Models

Phillip Ayieko

Simple or general linear model?

- Here **general** refers to the dependence on potentially more than one explanatory variable (i.e. multiple linear regression), v.s. the **simple linear model**:

Simple or general linear model?

- The model is linear in the parameters, e.g.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon_i$$

- But not e.g.

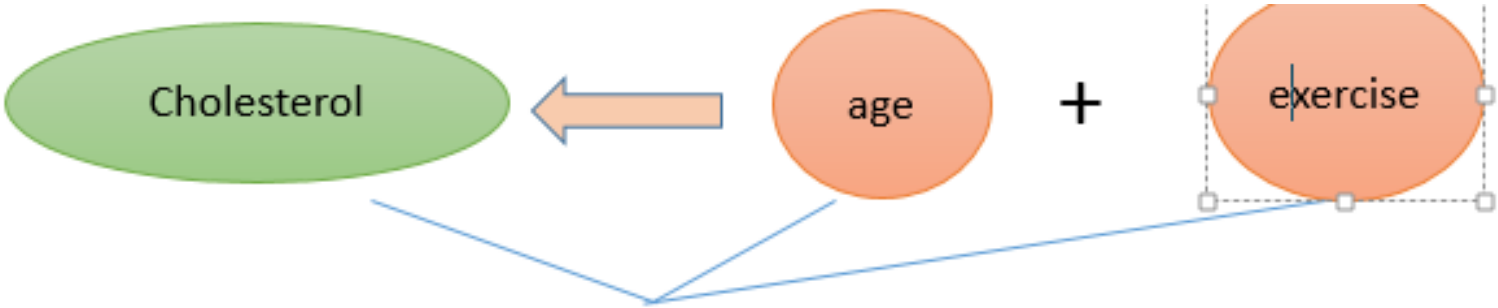
$$y_i = \beta_0 + \beta_1 x_1^{\beta_2} + \epsilon_i$$

General or generalized linear model?

- There are situations where, general linear models are inappropriate:
- The range of Y is restricted (e.g. binary, count)
- The variance of Y depends on the mean
- Generalized linear models extend the general linear model to address these issues

General or generalized linear model?

- A generalized linear model is made up of a linear predictor = relates mean to predictors
- and two functions:
- a link function (transform done on Y) = relates means of observations to predictors



- a variance function (the distribution) = relates the means to the variances

Poisson Regression or Regression of Counts (& Rates)

Poisson Regression

- is a generalized linear model:

$$linearpredictor = \eta$$

$$linkfunction(transformation) = logtransform$$

$$variancefunction(thedistribution) = V(\mu) = \mu$$

Poisson Regression

Effect of expanded insecticide-treated bednet coverage on child survival in rural Kenya: a longitudinal study

Greg W Fegan, Abdisalan M Noor, Willis S Akhwale, Simon Cousens, Robert W Snow

Summary

Background The potential of insecticide-treated bednets (ITNs) to contribute to child survival has been well documented in randomised controlled trials. ITN coverage has increased rapidly in Kenya from 7% in 2004 to 67% in 2006. We aimed to assess the extent to which this investment has led to improvements in child survival.

Methods A dynamic cohort of about 3500 children aged 1–59 months were enumerated three times at yearly intervals in 72 rural clusters located in four districts of Kenya. The effect of ITN use on mortality was assessed with Poisson regression to take account of potential effect-modifying and confounding covariates.

Lancet 2007; 370: 1035–39

See Editorial page 1007

See Comment page 1009

Malaria Public Health and Epidemiology Group, Centre for Geographic Medicine Research—Coast, Kenya Medical Research Institute/Wellcome Trust

Poisson Regression

- is a generalized linear model:

$$linearpredictor = \eta$$

$$linkfunction(transformation) = logtransform$$

$$variancefunction(thedistribution) = V(\mu) = \mu$$

Poisson Regression

OPEN ACCESS Freely available online

PLoS one

Evidence for Over-Dispersion in the Distribution of Clinical Malaria Episodes in Children

Tabitha Wanja Mwangi[1]*, Gregory Fegan[1,2], Thomas Neil Williams[1,3], Sam Muchina Kinyanjui[1,3], Rober William Snow[3,4], Kevin Marsh[1,3]

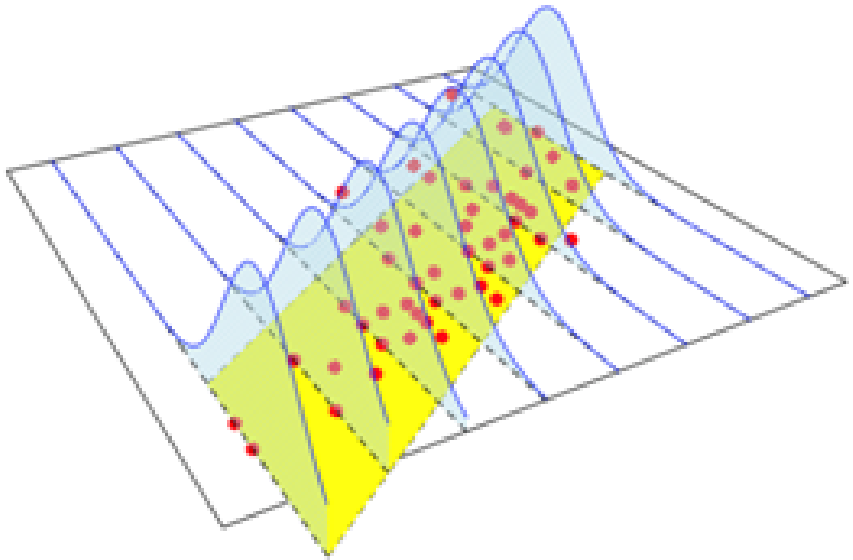
1 Kenya Medical Research Institute, Centre for Geographic Medicine Research, Coast/Wellcome Trust Collaborative Program, Kilifi, Kenya, **2** Infectious Disease Epidemiology Unit, Department of Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom, **3** Centre of Tropical Medicine, Nuffiel Department of Clinical Medicine, John Radcliffe Hospital, Oxford, United Kingdom, **4** Kenya Medical Research Institute/Wellcome Trust Collaborative Program, Nairob Kenya

