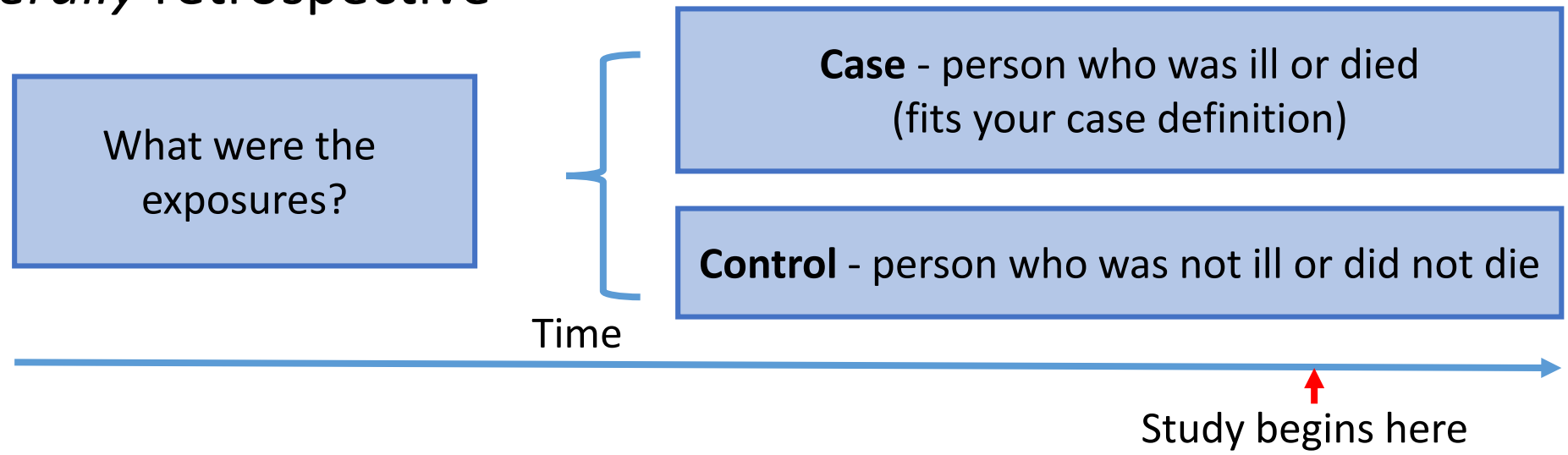


Case reports or case series

- What a clinician sees
- Unexpected observation in one or a series of patients, e.g. the first observation of a rare or previously unreported occurrence
- Can generate ideas for research or hypotheses
- Can communicate an important clinical lesson
- A single case can be misleading...
  - The exceptional case is not always generalizable
  - Cannot identify associations or risk factors or causation

Case control

- Group people on disease (outcome)
  - case has disease (meets ‘case definition’)
  - control does not have disease
  - look for differences in exposure between the groups (Odds ratio)
- *Generally* retrospective



Case control

- Control selection is crucial, should be from the same population:
  - Matching, e.g. age, date of birth, place, socioeconomic status, ethnicity
- Often some unknown confounding (as well as known confounding)
- Because retrospective: high probability of selection, measurement and recall biases
- Case control studies good for uncommon diseases (cf cohort studies which take a very long time if a disease is rare).
- Odds ratio (not relative risk)

Odds ratio

- The odds is the number of events / the number of non-events (similar but different to risk)
- Odds *Ratio* = odds of being exposed if you have the disease compared to the odds of being exposed if you don’t have the disease
- OR = 1, no association
- OR >>>1 = "those with the disease are more likely to have been exposed“
- OR <<<1 = "those with the disease are less likely to have been exposed“ exposure may be a protective factor in the causation of the disease
- 95% confidence intervals – do they overlap with 1?

Rosewell *et al. BMC Infectious Diseases* 2012, **12**:287
http://www.biomedcentral.com/1471-2334/12/287

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

Cholera risk factors, Papua New Guinea, 2010

Alexander Rosewell[1,2*], Benita Addy[3], Lucas Komnapi[3], Freda Makanda[3], Berry Ropa[4], Enoch Posanai[4], Samir Dutta[5], Glen Mola[6], WY Nicola Man[2], Anthony Zwi[2] and C Raina MacIntyre[2]

- First cases ever of cholera in PNG in July 2009
- 15,000 cases, case fatality proportion of 3.2%
- Case control study April – June 2010
- Confirmed case definition – suspected case with *V. cholerae* isolated in stool

Rosewell et al. BMC Infectious Diseases 2012, 12:287
http://www.biomedcentral.com/1471-2334/12/287

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

Cholera risk factors, Papua New Guinea, 2010

Alexander Rosewell[1,2*], Benita Addy[3], Lucas Komnapi[3], Freda Makanda[3], Berry Ropa[4], Enoch Posanai[4], Samir Dutta[5], Glen Mola[6], WY Nicola Man[2], Anthony Zwi[2] and C Raina MacIntyre[2]

Method

- Prospective
- Hospital-based (Angau)
- 3 controls per case interviewed within 48 hours of a case
- Controls had pneumonia or malaria (hospital admission register)
- Unmatched

