Basic research methods 1: Designing, analysing and writing up your project

Prof Trevor Duke

Sept 6, 2021

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

How to search the literature

- https://pubmed.ncbi.nlm.nih.gov/
- Pubmed: 32 million papers, 14,000 journals

How to read a paper - structure

1. Title and Abstract		**Objective**	Relevant Y /N
2. Introduction	Why I did it	Subjective	
3. Method	**How I did it**	**Objective**	Quality / valid
4. Results	**What I found**	**Objective**	Quality / valid
5. Discussion	What it means	Subjective	
6. Conclusion		Subjective	

Epidemiology

- Basic epidemiology
- Types of studies
- Basic statistics – mean, median, incidence, prevalence, OR, RR

Epidemiology

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

Types of epidemiology

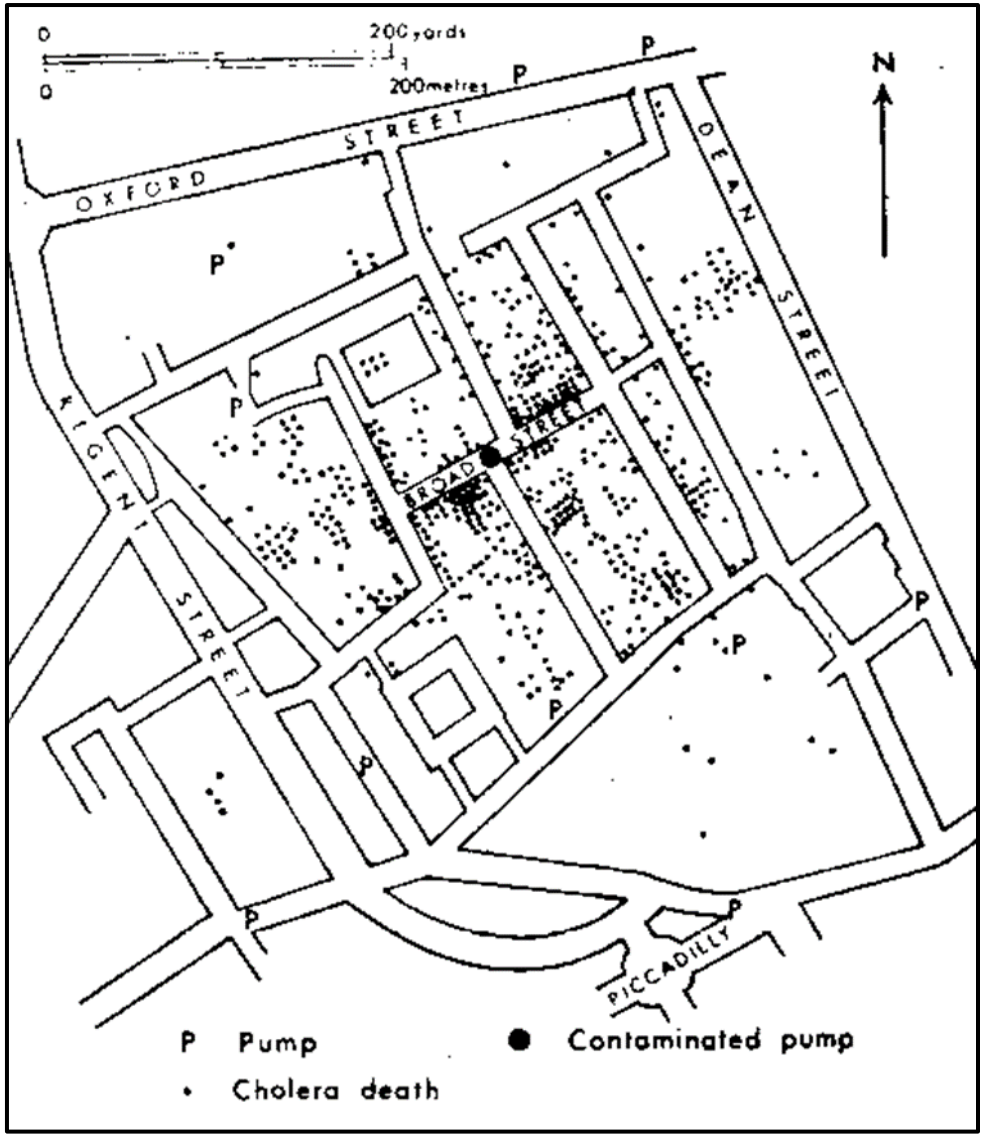
- Descriptive
  - Describing disease by time, place, person
  - Measuring the burden of disease
- Analytical
  - Looking for associations between exposures and outcomes, and between comorbidities and outcomes
- Interventional
  - Evaluating interventions
- Clinical
- Public health