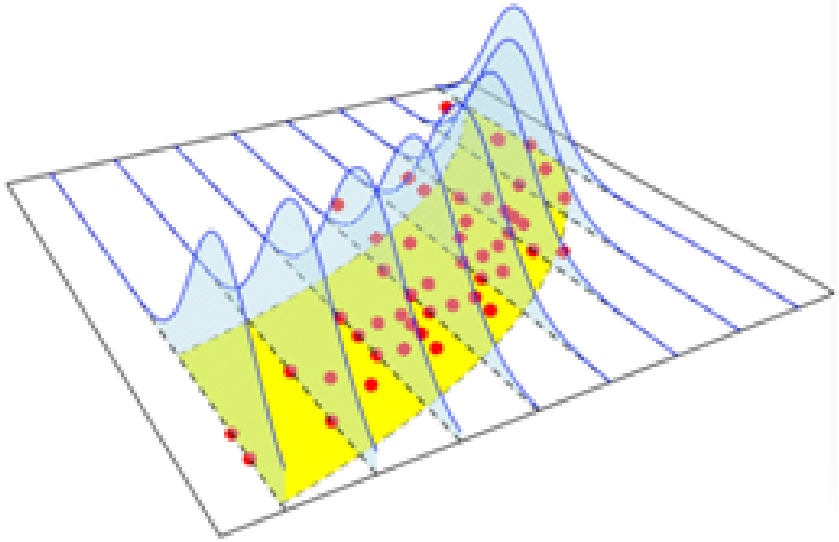
Abstract

Background: It may be assumed that patterns of clinical malaria in children of similar age under the same level of exposure would follow a Poisson distribution with no over-dispersion. Longitudinal studies that have been conducted over many years suggest that some children may experience more episodes of clinical malaria than would be expected. The aim of this study was to identify this group of children and investigate possible causes for this increased susceptibility.

Methodology and Principal Findings: Using Poisson regression, we chose a group of children whom we designated as 'more susceptible' to malaria from 373 children under 10 years of age who were followed up for between 3 to 5 years from 1998–2003. About 21% of the children were categorized as 'more susceptible' and although they contributed only 23% of

Poisson Regression





- *L*inearity
- *I*ndependence
- *N*ormality
- *E*qual variance

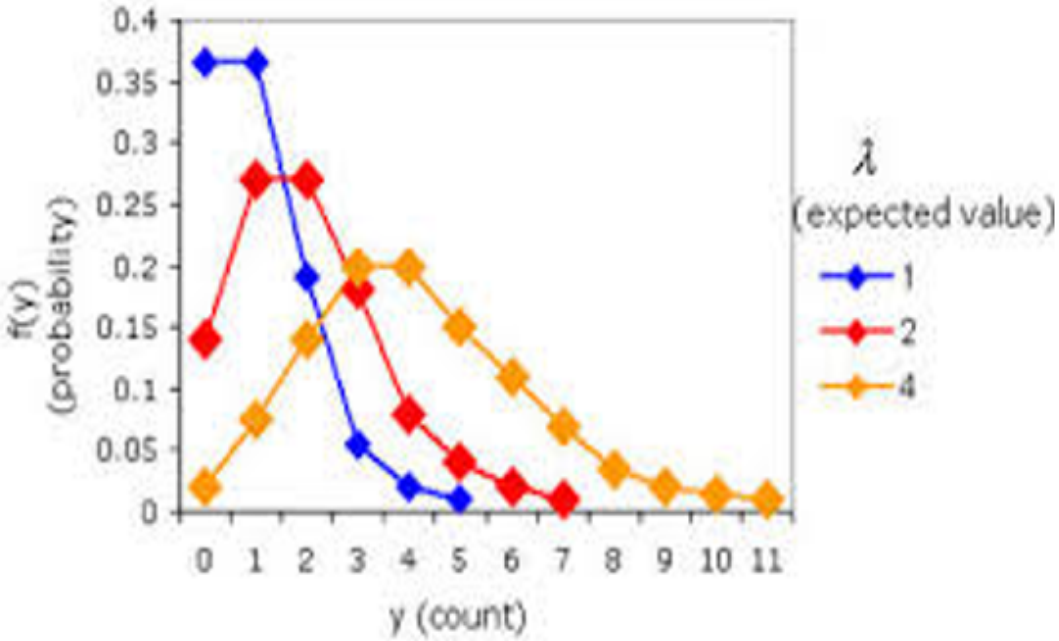
Link function – Poisson regression

Table 13.1 glm() parameters

Family	Default link function
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

$$poisson.model = glm(Y \sim x_1 + x_2 + \ldots + x_n, data = dataframe, family = poisson(link = "log")) \quad (1)$$