Table 2 Cholera risk factors, Angau General Hospital, Papua New Guinea, 2010

Risk Factor	Cases n=54		Controls n=122		Odds ratio	Confidence interval	p value	Adjusted odds ratio	Confidence interval	p value
	n	(%)	n	(%)	(OR)	(95%CI)		(aOR)	(95%CI)	
Over 20 years of age	17	(32)	65	(53)	2.4	(1.2, 5.2)	0.007	2.7	(1.2, 5.8)	0.012
Resides in an informal settlement	44	(82)	70	(57)	3.2	(1.4, 7.9)	0.002	NA	NA	NA
River as drinking water source	25	(46)	31	(25)	2.5	(1.2, 5.2)	0.006	NA	NA	NA
Defecates in open air (or river)	13	(24)	5	(4)	7.4	(2.3, 27.9)	0.0001	4.6	(1.4, 14.9)	0.011
Has soap for hand washing at home	18	(33)	66	(54)	0.42	(0.20 0.87)	0.01	0.41	(0.19, 0.87)	0.021
Chews betel nut	30	(75)	43	(63)	1.74	(0.68, 4.67)	0.2	NA	NA	NA
Washes hands before eating	17	(32)	54	(49)	0.48	(0.23, 1.01)	0.035	NA	NA	NA
Knows case of cholera	16	(30)	11	(9)	4.3	(1.7, 11.0)	0.0005	2.4	(0.9, 6.2)	0.075
Attended funeral	5	(9)	12	(10)	0.93	(0.24, 3.02)	0.9	NA	NA	NA
Knows someone who travelled to cholera area	47	(87)	68	(56)	5.3	(2.2, 15.0)	0.0001	4.5	(1.8, 11.7)	0.002
Shares housing with diarrhoea case	11	(20)	6	(5)	5.0	(1.6, 17.2)	0.001	NA	NA	NA

Odds ratio calculation

		Disease (cholera)		
		Cases (n=54)	Controls (n=122)	Total
Exposure: Open defecation	Open defecation	13 (24%) **a**	5 (4%) **b**	18
	No open defecation (unspecified)	41 (76%) **c**	117 (96%) **d**	158
	Total	54	122	176

OR (the ratio of 2 odds)

= (a/b) / (c/d)

= ad / bc

= (13 x 117) / (41 x 5)

= 1521 / 205 = 7.4

Interpretation: "people who had cholera had 7 times the odds of practicing open defecation than those who did not get cholera"

Table 2 Cholera risk factors, Angau General Hospital, Papua New Guinea, 2010

Risk Factor	Cases n=54		Controls n=122		Odds ratio	Confidence interval	p value	Adjusted odds ratio	Confidence interval	p value
	n	(%)	n	(%)	(OR)	(95%CI)		(aOR)	(95%CI)	
Over 20 years of age	17	(32)	65	(53)	2.4	(1.2, 5.2)	0.007	2.7	(1.2, 5.8)	0.012
Resides in an informal settlement	44	(82)	70	(57)	3.2	(1.4, 7.9)	0.002	NA	NA	NA
River as drinking water source	25	(46)	31	(25)	2.5	(1.2, 5.2)	0.006	NA	NA	NA
Defecates in open air (or river)	13	(24)	5	(4)	7.4	(2.3, 27.9)	0.0001	4.6	(1.4, 14.9)	0.011
Has soap for hand washing at home	18	(33)	66	(54)	0.42	(0.20 0.87)	0.01	0.41	(0.19, 0.87)	0.021
Chews betel nut	30	(75)	43	(63)	1.74	(0.68, 4.67)	0.2	NA	NA	NA
Washes hands before eating	17	(32)	54	(49)	0.48	(0.23, 1.01)	0.035	NA	NA	NA
Knows case of cholera	16	(30)	11	(9)	4.3	(1.7, 11.0)	0.0005	2.4	(0.9, 6.2)	0.075
Attended funeral	5	(9)	12	(10)	0.93	(0.24, 3.02)	0.9	NA	NA	NA
Knows someone who travelled to cholera area	47	(87)	68	(56)	5.3	(2.2, 15.0)	0.0001	4.5	(1.8, 11.7)	0.002
Shares housing with diarrhoea case	11	(20)	6	(5)	5.0	(1.6, 17.2)	0.001	NA	NA	NA

Odds ratio calculation

		Disease (cholera)		
		Cases (n=54)	Controls (n=122)	Total
Exposure: **Soap for handwashing at home**	Soap	**a**	**b**	
	No soap	**c**	**d**	
	Total			

OR (the ratio of 2 odds)

= (a/b) / (c/d)

= ad / bc

= Interpretation -

Odds ratio calculation

		Disease (cholera)		
		Cases (n=54)	Controls (n=122)	Total
Exposure: **Soap for handwashing at home**	Soap	18 **a**	66 **b**	84
	No soap	36 **c**	56 **d**	92
	Total	54	122	176

OR (the ratio of 2 odds)

= (a/b) / (c/d)

= ad / bc

= Interpretation –

Odds ratio calculation

		Disease (cholera)		
		Cases (n=54)	Controls (n=122)	Total
Exposure: **Soap for handwashing at home**	Soap	18 **a**	66 **b**	84
	No soap	36 **c**	56 **d**	92
	Total	54	122	176

OR (the ratio of 2 odds)

= (a/b) / (c/d)

= ad / bc

= (18 x 56) / (66 x 36) = 1008 / 2376

= 0.42

Interpretation – ??

