

Statistical Modelling

Proportions, odds, risk and hazard ratios

Søren Lophaven

Agenda

- Continuous outcomes (t-test, Analysis of Variance, Linear regression, Analysis of Covariance)
- Binary outcomes (Logistic regression)
- Time-to-event outcomes (Cox proportional hazard model)
- Poisson regression

Statistical analysis of continuous outcomes

The basic model

$$y \sim f(x_{primary}, x_i)$$

- y : continuous outcome variable
- $x_{primary}$: explanatory variable of primary interest (categorical or continuous)
- x_i : other explanatory variables that we adjust for (categorical or continuous)

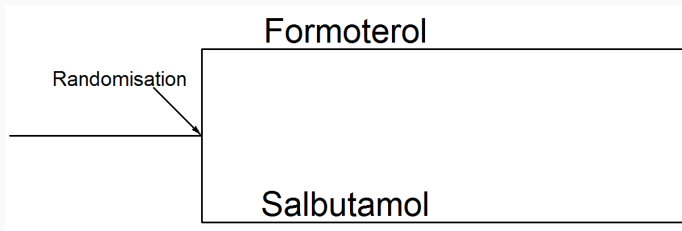
Drug treatments of asthma

- Chronic condition
- Abnormal degree of wheezing or breathlessness
- Severity of symptoms ebbs and flows over time
- Outcome: Maximum rate at which the patient is able to exhale a fixed time after administration of the drug (PEF)
- Comparison of the two drugs Salbutamol and Formoterol for short-term relief of the symptoms

PEF data

Formoterol	Salbutamol
310	270
310	260
370	300
410	390
250	210
380	350
330	365
385	370
400	310
410	380
320	290
340	260
220	90

Design of study



The analysis model

$$y \sim f(x_{\text{primary}}, x_i)$$

$$PEF \sim f(\text{treatment})$$

- The x -variable is categorical
- This is a so-called t-test

PEF data - some calculations

Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

Variance: $var(x) = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, measures the total squared average "distance" from the mean to the observations

Standard deviation: $SD = \sqrt{var(x)}$, measures the total average "distance" from the mean to the observations

Standard error: $SE = \frac{SD}{\sqrt{n}}$, measures the variability of the mean

Formoterol: Mean=341.2, variance=3559.32, SD=59.66

Salbutamol: Mean=295.8, variance=6865.78, SD=82.86

PEF data - some calculations for the difference

Mean: $341.2 - 295.8 = 45.4$

Pooled variance: $(3559.32 + 6865.78)/2 = 5212.84$

Standard error: $SE = \sqrt{\frac{2 \cdot \text{pooled variance}}{n}} = \sqrt{\frac{2 \cdot 5212.84}{13}} = 28.3$

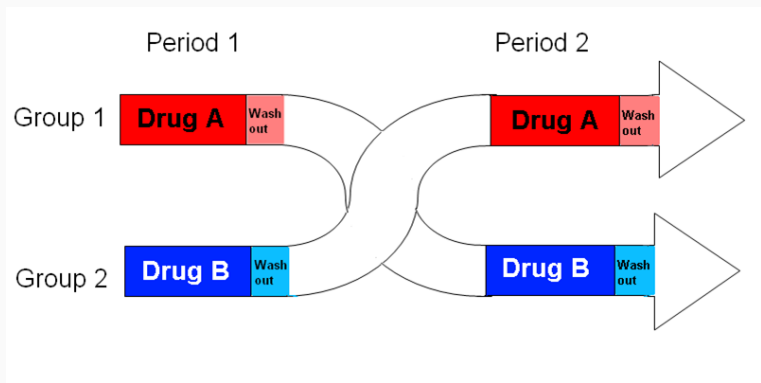
95 % confidence interval for the difference:

$$\text{mean} \pm 2 \cdot SE = [-11.3; 102.0]$$

PEF data

Subject ID	Formoterol	Salbutamol	Difference
1	310	270	40
2	310	260	50
3	370	300	70
4	410	390	20
5	250	210	40
6	380	350	30
7	330	365	-35
8	385	370	15
9	400	310	90
10	410	380	30
11	320	290	30
12	340	260	80
13	220	90	130

