GLOSSARY AND TERMINOLOGY

There is considerable variation in terminology and methods of presenting data among epidemiology texts and other information sources. In general, the terminology and data layouts used in this book will conform to those used in Modern Epidemiology, 3rd edition (Rothman and Greenland, 2008).

GT. 1 DATA LAYOUT

The outcome variable is listed in the rows of the table, the predictor variable is listed in the columns.

Risk calculations (2X2 table)

	Exposure		
	Exposed	**Non-exposed**	
Diseased	a_1	a_0	m_1
Non-diseased	b_1	b_0	m_0
	n_1	n_0	n

where:

a_1 = the number of subjects that have both the disease and the risk factor.
a_0 = the number of subjects that have the disease but not the risk factor.
b_1 = the number of subjects that have the risk factor but do not have the disease.
b_0 = the number of subjects that have neither the disease nor the risk factor.
m_1 = the number of diseased subjects.
m_0 = the number of non-diseased subjects.
n_1 = the number of exposed subjects.
n_0 = the number of non-exposed subjects.
n = the number of study subjects.

In general, no distinction is made between values derived from a sample and population values as it is usually easy to determine what is being referred to from the context. In select situations where the distinction is necessary, upper-case letters (*eg* A_1) will be used for population values and lower case (*eg* a_1) for sample values.

Rate calculations (2X2 table)
Here, subject-time replaces the number of non-diseased.

	Exposure		
	Exposed	**Non-exposed**	
Number of cases	a_1	a_0	m_1
Animal-time at risk	t_1	t_0	t

where:

a_1 = the number of cases of disease in the exposed group.
a_0 = the number of cases of disease in the non-exposed group.
t_1 = the animal-time accumulated in the exposed group.
t_0 = the animal-time accumulated in the non-exposed group.
t = the total animal-time accumulated by the study subjects.

Diagnostic tests (2X2 tables)
Gold standard layout

	Test result		
	Positive	**Negative**	
Disease positive	a	b	m_1
Disease negative	c	d	m_0
	n_1	n_0	n

Note The marginals are the same as for risk calculations; the inner cell values are denoted as a, b, c, d.

Test comparison layout

	Test 2 positive	**Test 2 negative**	**Total**
Test 1 positive	n_{11}	n_{12}	$n_{1.}$
Test 1 negative	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Correlated data
Matched-pair case-control data layout

		Control pair		**Case totals**
		Exposed	**Non-exposed**	
	Exposed	t	u	$t+u = a_1$
Case pair	**Non-exposed**	v	w	$v+w = a_0$
	Control totals	$t+v = b_1$	$u+w = b_0$	

Note If pair-matching is used in a cohort study, the same format is used but the case (rows)-control(columns) status is replaced by exposed (rows) non-exposed (columns) and the exposure status is replaced by disease status.

Significant digits
Throughout the text, data are often presented with more significant digits than normally would be warranted. This is done for clarity and to avoid rounding errors.

GT. 2 Multivariable models

In general, multivariable models will be presented as follows, with explicit subscripting (*eg* for observation number) used only if absolutely necessary for clarity:

$$\text{outcome} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where the outcome may be a variety of parameters (*eg* for logistic regression outcome = $\ln(p/1\text{-}p)$ and k is the number of parameters in the model (excluding the intercept).

In some situations, βX or μ will be used to represent the entire right-hand side of the model (*ie* the linear predictor) to simplify presentation:

$$\beta X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

The terms predictor, exposure, risk factor and independent variable will all be used to designate factors that 'cause' the outcome of interest, although in general we prefer to use one of the first

two terms. These will be designated *X*.

The terms outcome and dependent variable will both be used for the response, but the former term is used most commonly. These will be designated *Y*.

GT. 3 Multilevel models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{fam(i)} + \varepsilon_i$$

Note For the sake of simplicity, a single index notation will be used for all multilevel data. The subscript *i* denotes the individual (lowest level) observation. In the example above, $u_{fam(i)}$ refers to the family containing the i^{th} individual. If there are 40 families in the study, *u* could have one of 40 values. An alternative notation, used in some texts, has multiple indices such as $u_j + i_j$ where *j* refers to the family and *i* to the i^{th} individual in the j^{th} family. We will use this notation for repeated measures data where Y_{ij} = measurement for subject *i* at time *j*.