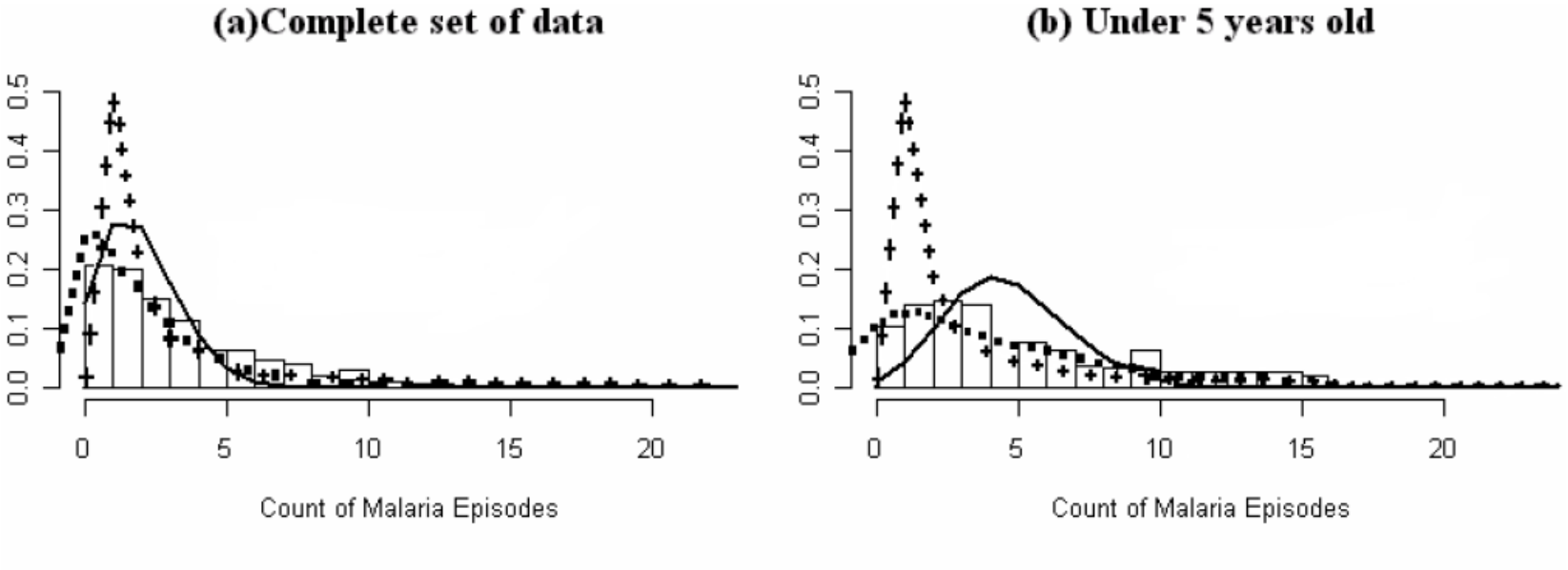
Poisson distribution (discrete probability distribution)



- Not a single shape
- Described by a single parameter (mean - lambda)
- The mean = variance
- As the mean increases the distribution spreads out

Poisson regression (Mwangi et al)

- What is the difference in malaria episodes in children < 5 yrs. and all children (including > 5 yrs.) in Kilifi?



Outcome in regression models

Type of regression	Typical use
Simple linear	Predicting a quantitative response variable from a quantitative explanatory variable
Polynomial	Predicting a quantitative response variable from a quantitative explanatory variable, where the relationship is modeled as an nth order polynomial
Multiple linear	Predicting a quantitative response variable from two or more explanatory variables
Multivariate	Predicting more than one response variable from one or more explanatory variables
Logistic	Predicting a categorical response variable from one or more explanatory variables
Poisson	Predicting a response variable representing counts from one or more explanatory variables
Cox proportional hazards	Predicting time to an event (death, failure, relapse) from one or more explanatory variables
Time-series	Modeling time-series data with correlated errors
Nonlinear	Predicting a quantitative response variable from one or more explanatory variables, where the form of the model is nonlinear

R syntax

$$poisson.model = glm(Y \sim x_1 + x_2 + \ldots + x_n, data = dataframe, family = poisson(link = "log")) \quad (2)$$