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

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

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

- 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


Reprints and permissions:
sagepub.com/journalsPermissions.nav
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]

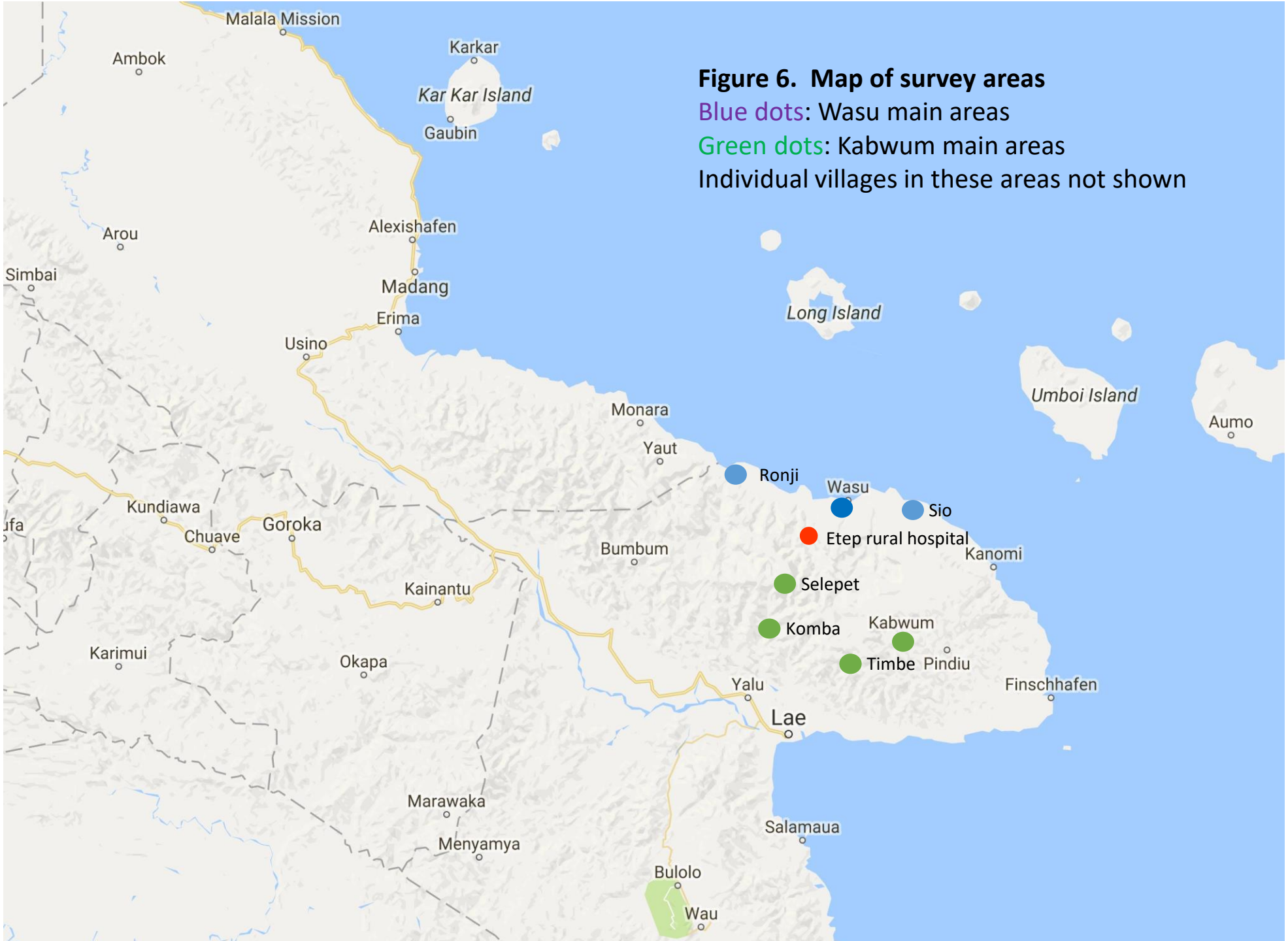
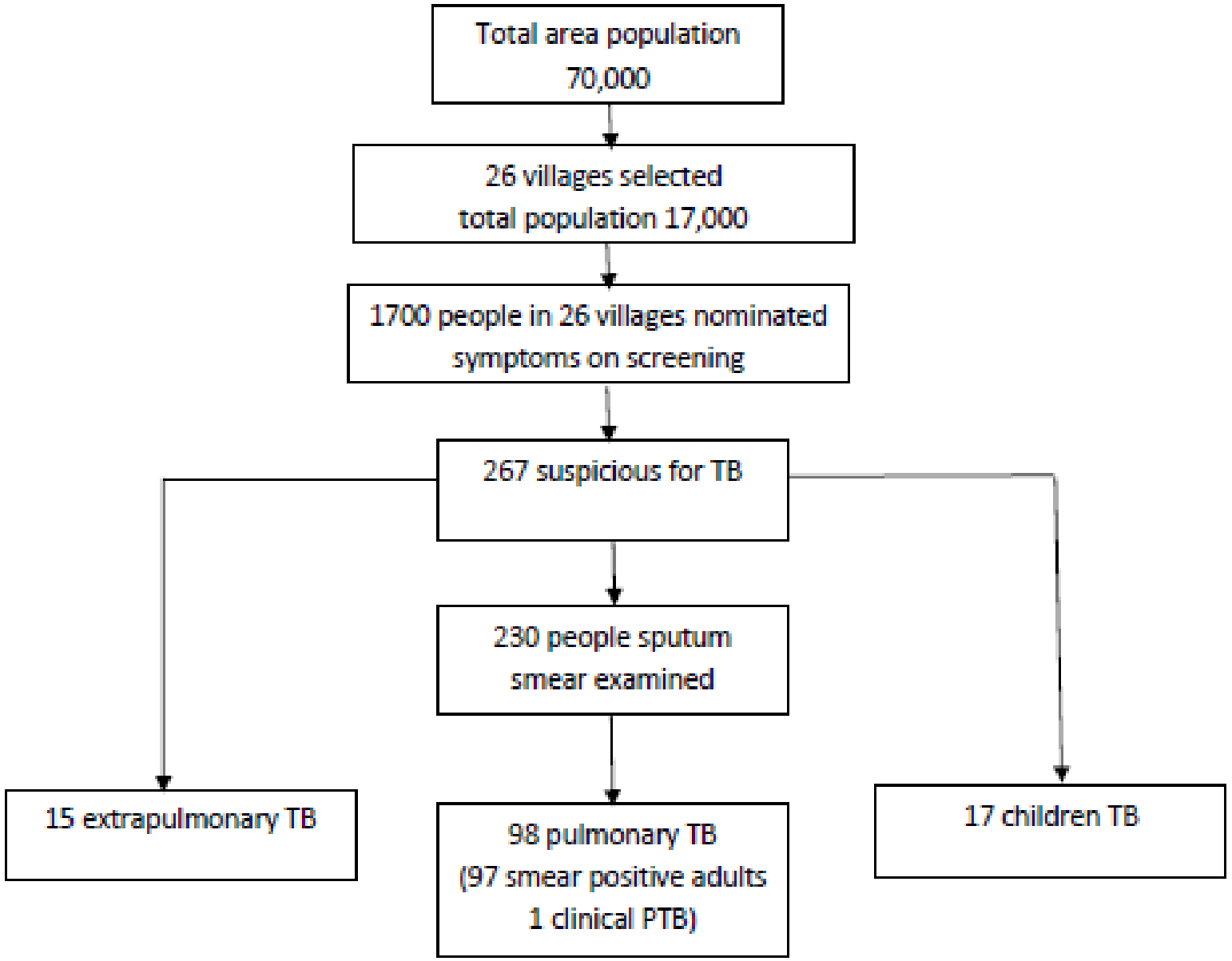