Odds ratio calculation

		Disease (cholera)		
		Cases (n=54)	Controls (n=122)	Total
Exposure: **Soap for handwashing at home**	Soap	18 **a**	66 **b**	84
	No soap	36 **c**	56 **d**	92
	Total	54	122	176

OR (the ratio of 2 odds)
= (a/b) / (c/d)
= ad / bc
= (18 x 56) / (66 x 36) = 1008 / 2376
= 0.42

Interpretation – "people with cholera were 58% less likely to have soap at home for handwashing." Handwashing with soap and water protects against cholera

Odds ratio – 3 more concepts

- **Confidence intervals**
  - CI indicates the level of uncertainty around the measure of effect, in this case OR (precision of the OR estimate).
    - Takes account of sample size: small studies, wide CI; large studies, narrow CI for a given true effect size.
    - 95% CI means the true population effect is 95% likely to lie between these two points
- "**Adjusted Odds ratio**"
  - Multi-variable analysis compares several variables that may be associated with or predictive of a certain outcome.
    - Takes into account confounding
    - Allows the minimum number of predictive variables to be identified
- P-value
  - The probability that the true population estimate falls *outside* the 95% CI
  - *Not* precise, better to use OR (95% CI)

Table 2 Cholera risk factors, Angau General Hospital, Papua New Guinea, 2010

Risk Factor	Cases n=54		Controls n=122		Odds ratio	Confidence interval	p value	Adjusted odds ratio	Confidence interval	p value
	n	(%)	n	(%)	(OR)	(95%CI)		(aOR)	(95%CI)	
Over 20 years of age	17	(32)	65	(53)	2.4	(1.2, 5.2)	0.007	2.7	(1.2, 5.8)	0.012
Resides in an informal settlement	44	(82)	70	(57)	3.2	(1.4, 7.9)	0.002	NA	NA	NA
River as drinking water source	25	(46)	31	(25)	2.5	(1.2, 5.2)	0.006	NA	NA	NA
Defecates in open air (or river)	13	(24)	5	(4)	7.4	(2.3, 27.9)	0.0001	4.6	(1.4, 14.9)	0.011
Has soap for hand washing at home	18	(33)	66	(54)	0.42	(0.20 0.87)	0.01	0.41	(0.19, 0.87)	0.021
Chews betel nut	30	(75)	43	(63)	1.74	(0.68, 4.67)	0.2	NA	NA	NA
Washes hands before eating	17	(32)	54	(49)	0.48	(0.23, 1.01)	0.035	NA	NA	NA
Knows case of cholera	16	(30)	11	(9)	4.3	(1.7, 11.0)	0.0005	2.4	(0.9, 6.2)	0.075
Attended funeral	5	(9)	12	(10)	0.93	(0.24, 3.02)	0.9	NA	NA	NA
Knows someone who travelled to cholera area	47	(87)	68	(56)	5.3	(2.2, 15.0)	0.0001	4.5	(1.8, 11.7)	0.002
Shares housing with diarrhoea case	11	(20)	6	(5)	5.0	(1.6, 17.2)	0.001	NA	NA	NA

“Dummy tables” – draft them early…

Characteristic	Total n=
Male / Female	
Age in months: median (IQR)	
Duration of cough in days: median (IQR)	
Temperature ≥38 C, n (%)	
Apnea, n (%)	
Poor feeding, n (%)	
Severe chest in drawing, n (%)	
Tracheal tugging, n (%)	
Heart rate, median (IQR)	
Oxygen saturation %, median (IQR)	
SpO_2 <85%, n (%)	
Chest x-ray done, n (%)	
Radiographic signs, present, n (%)	
Radiographic signs, absent, n (%)	

Table 1: Clinical characteristics at enrolment

Cohort studies
Randomised trials
Relative risk
Bias and confounding

Lecture 3

Cohort studies

- Cohort: “a group of people with a shared characteristic”
- Cohort studies can be observational or intervention studies
- Detailed longitudinal recording of data

Cohort studies

- Involves follow-up of people with a common characteristic: and comparison of outcomes by exposure to a possible risk factor(s).
- Direction of study is always *forward* in time (after the exposure), whether the study is prospective or retrospective
- The incidence of an outcome is determined, and compared between those exposed and those not exposed to a risk factor during the study time
- Provides good evidence of cause and effect relationship

Types of cohorts

- Birth cohort
- Age cohort – “7-Up”, “adolescent cohort”
- School class cohort
- Professional group cohort
- Disease cohort, e.g. a cohort of children with epilepsy, or HIV…
- Social group cohort, e.g. a cohort of adopted children…

Cohort studies

- Advantages
  - Describe the varied influences on a group of people over time, and their effects
  - Can explore multi-dimensional effects, such as biological, social, economic, educational influences on disease and other outcomes
  - Disease cohort can describe the natural history of a condition over time, and how it is influenced by treatment and other factors (social, environmental)
  - Describe the temporal sequence between cause and outcome
  - Identify the incidence (within that cohort)

Cohort studies

- Limitations:
  - loss to follow up common (especially the longer a study goes on, and if routine data used)
  - time consuming (longitudinal)
  - sometimes insufficient numbers to study the cause of rare diseases (e.g. IM vitamin K and childhood leukaemia).

Examples of observational cohort studies

- Bradford-Hill – 40,000 British doctors from 1951-2001
- BT20 Birth to 20 study (“Mandela's children”) in South Africa – 3000 births (1990)
- Nurses health study – UK 120,000 women, cardiovascular risk
- Dunedin Multidisciplinary Health and Development Study – 1000 births

In PNG?