Hypothesis testing in Poisson regression

Count

Rates

	Dec 04–Jan 05 to Dec 05–Jan 06			Dec 05–Jan 06 to Dec 06–Jan 07			Combined follow-up time		
	Total*	With ITN	No ITN	Total*	With ITN	No ITN	Total*	With ITN	No ITN
Bondo									
1–5 months	8/64·2 (124·6)	1/15·4 (65·0)	7/45·7 (153·2)	4/54·3 (73·7)	2/18·5 (108·1)	2/35·7 (56·0)	12/118·5 (101·3)	3/33·9 (88·48)	9/81·4 (110·6)
6–11 months	9/81·2 (110·8)	3/20·4 (147·4)	6/57·9 (103·6)	8/61·5 (130·0)	2/32·7 (61·2)	6/28·9 (207·6)	17/142·8 (119·0)	5/53·0 (94·3)	12/86·8 (138·2)
1–5 years	13/563·2 (23·1)	0/111·2 (0·0)	13/442·6 (29·4)	16/465·1 (34·4)	6/197·7 (30·4)	10/265·4 (37·7)	29/1028·3 (28·2)	6/308·9 (19·4)	23/708·0 (32·5)
Rate ratio†		0·50 (0·17–1·47; p=0·20)			0·70 (0·33–1·48; p=0·35)			0·62 (0·33–1·14; p=0·12)	
Kisii									
1–5 months	1/87·0 (11·5)	0/10·5 (0·0)	1/75·0 (13·3)	2/65·5 (30·5)	2/27·4 (72·9)	0/38·1 (0·0)	3/152·6 (19·7)	2/37·9 (52·8)	1/113·1 (8·8)
6–11 months	0/109·4 (0·0)	0/14·1 (0·0)	0/94·2 (0·0)	0/72·9 (0·0)	0/33·4 (0·0)	0/39·5 (0·0)	0/182·3 (0·0)	0/47·5 (0·0)	0/133·7 (0·0)
1–5 years	5/708·9 (7·1)	1/93·3 (10·7)	4/607·6 (6·6)	1/639·6 (1·6)	0/260·8 (0·0)	1/375·8 (2·7)	6/1348·5 (4·4)	1/354·1 (2·8)	5/983·4 (5·1)
Rate ratio†		1·32 (0·15–11·39; p=0·80)			2·84 (0·25–32·04; p=0·38)			1·91 (0·39–9·30; p=0·42)	
Kwale									
1–5 months	2/93·8 (21·3)	0/9·1 (0·0)	2/83·0 (24·1)	0/73·2 (0·0)	0/21·5 (0·0)	0/51·6 (0·0)	2/167·0 (12·0)	0/30·6 (0·0)	2/134·6 (14·9)
6–11 months	2/132·2 (15·1)	0/11·6 (0·0)	2/117·8 (17·0)	2/101·0 (19·8)	0/31·1 (0·0)	2/69·0 (29·0)	4/233·2 (17·1)	0/42·7 (0·0)	4/186·8 (21·4)
1–5 years	9/1003·6 (9·0)	0/78·1 (0·0)	9/904·2 (10·0)	14/898·5 (15·6)	1/258·3 (3·9)	13/638·2 (20·4)	23/1902·1(12·1)	1/336·4 (3·0)	22/1542·4 (14·3)
Rate ratio†		0 (0–3·67; p=0·27)‡			0·16 (0·02–1·23; p=0·04)§			0·13 (0·02–1·02; p=0·02)§	
Makueni									
1–5 months	1/66·8 (15·0)	1/7·7 (129·9)	0/58·4 (0·0)	0/47·7 (0·0)	0/20·7 (0·0)	0/27·0 (0·0)	1/114·5 (8·7)	1/28·4 (35·2)	0/85·4 (0·0)
6–11 months	0/87·4 (0·0)	0/11·8 (0·0)	0/74·0 (0·0)	0/59·9 (0·0)	0/26·4 (0·0)	0/33·5 (0·0)	0/147·3 (0·0)	0/38·2 (0·0)	0/107·5 (0·0)
1–5 years	2/595·7 (3·4)	0/55·6 (0·0)	2/521·1 (3·8)	1/572·8 (1·7)	0/218·7 (0·0)	1/350·5 (2·9)	3/1168·4 (2·6)	0/274·3 (0·0)	3/871·6 (3·4)
Rate ratio†		4·58 (0·38–55·8; p=0·19)			0 (0–60·3; p=0·43)‡			1·53 (0·21–10·95; p=0·67)	
Overall									
1–59 months	52/3593·6 (14·5)	6/438·9 (13·7)	46/3081·7 (14·9)	48/3112·9 (15·4)	13/1147·2 (11·3)	35/1953·2 (17·9)	100/6705·6 (14·9)	19/1586·1 (12·0)	81/5034·9 (16·1)
Rate ratio¶		0·60 (0·25–1·43; p=0·24)			0·57 (0·30–1·09; p=0·09)			0·58 (0·35–0·98; p=0·04)	