Figure 6. Map of survey areas

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 - quiz

“Do you currently have asthma?”	
“Have you had asthma during the last 2 years?”	
“Have you ever had asthma?”	

Life-time cumulative prevalence?
Point prevalence?
Period prevalence?

Several types of prevalence - quiz

“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

Data collection forms and spreadsheets

- The questions should be objective
- The method should be appropriate to the questions
- The data collection form should reflect your questions
- A spreadsheet should reflect your data collection form

Data collection form → Spreadsheet

- A spreadsheet should reflect your data collection form
- The same order so it is easy to enter data
- Types of variables:
  - Continuous
  - Binary (yes / no)
  - Categorical
- Yes or no responses should be represented as 1 or 0.
- Continuous variables such as weight, length, head circumference, MUAC, duration of illness should be numerical to a fixed number of decimal places.

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 – better

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

Spreadsheets – ideal

SPO2 (%)	RR (bpm)	PR bpm	BT (degC)	Pallor	Edema	Hepatomeg	SevAnaemia	HIV	TB	CHD	Malaria	Meningitis	Others4	Outcome	Date of Adm
99	38	120	38	0	0	0	0	0	0	0	0	0	0	1	19.5.21
99	40	150	40	1	1	1	1	0	1	0	0	0	0	1	20.5.21
92	36	138	37.5	0	0	0	0	0	0	0	0	0	0	1	20.5.21
100	30	102	38	0	0	0	0	0	1	0	0	0	0	1	23.5.21
99	28	100	36.7	1	0	1	1	1	1	0	0	0	0	1	17.5.21
99	34	110	36.5	1	0	0	0	0	0	0	0	0	0	1	24.5.21
97	24	124	36.6	0	0	0	0	0	0	0	0	0	0	1	24.5.21
99	30	105	36.8	0	0	0	0	0	1	0	0	0	0	1	25.5.21
100	30	110	37.8	0	0	0	0	0	1	0	0	0	0	1	22.5.21
100	24	134	36.2	1	0	0	0	0	0	0	0	0	0	0	26.5.21
99	30	86	36.6	1	1	0	1	0	0	0	0	0	0	1	27.5.21
97	28	124	35	0	0	0	0	0	0	0	0	0	0	1	22.5.21
98	24	100	36.4	1	0	0	1	0	0	0	0	0	Hookworm	1	31.5.21
99	28	113	37.3	1	0	0	0	1	1	0	0	0	0	1	31.5.21
100	26	102	37.9	0	0	0	0	0	0	0	0	0	0	1	2.6.21
97	30	120	37.1	0	0	0	0	0	0	0	0	0	0	1	2.5.21
98	32	128	36.8	0	0	0	0	0	0	0	0	0	0	1	2.6.21
100	18	102	36.8	0	0	0	0	0	0	0	0	0	0	1	2.6.21
98	30	120	37.8	0	0	0	0	0	0	0	0	0	0	1	2.6.21
99	28	110	36.7	1	0	0	0	0	0	0	0	0	0	1	2.6.21
96	30	120	36.2	1	0	0	1	0	1	0	0	0	0	1	8.6.21
97	28	102	36.8	1	0	0	0	0	0	0	0	0	0	1	9.6.21
98	30	110	36.3	0	0	0	0	0	1	0	0	0	0	1	6.6.21
97	28	128	36.3	1	0	0	0	0	1	0	0	0	RTA	1	30.5.21
97	34	114	37.5	0	0	0	0	0	0	0	0	0	0	1	10.6.21

Making up a spreadsheet

- **Do not use categories in your data collection form or spreadsheet**: use numbers for continuous variables
- E.g. number of people in a household …<4…. 5-8….>8”
- Why not?
  - When you record data as categorical it cannot be analysed
  - You lose information / precision
  - A computer cannot understand > or <
- There may be value in categorising later, but not for data entry and analysis.