- Longitudinal study of a cohort of children with epilepsy, looking at risk factors for death / poor control. Or protective factors for good control?
- Longitudinal follow-up study of a cohort of low birth weight babies, looking at risk factors for developmental delay. Or protective factors for normal development?
- Cohort study of children with HIV – from birth to adolescence.

Relative risk

- Relative Risk or Risk Ratio

  Risk in exposed / Risk in unexposed =

  a / (a + b)

  ———————

  c / (c + d)

		Disease / outcome		
		Disease	No disease	Total
Exposure:	Exposed	**a**	**b**	
	Unexposed	**c**	**d**	
	Total			

  The RR takes into account prevalence

  The OR and the RR are very similar if the prevalence of the outcome is low (for rare outcomes). Where the outcome is common (>10%) the OR over-estimates the RR.

Reducing hospital-acquired infections and improving the rational use of antibiotics in a developing country: an effectiveness study

Kartika Ita, et al Archives Dis Child 2014.

- I**ntervention study** of two cohorts: before and after introduction of a multi-faceted intervention to reduce nosocomial infections in Indonesia
  - Hand hygiene
  - Antibiotic stewardship
  - Guidelines for aseptic procedures
- In this case the “exposure” was an intervention, a better way of doing a certain thing
- Relative risk is a valid measure of the effect of the exposure, as the study follows 2 cohorts *prospectively* (which means the incidence of nosocomial infection can be defined by the study).

Relative risk calculation

a / (a + b)

_____

c / (c + d)

RR =

Interpretation:

		Disease (nosocomial infection)		
		Nosocomial infection	No nosocomial infections (n=122)	Total
Exposure: Package of intervention to reduce nosocomial infections	Intervention-era "exposed"	123 **a**	1296 **b**	1419
	Before interventions "unexposed"	277 **c**	950 **d**	1227
	Total	400	2246	2646

Relative risk calculation

$$\frac{a / (a + b)}{c / (c + d)}$$

$$\frac{123 / (123 + 1296)}{277 / (277 + 950)}$$

0.086680 / 0.225755

RR = 0.38

		Disease (nosocomial infection)		
		Nosocomial infection	No nosocomial infections (n=122)	Total
Exposure: Package of intervention to reduce nosocomial infections	Intervention-era "exposed"	123 **a**	1296 **b**	1419
	Before interventions "unexposed"	277 **c**	950 **d**	1227
	Total	400	2246	2646

Interpretation: "those who were exposed to multi-faceted intervention to prevent nosocomial infection (hand hygiene, antibiotic guidelines) had a RR of infection of 0.38 (or 38%)"
Relative risk reduction of 62%.

Risk factors and causation

- Causation: something that either alone or in combination with another factor results in disease. Often multi-factorial
- Attributable fraction: quantify the likely preventive impact of eliminating a specific causal factor

Case control and cohort studies

- Can identify *associations*
- Rules for evidence of *causation* (Bradford Hill):
  - Temporal relationship: cause must precede effect
  - Plausibility: consistent with other knowledge (but other evidence may just be lacking)
  - Consistency / reproducibility : several studies give the same finding
  - Strength: a weak relationship does not mean a factor is not casual
  - Dose-response: increased exposure increases your risk
  - Reversibility: does not always apply

- Is there an association between a possible cause and an effect?
  - Could it be due to bias?
  - Could it be due to confounding?
  - Could it be the result of chance?
  - Is the relationship casual?

“Infectious meningitis in Japan”

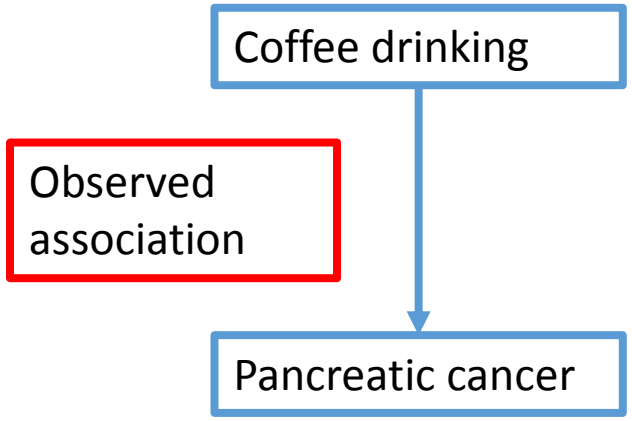
- Encephalopathy and deaths thought to be infectious meningitis…
- Epidemiological associations and proof of causation:
  - Most sufferers were found to reside close to Minamata Bay
  - Affected people were mostly from families involved in fishing trade
  - Those ingesting only small quantities of the fish did not get sick (dose effect)
  - Mercury found in fish (biological plausibility based on previous known information)
  - Identified as methyl-mercury poisoning…