Design of study



The analysis model

$$y \sim f(x_{\text{primary}}, x_i)$$

$$PEF \sim f(\text{treatment}, \text{subject})$$

- Both x-variables are categorical
- This is a paired t-test

PEF data - revisited

Mean: 45.4

SD: 40.59

Standard error: $SE = \frac{SD}{\sqrt{n}} = \frac{40.59}{\sqrt{13}} = 11.26$

95 % confidence interval for the difference:

$$mean \pm 2 \cdot SE = [22.9; 67.9]$$

Analysis of Variance (ANOVA)

- The x -variables are categorical
- The x -variable of primary interest has more than two levels
- Overall F-test for significance
- Pairwise tests for significance
- Analysis of means

Linear regression

$$y \sim f(x_{\text{primary}}, x_i)$$

- All x -variables are continuous

FEV1 in children - first 10 of 654 observations

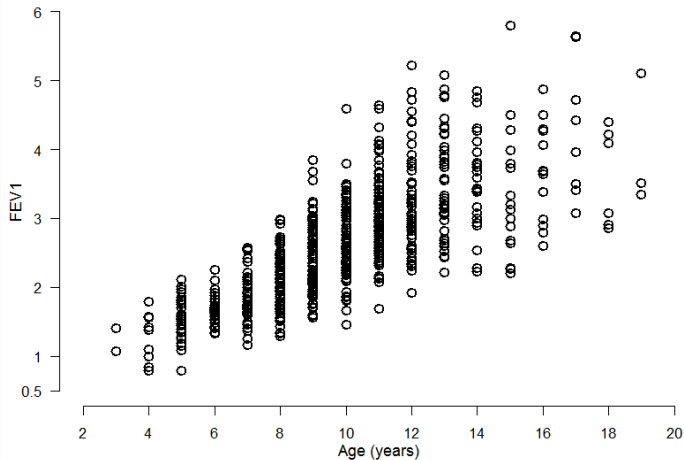
ID	Age	FEV	Height	Sex	Smoker
301	9	1.708	57	Female	Non
451	8	1.724	67.5	Female	Non
501	7	1.72	54.5	Female	Non
642	9	1.558	53	Male	Non
901	9	1.895	57	Male	Non
1701	8	2.336	61	Female	Non
1752	6	1.919	58	Female	Non
1753	6	1.415	56	Female	Non
1901	8	1.987	58.5	Female	Non
1951	9	1.942	60	Female	Non

FEV1 in children - The influence of age

$$y \sim f(x_{\text{primary}}, x_i)$$

$$FEV1 \sim f(\text{age})$$

FEV1 in children - The influence of age



FEV1 in children - The influence of age

```
summary(lm(FEV~Age, data=fev))
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.431648	0.077895	5.541	4.36e-08	***
Age	0.222041	0.007518	29.533	< 2e-16	***

FEV1 in children - Descriptive statistics by smoking status

	Non-smoker	Smoker
N	589	65
Mean	2.57	3.28
SD	0.85	0.75
Median	2.47	3.17
Min	0.79	1.69
Max	5.79	4.87

- Why do the smokers have higher lung capacity than non-smokers?
- Is it a chance finding?

Analysis of Covariance (ANCOVA)

$$y \sim f(x_{\text{primary}}, x_i)$$

- Both continuous and categorical x -variables

FEV1 in children - The influence of age and smoking status

```
summary(lm(FEV~Age+Smoker+Age:Smoker, data=fev))
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.19697	0.40596	5.412	8.78e-08	***
Age	0.07986	0.02959	2.699	0.00713	**
SmokerNon	-1.94357	0.41428	-4.691	3.31e-06	***
Age:SmokerNon	0.16270	0.03074	5.293	1.65e-07	***

What did we learn so far?