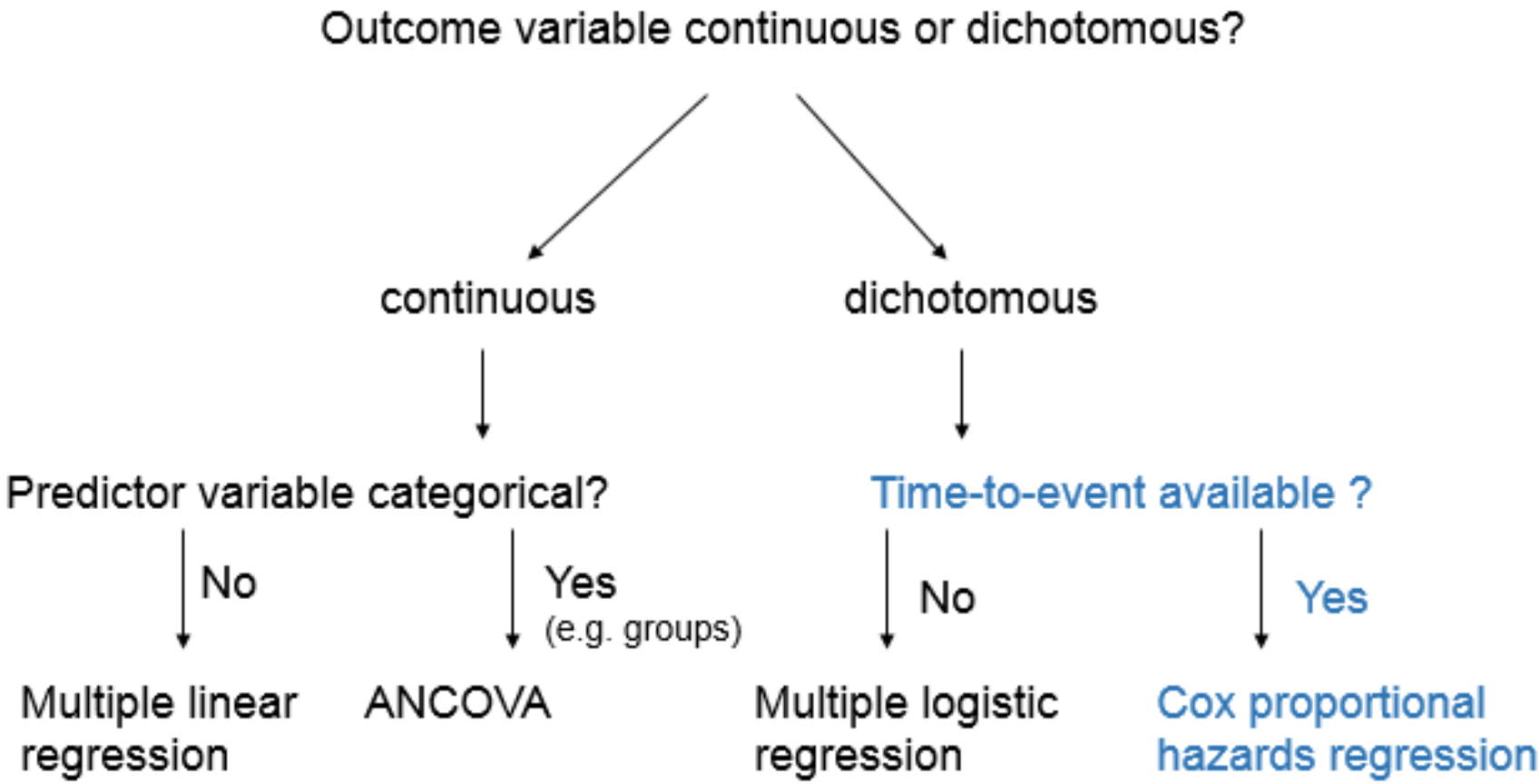
Data are deaths/person-years (rate per 1000 person-years) or rate ratio (95% CI; p value). *Person-year totals are larger than the sum of the nets and non-nets person times due to missing data for net status. †Mantel-Haenszel adjusted on age-group and also year for combined follow-up time. ‡Not adjusted for age due to lack of data but exact values computed. §Test is approximate, hence discrepancy between 95% CI including one and p value being less than 0·05. ¶Adjusted for age-group and district for the yearly surveys and also for year when combined.

Table 1: **Deaths, person-years of observation, and mortality rates by period of follow-up, district, and age**

Hypothesis testing and test for linear trend in Poisson regression

- **Recap of time-to-event (TTE) data**

Flow chart for regression models



Recap time-to-event data

MAJOR ARTICLE

Rates of Acquisition of Pneumococcal Colonization and Transmission Probabilities, by Serotype, Among Newborn Infants in Kilifi District, Kenya

Caroline C. Tigoi,[1] Hellen Gatakaa,[1] Angela Karani,[1] Daisy Mugo,[1] Stella Kungu,[1] Eva Wanjiru,[1] Jane Jomo,[1] Robert Musyimi,[1] John Ojal,[1] Nina E. Glass,[2] Osman Abdullahi,[1,3] and J. Anthony G. Scott[1,4]

[1]Kenya Medical Research Institute–Wellcome Trust Research Programme, Kilifi; [2]Department of Surgery, Langone Medical Center, New York University; [3]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts; and [4]Nuffield Department of Clinical Medicine, Oxford University, United Kingdom

Background. Herd protection and serotype replacement disease following introduction of pneumococcal conjugate vaccine (PCV) are attributable to the vaccine's impact on colonization. Prior to vaccine introduction in Kenya, we did an epidemiological study to estimate the rate of pneumococcal acquisition, by serotype, in an uncolonized population.