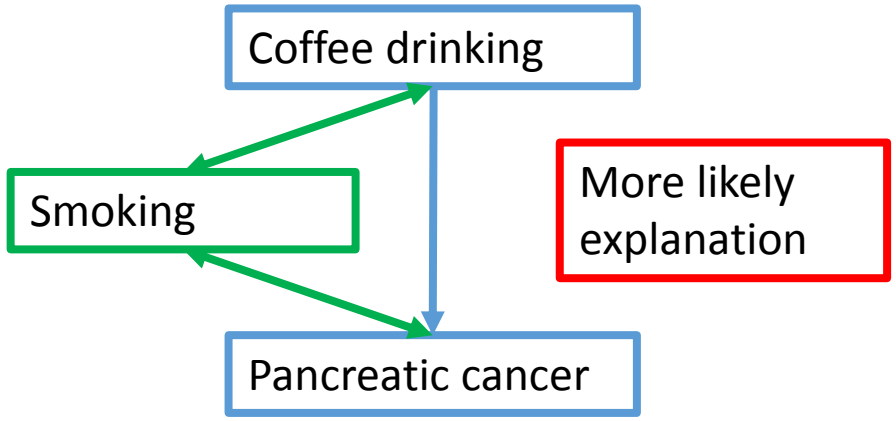
Bias

- The difference between results and population value due to incorrect measurements being taken or measurements being taken on a non-representative sample
  - Selection bias: systematic difference between the baseline characteristic of the groups compared
  - Measurement bias: a systematic error in the measurement of information on the exposure or outcome, sometimes called ascertainment bias
  - Responder/recall bias: a systematic error caused by differences in the accuracy or completeness of the recollections retrieved by study participants regarding events or experiences from the past

Confounding

- Situation in which a non-casual association between a given association is observed due to the influence of a third variable
  - Bias creates an association that is not true
  - Confounding describes an association that is true, but potentially misleading





How to control for confounding

- Design stage:
  - Randomisation: equal distribution of groups
  - Matching: match for age, sex, social class, other potential confounders in a case control study
- Analysis stage:
  - Stratification: tables of exposure vs outcome, one for each level or type of confounder
  - Statistical adjustment: can adjust for multiple factors

Randomised controlled trial

- Gold Standard for attributable risk or benefit of any intervention:
  - A new drug
  - A new type of surgical procedure
  - A complex intervention: such as a protocol of management for severe malnutrition, or a multi-faceted intervention to reduce nosocomial sepsis
  - A community-based intervention: cash transfers for completed immunisation, a school nutrition program

Randomised controlled trial

- Eliminates bias and confounding
- Measures the incidence of an outcome
- However...
  - Need to be evaluated for quality and relevance
    - Validity?
    - Applicability?
    - Efficacy vs effectiveness?
    - Sustainability?

Randomised controlled trial: PICOT

- Population
  - In children with disease X (or at risk of disease X)
- Intervention
  - Does treatment with Y…
- Comparator
  - Compared with Gold Standard…
- Outcome
  - Improve predefined outcome…
- Time
  - Over a predefined time period...

Types of RCTs

- Open: everyone involved knows which intervention is given to each patient
- Single-blind: one group of individuals does not know the identity of the intervention given to participants
- Double-blind: two groups of individuals do not know the identity of the intervention given to the participants. Performance and detection bias are minimised.

Randomised controlled trial

- Advantages:
  - Less risk of bias and confounding than any other epidemiological study
  - Provide strong evidence of causal relationships
  - Can be used to study multiple outcomes
  - Measures the incidence rate of an outcome
- Limitations:
  - Expensive
  - Long follow up period
  - Ethical issues
  - Outcomes must be measureable

Randomised controlled trial

- *Average* treatment effects for one group might not apply to another group, or even to subgroups, or individuals
- RCTs don't necessarily tell you *how* it works, or in *what context* it works

Randomised controlled trial: PICOT

- Population
  - In children with disease X (or at risk of disease X)
- Intervention
  - Does treatment with Y…
- Comparator
  - Compared with Gold Standard…
- Outcome
  - Improve predefined outcome…
- Time
  - Over a predefined time period...

Treatment of acute seizures: an RCT

J Child Neurol. 2014 Jul;29(7):895-902

Efficacy of sublingual lorazepam versus intrarectal diazepam for prolonged convulsions in Sub-Saharan Africa.

- Trial in paediatric emergency departments of 9 hospitals.
- 436 children aged 5 months to 10 years with convulsions persisting for more than 5 minutes assigned to receive intra-rectal diazepam (0.5 mg/kg, n = 202) or sublingual lorazepam (0.1 mg/kg, n = 234)
- Cessation of seizures within 10 minutes
- **Sublingual lorazepam 56% vs Intra-rectal diazepam in 79%**
- Probability of treatment failure higher with sublingual lorazepam (OR = 2.95, 95% CI = 1.91-4.55, $p<0.001$)
- Sublingual lorazepam is less effective in stopping paediatric seizures than intra-rectal diazepam, and intra-rectal diazepam should thus be preferred as a first-line medication in this setting.

Randomised controlled trial: PICOT

- Population
  - In children aged 5 months to 10 years with convulsions persisting for more than 5 minutes
- Intervention
  - Does treatment with lorazepam
- Comparator
  - Compared with intra-rectal diazepam
- Outcome
  - Increase the probability of cessation of seizures (over 10 minutes)
  - (Increase the probability of treatment failure: persistence of seizures longer than 10 minutes)
- Time
  - Over 10 minutes…

Precision of diagnostic tests
Sensitivity / specificity, PPV, NPV
Screening tests
Quality improvement research

Lecture 4

Assessment of precision of diagnostic measures

- Sensitivity: proportion with the disease who test positive
- Specificity: proportion without the disease who test negative
- Positive predictive value: proportion with a positive test who have the disease
- Negative predictive value: proportion with a negative test who do not have the disease

Validity and Reliability of Clinical Signs in the Diagnosis of Dehydration in Children

Marc H. Gorelick, MD, MS*§; Kathy N. Shaw, MD, MS*§; and Kathleen O. Murphy, RN, MSN‡

- 186 children with diarrhoea, vomiting and poor oral intake
- All children evaluated for 10 clinical signs before treatment
- Fluid deficit determined by serial weight gain after treatment (Gold Standard *)
- 63 children had dehydration (5% or greater body weight)
- Individual signs had low SENSITIVITY and high SPECIFICITY
- 4 clinical signs predicted diarrhoea as well as all others
  - Capillary refill >2 seconds
  - Absent tears
  - Dry mucous membranes
  - Ill general appearance

* Validated during the study with pre- and post-illness weights in 19 children – Fig 1.