- For continuous outcomes the basic analysis model is the same
- Method of analysis named according to the types of x -variables (t-test, ANOVA, linear regression, ANCOVA)
- Ordinary least squares used for estimation
- Estimation of means, mean changes, mean differences
- The general linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

SAS terminology

- `proc ttest`, `proc anova`, `proc reg`
- The common procedure is `proc glm`

Statistical analysis of binary outcomes

Binary outcomes

- Two possible outcomes (e.g. dead/alive, response to treatment/no response)
- Could in general be coded as $Y=0$ or $Y=1$
- What about $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$?
- We are interested in $\text{Prob}(Y=1)$
- The fitted $\text{Prob}(Y=1)$ will be outside $[0;1]$ for some combinations of x -variables
- Something needs to be done

Logistic regression

Cox (1958), Walker and Duncan (1967):

$$Prob(Y = 1|X) = \frac{1}{[1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)]}$$

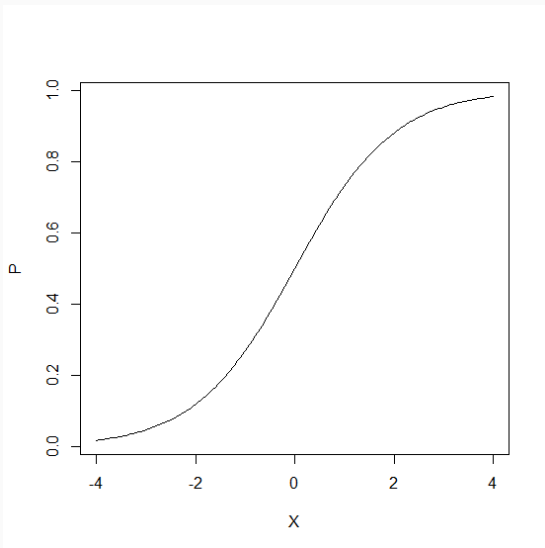
$$\begin{aligned} \text{logit}(Y = 1|X) &= \text{logit}(P) = \log(P/1 - P) = \log(\text{odds}) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \end{aligned}$$

A generalized linear model (in SAS: proc genmod, proc logistic)

D.R. Cox. The regression analysis of binary sequences (with discussion). Journal of the Royal Statistical Society B

S.H. Walker and D.B. Duncan. Estimation of the probability of an event as a function of several independent variables

The logistic function



Montgomery and Åsberg Depression Rating Scale (MADRS)

- MADRS: rating scale designed to assess the severity of depressive symptoms
- Based on a clinical interview
- Administered by trained psychiatrists
- MADRS total score is the sum of the score of the 10 individual items
- Symptoms rated on 7-point scales from 0 (no symptom) to 6 (severe symptom) with detailed anchor points
- MADRS total score goes from 0 to 60
- Response to treatment: at least 50 % reduction from baseline

Two MADRS items

1. Apparent sadness

Representing dependency, gloom and despair, (more than just ordinary transient low spirits) reflected in speech, facial expression, and posture. Rate by depth and inability to brighten up.

- 0 No sadness.
- 1
- 2 Looks dispirited but does brighten up without difficulty.
- 3
- 4 Appears sad and unhappy most of the time
- 5
- 6 Looks miserable all the time. Extremely despondent.

4. Reduced sleep

Representing the experience of reduced duration or depth of sleep compared to the subject's own normal pattern when well.

- 0 Sleeps as usual.
- 1
- 2 Slight difficulty dropping off to sleep or slightly reduced, light or fitful sleep.
- 3
- 4 Sleep reduced or broken by at least two hours.
- 5
- 6 Less than two or three hours sleep.

Treatment response at week 8 in depression study

Treatment	Responders			OR [95% CI]	p-value
	N	n	(%)		
Placebo	165	54	(32.7)		
Low dose	174	88	(50.6)	2.10 [1.35 ; 3.27]	0.0009
High dose	181	119	(65.7)	3.94 [2.52 ; 6.16]	<.0001

- Adjustment for treatment and MADRS total score at baseline
- β_j is the change in log odds per unit change in X_j
- $\exp(\beta_j)$ is the relative change in odds, i.e. the odds ratio