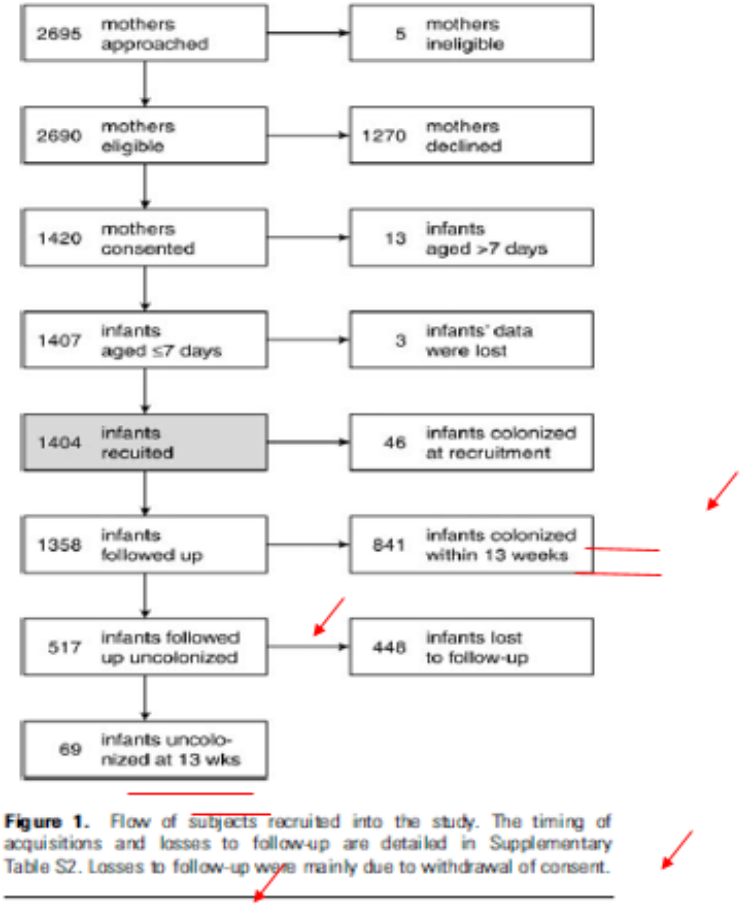


Figure 1. Flow of subjects recruited into the study. The timing of acquisitions and losses to follow-up are detailed in Supplementary Table S2. Losses to follow-up were mainly due to withdrawal of consent.

Time to event (TTE)

- Event: Pneumococcal acquisition among infants
- Time-to-event also available
- Statistical Modeling Approaches
- **Logistic Regression:**
  - Would do separate rate comparisons at distinct time points % with Pneumococcal acquisition by 60days, by 90 days
- **Cox Proportional Hazards Regression:**
  - Comparison of survival curves across all time points
    - Uses more information: Event (Yes/No), TTE
    - More powerful in identifying systematic differences