Making up a spreadsheet

- **Use the same metric of measurement consistently in a variable.** Do not record age in months for some and years for others, and some in days.
- If you record in months:
  - 6 months = 6
  - one year 8 months = 20
  - 5½ years = 66
  - 2 weeks = 0.5
  - Newborn day 1 = 0.03

Tables

- Create “dummy tables” to plan your data presentation
- Most studies have 2 or 3 tables:
  1. Demographics
  2. Results
  3. Results

	Baseline Survey February 15-16, 2014	First post-intervention survey August 8-9, 2014	Second post-intervention survey October 5-6, 2014
Total in-patients			
Severe malnutrition			
Age median months (IQR)			
Males			
Median length of stay days (IQR)			
Comorbidity			
Extra pulmonary TB			
Diarrhoeal disease			
Pulmonary TB			
ALRTI			
Others			
Primary malnutrition			
HIV/AIDS			

	Baseline Survey February 15-16, 2014	**First post-intervention survey August 8-9, 2014**	**Second post-intervention survey** **October 5-6, 2014**
Total in-patients	125	120	118
Severe malnutrition	43 (34.4)	38 (31.7)	35 (29.7)
Age median months (IQR)	24 (14 – 36)	17.5 (12-28)	17 (10-27)
Males	27 (62.8)	26 (68.4)	20 (57.1)
Median length of stay days (IQR)	16 (7-32)	8.5 (5-23)	8 (4-14)
Comorbidity			
Extra pulmonary TB	14 (32.6)	6 (15.8)	8 (22.9)
Diarrhoeal disease	10 (23.3)	5 (13.2)	10 (28.6)
Pulmonary TB	9 (20.9)	8 (21.1)	4 (11.4)
ALRTI	4 (9.3)	3 (7.9)	1 (2.9)
Others	3 (7)	8 (21.1)	5 (14.3)
Primary malnutrition	2 (4.7)	4(10.5)	5 (14.3)
HIV/AIDS	1 (2.3)	4 (10.5)	2 (5.7)

	Baseline Survey February 15-16, 2014	First post-intervention survey August 8-9, 2014	Second post-intervention survey October 5-6, 2014
Feeding			
Average day of initiation of feeds (IQR)			
Difference between the baseline survey and the two follow-up surveys			
Feeding volume in last 24 hours in ml: median (IQR)			
Difference between the baseline survey and the two follow-up surveys			
Percentage of required calories received in last 24 hours (IQR)			
Difference between the baseline survey and the two follow-up surveys			
Weight change			
Median weight gain in grams/kg/day (IQR)			
Difference between the baseline survey and the two follow-up surveys			

	Baseline Survey February 15-16, 2014	First post-intervention survey August 8-9, 2014	Second post-intervention survey October 5-6, 2014
Feeding			
Average day of initiation of feeds (IQR)	2 (1-5)	1 (1-4)	2 (1-2)
Difference between the baseline survey and the two follow-up surveys: p = 0.31			
Feeding volume in last 24 hours in ml: median (IQR)	356 (178-450)	820 (600-1110)	780 (480-900)
Difference between the baseline survey and the two follow-up surveys: p < 0.001			
Percentage of required calories received in last 24 hours (IQR)	31% (21-48%)	98% (67-100%)	86% (46-100%)
Difference between the baseline survey and the two follow-up surveys: p < 0.001			
Weight change			
Median weight gain in grams/kg/day (IQR)	1.55 (-4.3-6.0)	5.56 (-3.7-12)	10.19 (0-16)
Difference between the baseline survey and the two follow-up surveys: p = 0.013			

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

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 – single document: proposal, thesis
- Back-up your data
- Writing style – concise

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