Probability and odds

- Probability is intuitive (range from 0 to 1)
- Odds are not intuitive (range from 0 to ∞)
- The probability of the event occurring divided by the probability of the event not occurring, i.e.
probability / (1 - probability)
- Probability of an event occurring is 0.05, then the odds of that event is $0.05 / (1 - 0.05) = 0.053 = 1 / 19$

Odds ratio and risk ratio

- Probability of outcome is 0.66 in treatment group one and 0.33 in the second treatment group
- Treatment effect: The first group is twice as likely to have the outcome as the second group

$$RR = \frac{p_1}{p_2} = 2$$

- Odds ratio = the ratio of two odds

$$OR = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 \cdot (1 - p_2)}{p_2 \cdot (1 - p_1)} = \frac{0.66 \cdot (1 - 0.33)}{0.33 \cdot (1 - 0.66)} = 4$$

- Does not mean that the outcome is 4 times as likely in the first group as in the second group
- Odds of the outcome is 4 times higher in the first group than the second group

Why is odds ratios so frequently reported?

- Estimated by logistic regression
- Logistic regression always produce probabilities in the range 0 to 1
- Odds have an unlimited range, and any positive odds ratio will still yield a valid probability (RR=2 can only apply to probabilities below 0.5)
- Convergence problems are rare with logistic regression

Not all are happy about odds ratios

Annals of Internal Medicine

ARTICLE

Infection Risk with Nitrofurazone-Impregnated Urinary Catheters in Trauma Patients

A Randomized Trial

- A clinical trial to determine whether nitrofurazone-impregnated urinary catheters reduce the incidence of urinary tract infections in patients that were admitted directly from the accident scene to the Trauma Center in Copenhagen
- The primary endpoint was the proportion of patients developing an urinary tract infection after surgery
- Analyzed using a logistic regression model with treatment group and gender as explanatory factors
- Reviewer's comment: "We ask that you use log-binomial models because events were common, so ORs from a logistic regression model will overstate risk estimates"

What is the problem?

- Risk ratios has a natural interpretation. The odds ratios has not
- Many examples of researchers misinterpreting the odds ratio as a risk ratio
- Holcomb et al. (2001) report that 26% of authors in top-tier medical journals explicitly misinterpret odds ratios as though they were risk ratios
- Thomas Lumley: *Most people don't have any feel for the meaning of an odds ratio (and those that do mostly can't communicate it to those who don't)*
- If the outcome is common ($> 10\%$ with event) the odds ratio becomes a poor approximation of the relative risk
- In this case the odds ratio overestimates the relative risk

Example (Schulman et al., 1999)

Primary care physicians were shown videotaped interviews with an actor portraying a patient with chest pain, and given data on cardiac risk factors and results of thallium stress test. There were 18 scenarios, and each was portrayed by 8 actors (race \times sex \times age combinations). The outcome was whether the doctors recommended referral to cardiac catheterization.

	Odds ratio 95 % CI
White	1
Black	0.6 (0.4-0.9)
Men	1
Women	0.6 (0.4-0.9)

So the odds of referral were 40% lower for women than for men and for blacks than for whites.

Example (Schulman et al., 1999)

Nightline "In our main analysis we found that blacks were 40 % less likely to be referred for cardiac catheterization compared to whites"

USA Today "Heart Care Reflects Race and Sex, not Symptoms"

Washington Post "Doctors are far less likely to recommend sophisticated cardiac tests for blacks and women than for white men with identical complaints"

LA Times "Authors suggest the differences are the consequences of race and sex bias"

NY Times "Doctors are only 60 % as likely to order cardiac catheterization for women and blacks as for men and whites"

Example (Schulman et al., 1999)

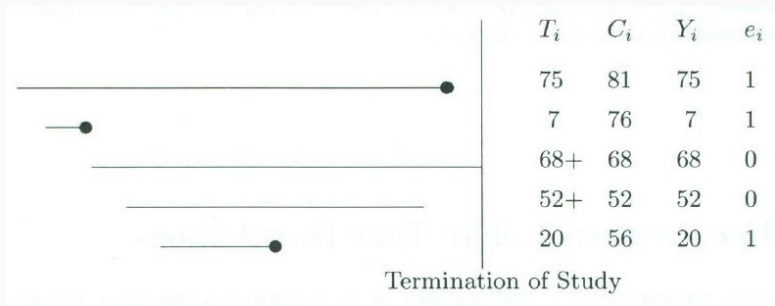
- After many readers pointed out that a relative risk of 0.93 isn't a 40 % reduction the New England Journal of Medicine and many of the news sources carried a revision of the story. The original article is still being cited as evidence of widespread and serious racial bias in medicine.
- This was a particularly public and embarrassing example, but the underlying problem is much more common.