Time to first episode of malaria

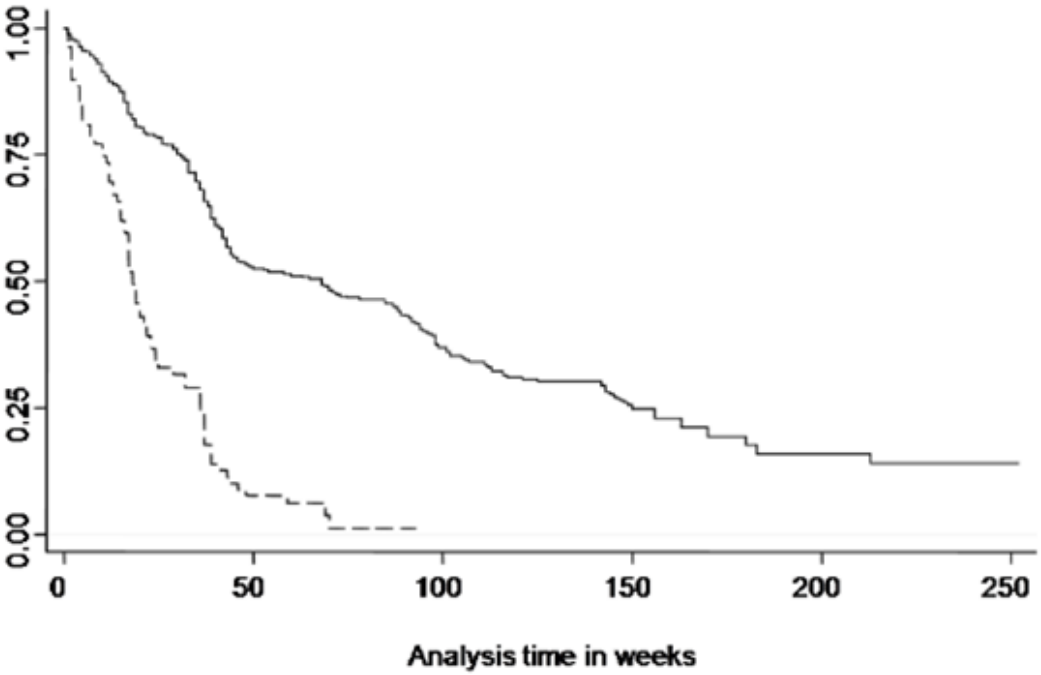


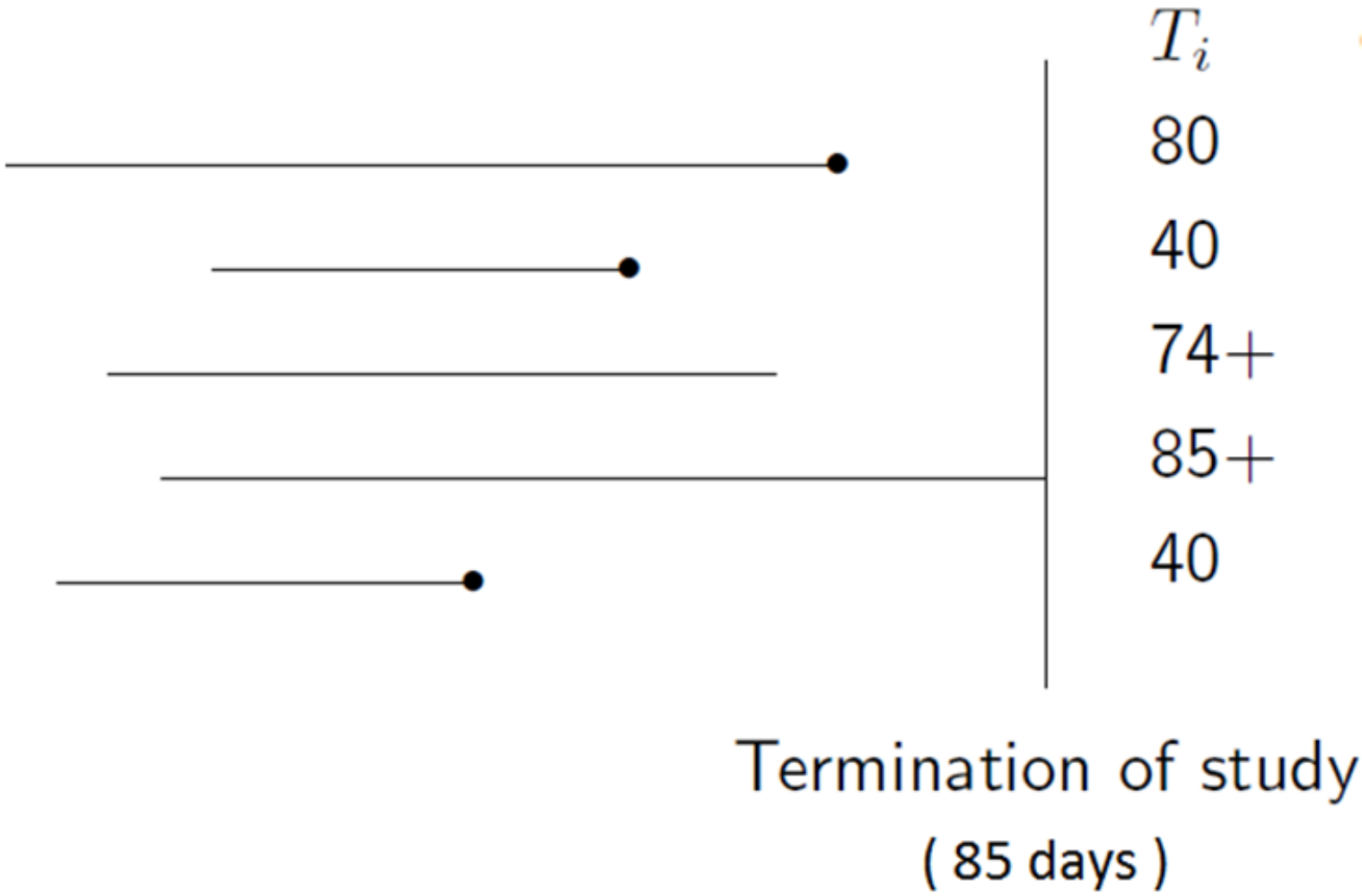
Figure 4. Kaplan-Meier survival curve of the time to first episode of clinical malaria. Dashed line represents the 'more susceptible' children and the solid line represents the time to first episode for the other children.
doi:10.1371/journal.pone.0002196.g004

Censoring

- Generally, three reasons why censoring might occur:
  - A subject does not experience the event before the study ends
  - A person is lost to follow-up during the study period
  - A person withdraws from the study These are all examples of right-censoring

Censoring

- Most typical to consider start of time-to-event "clock" as t=0



Classical analysis of rates

Life tables

Years	Alive at beginning	Deaths	Censored
0-1	146	27	3
1-2	116	18	10
2-3	88	21	10
3-4	57	9	3
4-5	45	1	3
		76	29
5-6	41	2	11
6-7	28	3	5
7-8	20	1	8
8-9	11	2	1
9-10	8	2	6

- Question: estimate the 5-year mortality rate? $S(5) = \Pr\{T \geq 5\}$
- Naive estimates
  1. 76 deaths/146 individuals=52.1%, $\hat{S}(5) = 47.9\%$,
  2. 76 deaths/(146-29)=65%, $\hat{S}(5) = 35\%$

Life tables

Years	Alive at beginning	Deaths	Censored
0-1	146	27	3
1-2	116	18	10
2-3	88	21	10
3-4	57	9	3
4-5	45	1	3
		76	29
5-6	41	2	11
6-7	28	3	5
7-8	20	1	8
8-9	11	2	1
9-10	8	2	6

- Question: estimate the 5-year mortality rate? $S(5) = \Pr\{T \geq 5\}$
- Naive estimates
  1. 76 deaths/146 individuals=52.1%, $\hat{S}(5) = 47.9\%$,
  2. 76 deaths/(146-29)=65%, $\hat{S}(5) = 35\%$

Life tables

- Life-table estimates:
  1. Assume censoring occurred at the right end of interval:

t	n	d	w	$q^r = d/n$	$p^r = 1 - q^r$	$\hat{S}^r = \prod p^r$
0-1	146	27	3	0.185	0.815	0.815
1-2	116	18	10	0.155	0.845	0.689
2-3	88	21	10	0.239	0.761	0.524
3-4	57	9	3	0.158	0.842	0.441
4-5	45	1	3	0.022	0.972	0.432