	Disease positive	Disease negative	Totals
Test positive	a	b	a+b
Test negative	c	d	c+d
Totals	a+c	b+d	a+b+c+d

- Sensitivity= a/(a+c) [proportion with the disease who test positive]
- Specificity= d/(b+d) [proportion *without* the disease who test negative]
- Positive predictive value= a/(a+b) [proportion with a positive test who have the disease]
- Negative predictive value=d/(c+d) [proportion with a negative test who do not have the disease]

	Dehydration >5%	No dehydration (<5%)	Totals
Capillary refill >2 sec	30 **a**	5 **b**	35
Capillary refil <2 sec	33 **c**	118 **d**	151
Totals	63	123	186

- Sensitivity= a/(a+c)
- Specificity= d/(b+d)
- Positive predictive value= a/(a+b)
- Negative predictive value=d/(c+d)

	Dehydration >5%	No dehydration (<5%)	Totals
Capillary refill >2 sec	30	5	35
Capillary refil <2 sec	33	118	151
Totals	63	123	186

- Sensitivity= 30/(30 + 33) = 0.48
- Specificity= 118/(5 + 118) = 0.96
- Positive predictive value= 30/(30 + 5) = 0.86
- Negative predictive value= 118/(33 + 118) = 0.78

- Sensitivity and specificity are unchanged by prevalence of disease
- PPV and NPV *do* change with prevalence
  - As the prevalence increases, the PPV of a test increases, and the NPV decreases. To understand this, see:
  - https://www.youtube.com/watch?v=SEcExAHTPqE

Requirements of screening test (WHO)

- The disease is well defined
- Screening detects a different spectrum of disease from the disease that presents clinically (length-time bias)
- In the case of cancer, screening will detect some slow growing cancer
- There is a long period between when disease can be first detected and when the disease will present clinically
- The disease is serious and there is effective treatment available
- The screening test is simple and safe
- The test result distinguishes clearly between those with and those without the disease
- Doing the screening test is cost effective
- The facilities needed to do both the screening test and deal with the positive results are available
- The path for dealing with a positive result is clear and is acceptable both to the people being screened and to the authorities doing the screening, and there is equity in access to the test

Screening test concepts

- Lead time: extra time during which you know you have the disease if it is diagnosed by screening rather than by clinical presentation. Because of lead time bias, survival will look longer in screened individuals even if the course of their disease is unaffected.

- Length time: screening tends to diagnose disease that is less aggressive then disease that presents clinically. Because of length time bias, some cases diagnosed by screening would never present clinically if they had not been detected by screening: over diagnosis.

Type I and II errors

- Type I error = we reject the null hypothesis when the null hypothesis is true (finding a difference when one does not exist)

- Type II error = we retain the null hypothesis when the null hypothesis is false (not finding a difference when one exists). Often related to sample size

Choice of study question

- Interesting and relevant to you, your patients and your community
- “Opportunity costs” – prioritise, with limited resources we must research the most important topics
- Do not just duplicate methodology or question from previous research – a lost opportunity to advance the science or explore a new dimension of a question or topic
- Think beyond the clinical biomedical model
- Consider multi-modal methodologies (quantitative and qualitative)
- Implementation science