Statistical analysis of time-to-event outcomes

Logistic regression or survival analysis?

- If time until the occurrence of the event is not important → binary endpoint → logistic regression
- If time until the occurrence of the event is important → time until event endpoint → survival analysis (Cox proportional hazards model)

Censoring

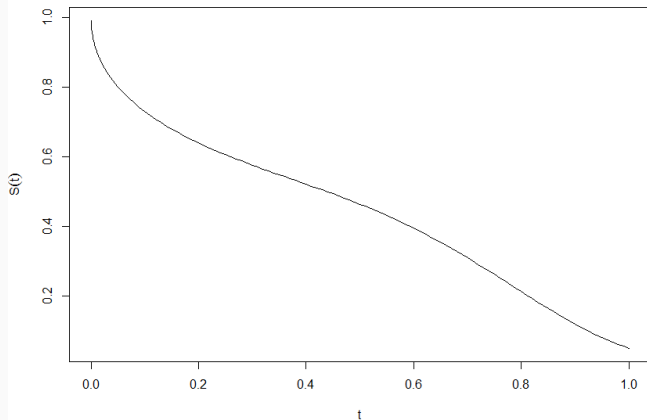


- Censored data: The time to event is only known to exceed some time t
- No responses censored \rightarrow standard regression models could be applied

Survival function

The probability that an event occurs after some time t :

$$S(t) = \text{Prob}\{T > t\}$$

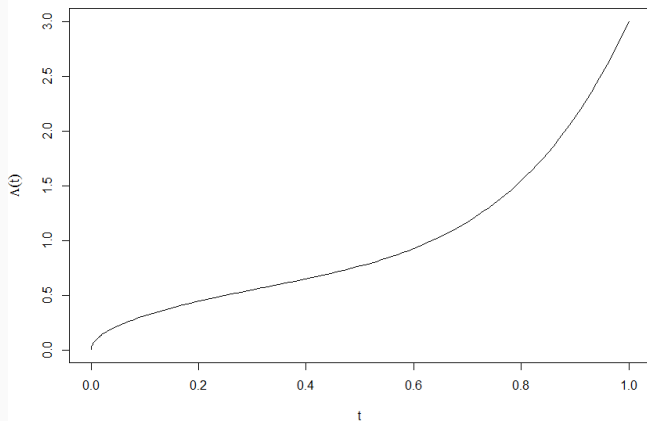


Survival function

- $S(t) = 1$ for $t = 0$, i.e. all subjects survived at least to time zero
- $S(t)$ is nonincreasing as t increases
- Very high risk of event in the beginning
- $S(t)$ flat for $0.1 \leq t \leq 0.6$ \longrightarrow low risk of event
- For $t > 0.6$ the risk increases again

Cumulative hazard function

Corresponding to survival function in previous slide

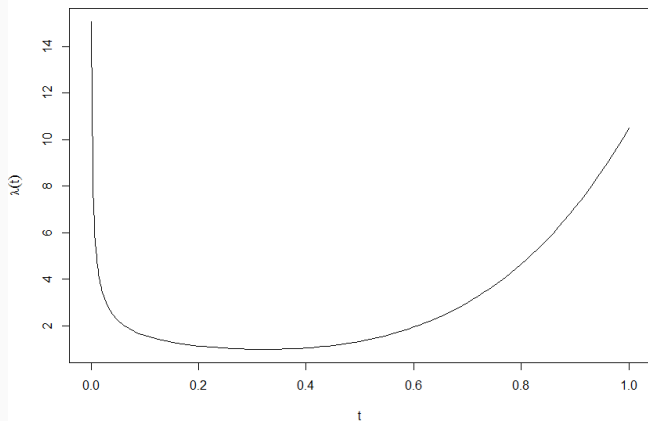


Cumulative hazard function

- $\Lambda(t)$ describes the accumulated risk up until time t
- $\Lambda(t)$ is nondecreasing as t increases
- $\Lambda(t) = -\log\{S(t)\}$