5-year survival rate estimate=0.432

- Censored observations are counted in the denominator of those "at risk" until they are censored

Life tables

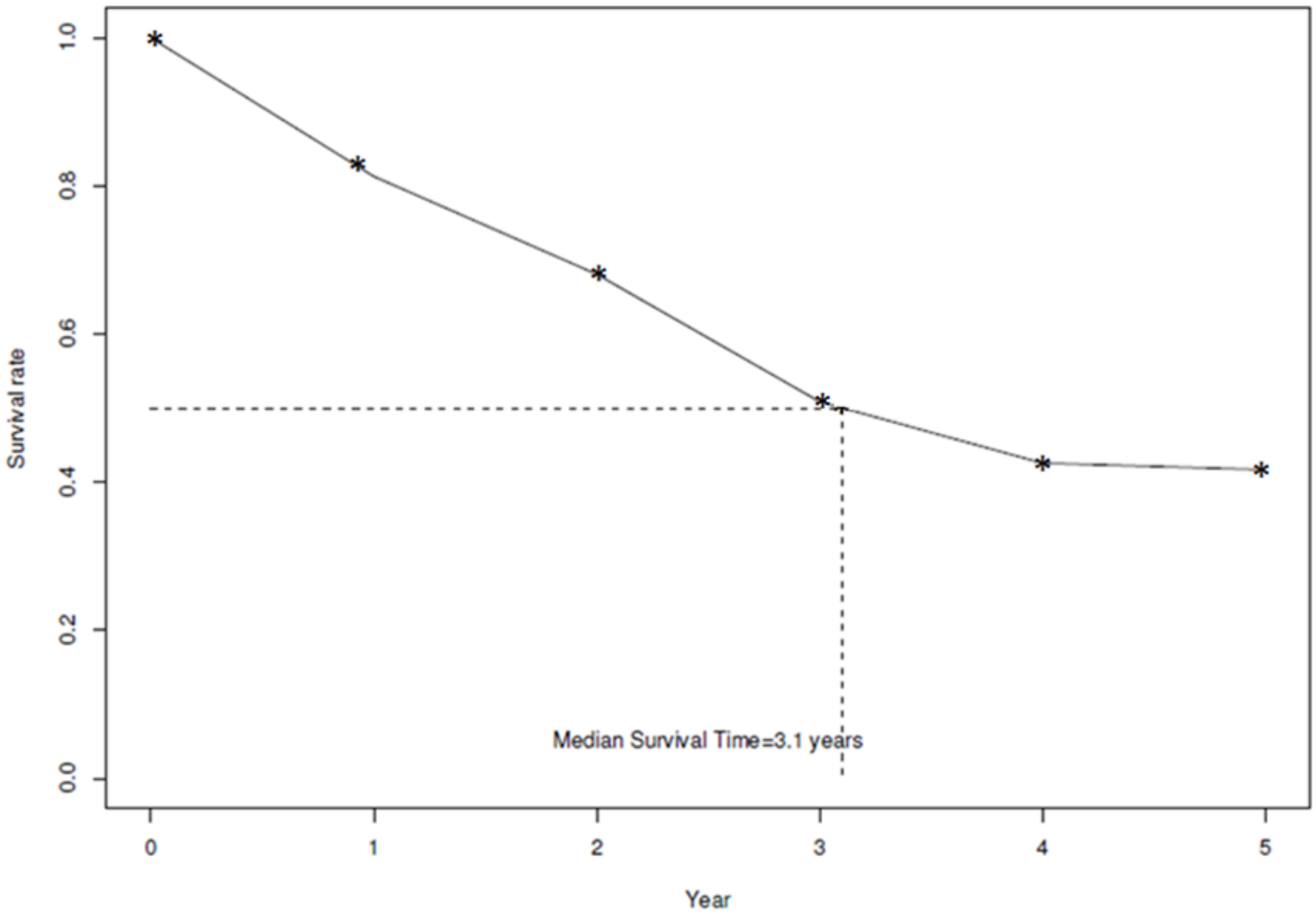
- Life-table estimates:
  1. Assume censoring occurred at the right end of interval:

t	n	d	w	$q^r = d/n$	$p^r = 1 - q^r$	$\hat{S}^r = \prod p^r$
0-1	146	27	3	0.185	0.815	0.815
1-2	116	18	10	0.155	0.845	0.689
2-3	88	21	10	0.239	0.761	0.524
3-4	57	9	3	0.158	0.842	0.441
4-5	45	1	3	0.022	0.972	0.432

5-year survival rate estimate=0.432

- Censored observations are counted in the denominator of those "at risk" until they are censored
  - Naive estimates
    1. 76 deaths/146 individuals=52.1%, $\hat{S}(5) = 47.9\%$,

Survival curve



Functions of interest in R

- Survival object: *Surv*
- Kaplan-Meier estimates: *survfit*
- The log-rank test: *survdiff*
- The Cox proportional hazards model: *coxph*
- The Accelerated failure time model: *survreg*
- Relevant R packages: *survival, survcomp, HMISC, Design, MASS*

Practical analysis of events over time