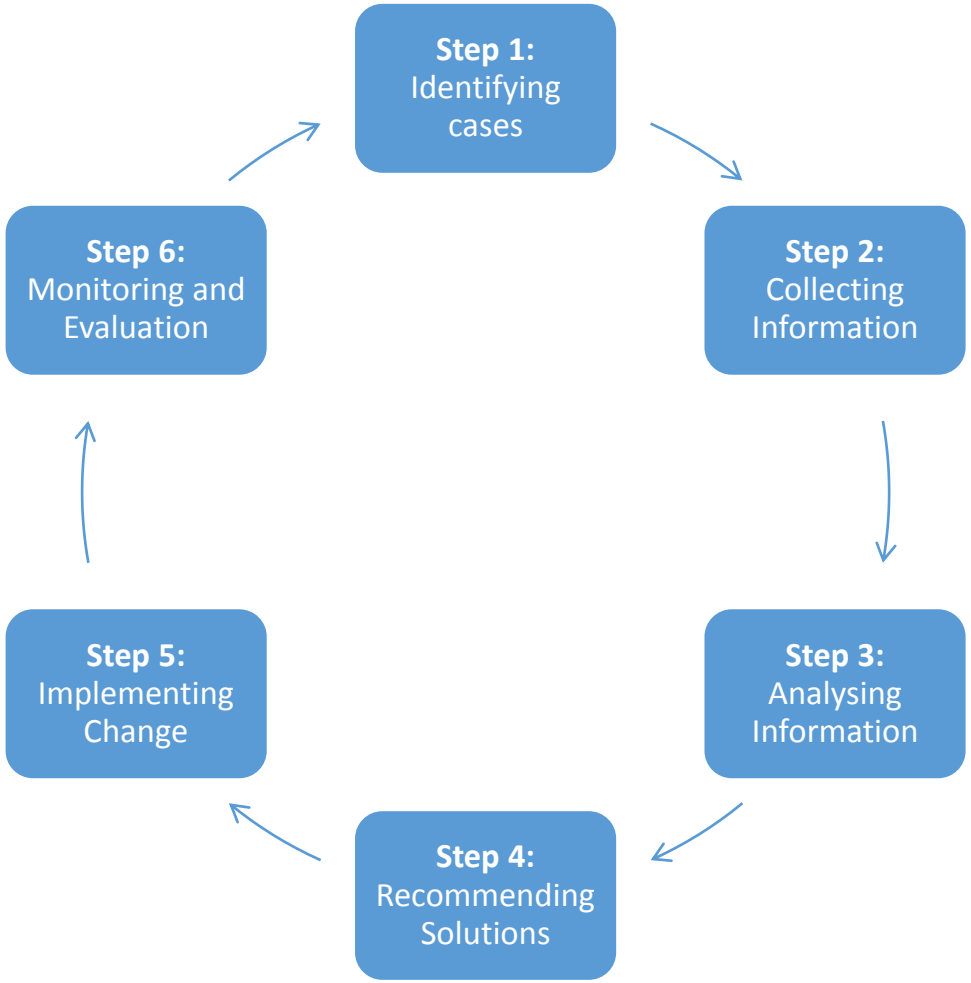
Implementation research

- Much evidence on efficacy of interventions to prevent child deaths, but varying degree of implementation and effectiveness – Why?
- Embed research in real-world practice
- Prioritise questions of local relevance
- Knowledge translation
- E.g. quality improvement research, mortality auditing

Quality improvement research

- Implementation of new clinical programs, approaches, evaluation of improvements to programs
- Many different study designs:
  - Before-and-after evaluation (historical controls)
  - Evaluate *whether* it works, *where* it works, *why* it works, and *what are the important ingredients* to make it work
- Multi-faceted interventions
  - E.g. How to reduce nosocomial infections, how to improve the management of severe malnutrition
- Incremental phased improvements and rigorous routine data for monitoring
- Mortality and morbidity auditing

Quality improvement cycle



Reducing hospital-acquired infections and improving the rational use of antibiotics in a developing country: an effectiveness study

Murni IK, *et al*. *Arch Dis Child* 2014;**0**:1–6. doi:10.1136/archdischild-2014-307297

Severe malnutrition in children in Papua New Guinea: effect of a multi-faceted intervention to improve quality of care and nutritional outcomes

Paediatrics and International Child Health 2015

Ethics in research
How to write a minor thesis

Lecture 5

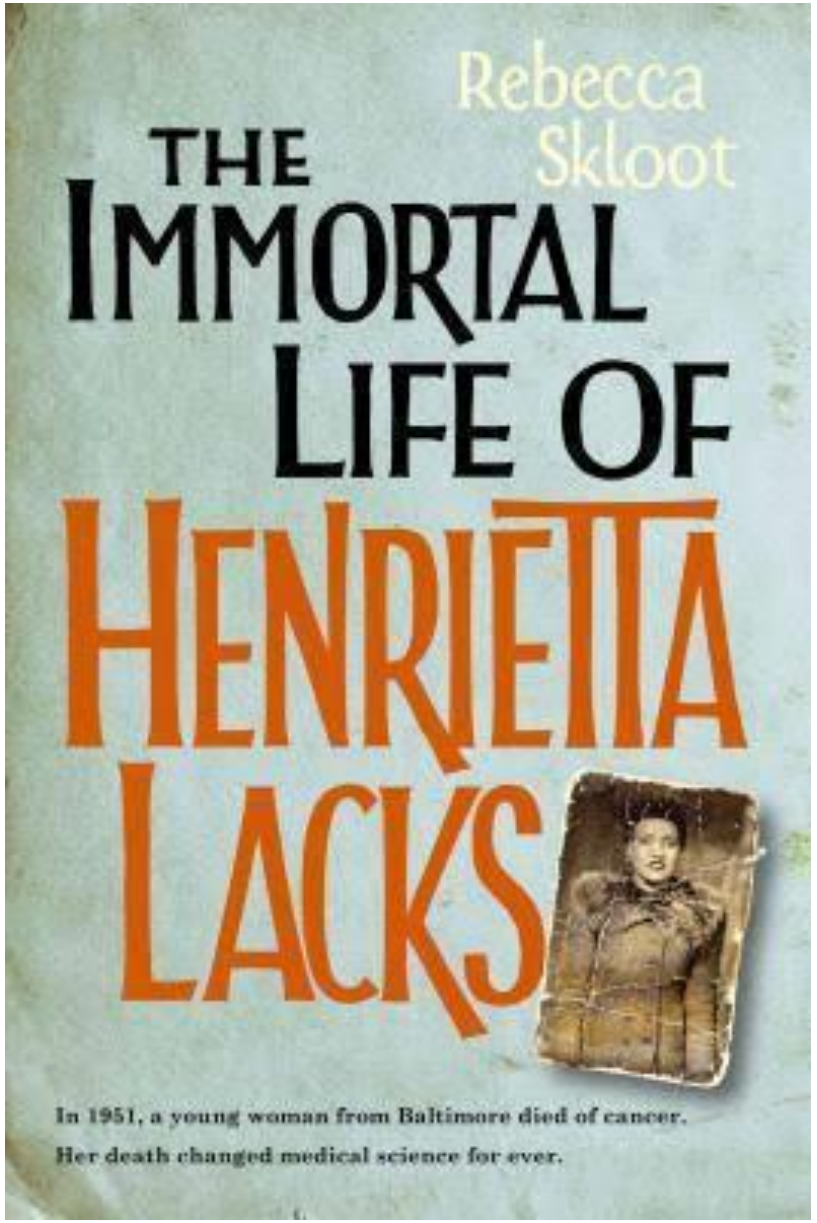
James Lind, HMS Salisbury, May 20th 1747