Hazard function

Corresponding to survival function in previous slide



Hazard function

- The hazard at time t is the probability that the event will occur in a small interval around t given that the event has not occurred before time t

The Kaplan-Meier estimator

- Estimation of survival function $S(t)$
- $S_{km}(t) = \prod_{i:t_i < t} (1 - d_i/n_i)$
- d_i : number of events at time t_i
- n_i : number at subjects at risk at time t_i

The Kaplan-Meier estimator

Event times: 1, 3, 3, 6⁺, 8⁺, 9, 10⁺

i	t_i	n_i	d_i	$(n_i - d_i)/n_i$
1	1	7	1	6/7
2	3	6	2	4/6
3	9	2	1	1/2

$$\begin{aligned}S_{km}(t) &= 1, \quad 0 \leq t < 1 \\&= 6/7 = 0.85, \quad 1 \leq t < 3 \\&= (6/7)(4/6) = 0.57, \quad 3 \leq t < 9 \\&= (6/7)(4/6)(1/2) = 0.29, \quad 9 \leq t < 10\end{aligned}$$

The Cox proportional hazards model

Stated in terms of the hazard function:

$$\lambda(t|X) = \lambda(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$$

- No assumption about specific shape of $\lambda(t)$
- No interest in estimating it

The Cox proportional hazards model

- The form of the true hazard is unknown \rightarrow Cox PH does not assume any particular form for the hazard function
- The form of the true hazard is usually not of primary interest \rightarrow Cox PH allows the analyst to ignore the hazard function
- In the Cox PH the explanatory variables are linearly related to the hazard function
- Hazard ratio = risk ratio obtained by exponentiating parameter estimates
- The survival functions for subjects with different values of X are powers of each other

Evaluating the proportional hazards assumption

Three graphical approaches:

- Plot survival curves for different levels of an explanatory variable X . The curves should not cross each other at any point in time
- Plot the cumulative hazard functions for different levels of an explanatory variable X . For all time points one should get from the lower curve to the upper by multiplication with the same constant:
 $\Lambda(\text{level}1) = C \cdot \Lambda(\text{level}2)$.
- Plot the logarithm of the cumulative hazard functions for different levels of an explanatory variable X . The curves should be parallel, i.e. for all time points one should get from the lower curve to the upper by adding the same constant: $\log(\Lambda(\text{level}1)) = C + \log(\Lambda(\text{level}2))$.

Night shift work and incidence of diabetes in the Danish Nurse Cohort

Objective:

- To study the association between shift work and incidence of diabetes in Danish nurses

Methodology:

- Data from the Danish Nurse Cohort with 28731 participating female nurses recruited in 1993 (19898) or 1999 (8833)
- Self-reported baseline information on diabetes prevalence, lifestyle and working time
- Self-reported whether they worked night, evening, rotating or day shifts
- Followed-up in the Danish Diabetes Register for incidence of diabetes until 2013
- The association between working time and diabetes incidence was analysed using a Cox proportional hazards model adjusted for diabetes risk factors, separately with and without adjustment for body mass index (BMI) which might be an intermediate variable
- 19873 nurses worked and were diabetes-free at recruitment