GT. 4 Glossary

Terms related to formulae and methods

a	number of cases
ACF	autocorrelation function
AF_e	attributable fraction in the exposed group
AF_p	attributable fraction in the population
AFT	accelerated failure time
AIC	Akaike's Information Criteria
ALR	alternating logistic regression
ANOVA	analysis of variance
AP	apparent prevalence
AR	autoregressive
ARMA	autoregressive moving average
AUC	area under ROC curve
BIC	Bayesian Information Criteria (Schwartz Bayesian Criteria)
BLUP	best linear unbiased predictor
BUGS	Bayesian analysis using Gibbs sampling
c	constant (*eg* baseline hazard)
c	cost of sampling
c	Geary's c for correlation between values at pairs of spatial points
c	rate of contacts an animal makes with other animals in one time period
CAR	conditional autoregressive
CCC	concordance correlation coefficient
chi^2	chi-square (χ^2)
CI	confidence interval
corr (*Y*)	correlation matrix of *Y*
cov (*Y*)	covariance matrix of *Y*
covar	covariance

covar (+)/covar(-)	covariance in test positive (+)/negative (-) sample results
cp	cutpoint
Cp	Mallow's statistic
cp	rate of 'effective' contacts
CRS	composite reference standard
CV	coefficient of variation
D	deviance statistic (-2*lnL)
D	minimum number diseased
D	duration
D	disease
d	duration = duration of the infectious period
d	standardised difference in propensity scores
D-	subjects not having a specified disease/condition
D'	classified (not necessarily correct) disease state
D(h)	difference function (between K-functions for cases and controls)
D+	subjects having a specified disease/condition
DAC	directed acyclic graph (aka cusal diagram)
DB	delta-beta
deff	design effect
df	degrees of freedom
DFITS	difference in fit statistic
d_i	Cohen's *d* for study *i* (meta-analysis)
DIC	deviance information criterion
$d_{ii'}$	distance between point *i* and *i'*
d_j	outcome events (failures) during the interval (actuarial life table) or number of events at time t_j (K-M life table)
DOR	diagnostic odds ratio
e	2.71828 (natural number)
E	expected value (*eg* $E(Y)$ = expected value of Y)
E	exposure factor
E-	subjects not exposed
E+	exposed subjects
ES	effect size
ESS	effective sample size
EV	extraneous variable
exp	expected cell number
exp	exponential function (*ie* $\exp(x) = e^x$)
f	proportion of population vaccinated
$F(t)$	failure function
$f(t)$	probability density function
$f(\theta)$	prior distribution for θ (Bayesian analysis)
$f(\theta\|Y)$	posterior distribution for θ (Bayesian analysis)
FNF	false negative fraction

FP	fractional polynomial
FPC	finite population correction
FPF	false positive fraction
G^2	likelihood ratio statistic
GEE	generalised estimating equations
g_i	Hedge's adjusted g for study *i* (meta-analysis)
GLM	generalised linear model
GLMM	generalised linear mixed model
GWR	geographically weighted regression
$h(t)$	hazard function
$H(t)$	cumulative hazard function
$h_0(t)$	baseline hazard function
h_i	leverage
H*j*	distribution of host factor and/or subject time in stratum j
HR	hazard ratio
Hs	standard population distribution of host factor
HSe	herd sensitivity
HSp	herd specificity
i	observation counter
I	incidence rate
I	Moran's I (spatial autocorrelation coefficient)
i	incidence = rate at which new infections are occurring in the population - this is the population incidence rate (designated *i* to differentiate it from *I*)
IC	information criteria
ICC	intra-class correlation coefficient
ID	incidence rate difference
ID_G	incidence rate difference based on group means
I_{dir}	directly standardised rate
I_e	expected incidence rate
I_h	indicator for K-function
IIA	independence of irrelevant alternatives
I_{ind}	indirectly standardised rate
IPTW	inverse probability of treatment weighted
IQR	interquartile range
IR	incidence rate ratio
IR_G	incidence ratio based on group-level data
Is	standard population incidence rates
j	designated for strata
j	designator for categories
j	designator for covariate patterns in a dataset
j	designator for time intervals (actuarial life table) or time points (KM life table)
j	sampling interval in systematic random sample
J	total number of *j*