- Many sailors dying from scurvy
- 12 sailors chosen from 30 who were sick with scurvy
  - 2: given 2 oranges and 1 lemon each day
  - Rest given other things, including 2 given sea-water
- Within a week, the 2 given citrus were healthy, the others sick or dying

- Nazi war experiments
- Tuskegee syphilis experiments (1930s)

Ethics of research

- The Nuremberg Code (1947): the first international statement on the ethical treatment of humans in research
  - Voluntary consent is essential
  - The research should be beneficial for society
  - Experiments should be well designed in line with current knowledge
  - Experiments should avoid unnecessary risk or suffering or injury to participants
  - Risk/benefit analysis should justify the research
  - Experiments should only be conducted by qualified scientists demonstrating “the highest degree of skill and care"
  - The research should cease if the subject withdraws consent or there is reason to believe the continuation of the research will be harmful


Rebecca
Skloot
THE
IMMORTAL
LIFE OF
HENRIETTA
LACKS
In 1951, a young woman from Baltimore died of cancer.
Her death changed medical science for ever.

Declaration of Helsinki (1964):

- Built on the Nuremberg code (which had been largely ignored)
- Patient welfare is the primary responsibility of all researchers and medical professionals
  - Needs ethics approval (Ethics Committees)
  - Includes
    - surveys/questionnaires
    - access to medical and other personal records
    - collection of body tissues and fluids

Principle of human research ethics

Consent
- Informed, voluntary, comprehension (plain language), right to refuse/withdraw (no reason required)

Maximise autonomy and human dignity
- Participants have the freedom to decide what will happen to them
- Respect for different cultural/religious beliefs
- Responsibility to protect those with diminished autonomy (children, medically-dependent people, confined populations)

Maintain confidentiality
- Ensure participant records are kept secure
- Autonomous decision-making (not possible in the absence of privacy)
- Identifiable, re-identifiable and non-identifiable records pose different problems for patient rights

Non-maleficence:
- Maintain confidentiality
- risk/benefit analyses
- Avoid psychological, physiological, and social harm to participants
- Participant welfare more important than scientific discovery

Principle of human research ethics

Beneficence

- Maximise possible benefits
- The research must not only avoid harm but must contribute something positive to society
- Risk must be kept to a minimum and must be justified in terms of potential benefits

Justice

- Fair selection of participants
- Fair distribution of burdens and benefits of the research
- Transparent, non-discriminatory recruitment procedures and inclusion/exclusion criteria

Scientific integrity:

- Publication of results for scrutiny
- methodology should be clearly explained so experiments can be independently repeated
- Results should never be fabrication/concealed
- Selection of participants should be justified and unbiased - no under or over representation
- Valid and rigorous methodology
- Sample sizes must be capable of yielding statistically significant results
- Poor research methodologies are unethical as they waste resources, time and show disrespect for participants

Ethical theories

Consequentialism

- Focuses on consequences
- Ends justify the means

Utilitarianism

- Focuses on achieving the greatest good for the greatest number
- Aims to maximise utility, which can be defined as achieving the most happiness, health
- Sometimes used as a basis for cost-benefit analyses

Deontology (Kantian)

- Focuses on rights, duties and other intrinsic moral features of actions, rather than the consequences of those actions
- The rightness or wrongness of actions does not depend on their consequences but on whether they fulfil our duty

Virtue ethics (a form of Deontology)

- Character matters above all else.
- Living an ethical life, or acting rightly, requires developing and demonstrating the virtues of courage, compassion, wisdom, and temperance, and avoidance of greed, jealousy, and selfishness

How to write a thesis

- Start early
- Set aside some time every week to do some work on your study and thesis
- Keep your supervisor informed and interested in your study and thesis progress
- Documents
- Back-up
- Writing style

Thesis structure

- Title page
- Declaration
- Acknowledgements
- Table of Contents
- Lists of Tables Figures and Diagrams
- Abstract
- Introduction – including objectives and specific research question(s)
- Literature review
- Methods
- Results
- Discussions
- Conclusions and recommendations
- Reference list
- Appendices

An idea or problem

A clear research question

Define objectives and hypotheses

Review of the relevant literature

Learn about End-Note

A valid methodology to address the question

Metrics of measurement

Data collection forms

Ethics proposal

Funding

Engaging others

A spread-sheet that reflects the data in the data collection form

Gather the data / conduct the study

Develop an analysis plan

Commence writing: intro / methods / dummy tables

Analysis and writing

Minor thesis / Publication