k	cutpoint for herd-level testing (number of positives required for positive herd classification)
k	number of predictors in a model (not including intercept)
k	number of spatial clusters or groups
k	number of studies in a meta-analysis
$K(h)$	K-function for spatial density of events per distance h
$K(ht)$	bivariate space-time K-function
KM	Kaplan-Meier (life table or survival model)
L	allowable error (one-half the length of a confidence interval)
L	likelihood function (*eg* $L(Y\|\theta)$
L	lag-time in case-crossover studies
L_0	null or baseline likelihood function
LCM	latent class model
L_{full}	likelihood function from full model
LISA	local indicator of spatial association
l_j	subjects at risk of failure at the start of the time interval (actuarial life table)
ln	natural log
lnL	ln (likelihood function)
log	natural log (also ln)
LR	likelihood ratio
LR_{cat}	likelihood ratio for defined category of result
LR_{cp}	likelihood ratio based on defined cutpoint(s)
L_{red}	likelihood function from reduced (smaller) model
LRT	likelihood ratio test
m	number of matched controls per case
m	number of observations in a covariate pattern
m	number of samples in a pooled sample
m	number of subjects per cluster (group)
MANOVA	multivariate analysis of variance
MAR	missing ar random
MAUP	modifiable areal unit problem
MCA	multiple correspondence analysis
MCAR	missing completely at random
MCMC	Markov chain Monte Carlo
MCSE	Monte Carlo standard error
MD_i	mean difference in study i in a meta analysis
ML	maximum likelihood
MM	method of moments
MNAR	missing not at random (also NMAR)
MOR	median odds ratio
MOR_c	cluster median odds ratio
MQL	marginal quasi-likelihood
MSE	mean square error

n	number
n	sample size
N	population size
n'	adjusted sample size
NB-1 etc	negative binomial models – see Chapter 18 for details
o	odds
obs	observed cell number
OD	optical density
OR	odds ratio
OR(ABC)	odds ratio for factor ABC
OR(ABC|D)	odds ratio for factor ABC conditional on D
OR_a	odds ratio - adjusted
OR_c	odds ratio - crude
OR_j	stratum-specific odds ratio
OR_{MH}	Mantel-Haenszel adjusted odds ratio
OR_{sf}	odds ratio of sampling fractions
p	probability as in p(D+|E+) or p(Y=1)
p	proportion as in 1n(p/1-p)
p	shape parameter for Weibull distribution
p	probability of transmission of the infection if one animal is infectious and one is susceptible
p'	classified (not necessarily correct) proportions with exposure or disease
p_j	probability of surviving interval j (actuarial life table) or survival at time t_j (KM life table)
P	P-value
P	prevalence
PA	population average
PACF	partial autocorrelation function
par	population at risk
par	parameter
PAR	population attributable risk
PD	prevalence difference
PE	prediction error
pl(λ)	profile likelihood function
PlSe	pooled-sample sensitivity
PlSp	pooled-sample specificity
PPV-	positive predictive value of a negative test
PQL	penalised quasi-likelihood
PR	prevalence ratio
PS	propensity score
PSU	primary sampling unit
PV	predictive value
PV-	negative predictive value

$PV+$	positive predictive value
q	$1-p$
q_j	risk of event during interval j (actuarial life table) or at time t_j (KM life table)
Q	Cochrane's Q statistic
QIC	quasi-likelihood under the independence model information criterion
r	correlation coefficient (ρ also used)
R	incidence risk
R	spatial region
R_0	R_0 = basic reproductive number = # of new cases that arise from an infectious individual in a completely susceptible population.
r^2	squared correlation (R^2 also used)
R^2	coefficient of determination (r^2 also used)
RCT	randomised controlled trial
RD	risk difference (also know as attributable risk)
RDD	random digit dialling
REML	restricted maximum likelihood
res_p	Pearson residual
r_i	raw residual
r_j	average number of subjects at risk during a time interval (actuarial life table) or at time t_j (KM life table)
ROC	receiver operating characteristics
RR	risk ratio (alternatively known as relative risk)
RR_a, RR_u	adjusted and unadjusted RR (meta-analysis)
Rs	standard population incidence risk
r_{si}	standardised residual
R_t	R_t = effective reproduction number = # of new cases arising from each infectious individual at time t.
r_{ti}	studentised residual
$s = S/N$	proportion of the population that is susceptible **Note**: in a completely susceptible population $S_0=N$ so $s_0=1$
S, I, R, N	the numbers of susceptible, infectious, removed and total number of animals in the population, respectively
$S(t)$	survivor function
SAR	Simultaneous autoregressive
SD	standard deviation
SE	standard error
Se	sensitivity
Se_{corr}/Sp_{corr}	corrected Se/Sp based on cross-sectional validation
Se_{new}/Sp_{new}	Se/Sp of current test adjusted for Se/Sp of referent test
Se_p/Sp_p	Se/Sp in parallel interpretation of test results
Se_s/Sp_s	Se/Sp in series interpretation of test results
sf	sampling fraction
sf_{T+}/sf_{T-}	sampling fractions for cross-sectional validation
S_i	value of latent variable for individual i

s_i	spatial points
SMR	standardised morbidity/mortality ratio
so	sampling odds
Sp	specificity
sr	sampling risk
SRR	standardized risk ratio
SS	subject specific
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
t or T	animal-time
TCE	true causal effect
t_d	doubling time
T_i	study outcome for study i in meta-analysis
t_j	time of event (KM life table)
t_{j-1},t_j	time span in the interval (actuarial life table)
Δt	length of period
TP	true prevalence
TR	time ratio
Ts	standard population animal-time at risk
T_{scan}	spatial scan statistic
TVC	time-varying covariate
U	measure of confounding bias (meta-analysis)
u_i	random effect of study i
var	variance
V_i	within study variance for study i in meta-analysis
VIF	variance inflation factor
W	sampling weights based on probability of exposure
w_j	subjects withdrawn during interval (censored observations) (actuarial life table) or censored observations at time t_j (KM life table)
X	predictor variable or design matrix of predictors
Y	outcome variable or vector of outcome values
Z	design matrix for random effects
Z	extraneous variable, factor or confounder
Z	standard normal deviate
Z_α	standard normal percentile for $\alpha/2$ Type I error (for sample size calculations)
Z_β	standard normal percentile for one-tailed β Type II error (for sample size calculations)

Note Acronyms are not italicised in Arial font (tables and figures).

Symbols

*	multiplication symbol
/	division
#	number
~	approximate symbol or distributed as (*eg* Y~N(0,1))

$\approx$	approximately equal to
α	level of significance (Type I error)
β	regression coefficient or vector (1*n) of coefficients
β	Type II error (power=1-β)
β	frailty factor
β_{aft}	coefficient from accelerated failure time model
β_{ph}	coefficient from proportional hazards model
γ	prior (spatial) disease rate
$\gamma(h)$	empirical semi-variogram
δ	spatial edge correction factor
Δ_i	Glass's Δ for study i (meta-analysis)
ε	error (or vector (1*n) of error values
θ	posterior value for local (spatial) disease rate
θ	a specified or assumed value
θ_0	null specified value
λ	kernel density
λ	hazard
λ	rate at which susceptible animals becomes infectious
λ	power transformation
μ	random group effect
μ	mean
π	3.14159 (natural number)
ρ	correlation - intra-class correlation coefficient (r also used)
ρ_{ce}	confounder-exposure correlation
σ	standard deviation
σ^2	variance
σ^2_h	herd variance
σ^2_i	random slope variance for β_1
σ^2_r	regional variance
τ	spatial bandwidth
τ	cutpoint for proportional odds
τ	distribution of survival times
τ^2	between study variance in meta-analysis
φ	dispersion parameter in GLM(M)
φ	variance of prior disease rate
χ^2	chi-square statistic
χ^2_{homo}	χ^2 test for homogeneity
χ_{Wald}	Wald chi statistic

Terms related to time, location and specific health problems

AID	autoimmune disease
BC	British Columbia (Canada's most westerly province)
bp	blood pressure
bwt	birth weight
CRD	childhood respiratory disease
d	day(s)
EIA	enzyme immunoassay
ELISA	enzyme-linked immunosorbent assay
HIV	human immunodeficiency virus
HPV	human papilloma virus
IFAT	indirect fluorescent antibody test
mi	mycardial infarction
MMR	measels, mumps and rubella
mo	month(s)
nv	norovirus
Ont.	Ontario (large province in Canada)
PCR	polymerase chain reaction
PEI	Prince Edward Island (smallest province in Canada)
ppb	parts per billion
ppm	parts per million
RSV	respiratory syncytial virus
SARS	severe acute respiratory syndrome
STREP	*Streptococcus pneumoniae*
VI	virus isolation
yr	year(s)

GT. 5 Probability notation

E(Y) = expected value of Y

p(D+) = probability of having the disease of interest

p(T+|D+) = probability of being test positive given the animal had the disease of interest

p(D+|E+) = probability of having the disease of interest in an exposed group

p(D+|T+) = probability of having the disease of interest given the animal was test positive

c^n_k = the number of combinations of k items from n items

GT. 6 Naming variables

Variable names in the text will be set between pairs of dashes (*eg* -varname-) but the dashes will not be included in tables and figures or if the variable is used in an equation.

Modifications of variables will generally (but not always – you wouldn't expect us to be totally consistent, would you?) be named by adding a suffix to the original variable name. For example:

varname_ct	centred version of the variable
varname_sq	squared version of the variable
varname_c#	a categorical version of -varname- with n = # categories
varname_ln	log transformed version of the variable

Indicator variables will usually be named by appending the category value (or left-hand end of the category range if it is a continuous variable). For example, a variable representing birth weight (-bwt-) broken into four categories (0-2499, 2500-2999, 3000-3499, 3500+) would result in the following four variables:

-meduc_0-
-meduc_2500-
-meduc_3000-
-meduc_3500-

Note Unless otherwise specified, values falling exactly on the dividing point will fall in the upper category.

INDEX

A

B

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

XYZ