Night shift work and incidence of diabetes in the Danish Nurse Cohort

	Total N=19 837	No diabetes N=19 000	Diabetes N=837
Person-years	308 078	298 541	9536
Baseline age, mean (SD)	51.4 (5.4)	51.4 (5.3)	53.6 (5.8)
Shifts			
Day shifts, n (%)	12 414 (62.4)	11 905 (62.6)	509 (58.3)
Evening shifts, n (%)	2022 (10.1)	1918 (10.1)	104 (11.9)
Night shifts, n (%)	1098 (5.5)	1014 (5.3)	84 (9.6)
Rotating shifts, n (%)	4339 (21.8)	4163 (21.9)	176 (20.1)
BMI			
Underweight (<18.5 kg/m ²), n (%)	401 (2.0)	393 (2.1)	8 (0.9)
Normal (18.5–25 kg/m ²), n (%)	14 124 (71.0)	13 775 (72.5)	349 (39.9)
Overweight (25–30 kg/m ²), n (%)	4320 (21.7)	3991 (21.0)	329 (37.7)
Obese (>30 kg/m ²), n (%)	1028 (5.1)	841 (4.4)	187 (21.4)
Smoking status			
Never smoked, n (%)	7185 (36.1)	6891 (36.2)	294 (33.7)
Currently smoker, n (%)	6854 (34.4)	6501 (34.2)	353 (40.4)
Previously smoked, n (%)	5834 (29.3)	5608 (29.5)	226 (25.9)
Smoking intensity* (g/day), mean (SD)	7.7 (9.0)	7.6 (8.9)	9.4 (10.2)
Alcohol consumption (g/week), mean (SD)	119.5 (125.8)	119.7 (124.8)	116.0 (144.6)

Night shift work and incidence of diabetes in the Danish Nurse Cohort

Physical activity			
Low, n (%)	1057 (5.3)	972 (5.1)	85 (9.8)
Medium, n (%)	13 105 (65.9)	12 510 (65.9)	595 (68.1)
High, n (%)	5711 (28.7)	5518 (29.0)	193 (22.1)
Consume vegetables and fruit regularly, n (%)	19 452 (97.8)	18 601 (97.9)	849 (97.2)
Avoid fatty meat, n (%)	18 097 (91.0)	17 337 (91.2)	760 (87.0)
Hypertension, n (%)	1970 (9.9)	1734 (9.1)	236 (27.1)
AMI, n (%)	60 (0.3)	54 (0.3)	6 (0.7)
Employment status			
Employed, n (%)	19 725 (99.2)	18 873 (99.3)	852 (97.5)
Retired, n (%)	99 (0.5)	83 (0.4)	16 (1.8)
Unemployed, n (%)	6 (0.03)	5 (0.03)	1 (0.11)
In rehabilitation, n (%)	43 (0.2)	39 (0.2)	4 (0.4)
Marital status			
Married, n (%)	14 827 (74.6)	14 218 (74.8)	609 (69.7)
Separated, n (%)	359 (1.8)	345 (1.8)	14 (1.6)
Divorced, n (%)	2319 (11.6)	2197 (11.5)	122 (13.9)
Unmarried, n (%)	1553 (7.8)	1471 (7.7)	82 (9.3)
Widow, n (%)	815 (4.1)	769 .0)	46 (5.2)

*In ever smokers.

AMI, acute myocardial infarction; BMI, body mass index.

Night shift work and incidence of diabetes in the Danish Nurse Cohort

Work shift	Person years Years	Cases N	Crude* model HR (95% CI)	Fully adjusted† model HR (95% CI)	Fully adjusted† model+BMI HR (95% CI)
Day shifts	193 624	509	1.00	1.00	1.00
Evening shifts	31 619	104	1.24 (1.01 to 1.53)	1.21 (0.98 to 1.50)	1.29 (1.04 to 1.59)
Night shifts	17 118	84	1.84 (1.46 to 2.31)	1.73 (1.37 to 2.19)	1.58 (1.25 to 1.99)
Rotating shifts	65 716	176	1.04 (0.87 to 1.23)	1.06 (0.89 to 1.26)	1.08 (0.91 to 1.28)

*Adjusted for age.

†Adjusted for age, smoking status, smoking intensity, physical activity, alcohol consumption (g/week), intake of fatty meat, marital status, employment status, acute myocardial infarction, hypertension, fruit and vegetables intake.

BMI, body mass index.

Poisson Regression

Statistical models for different outcomes

- The General Linear Model for continuous outcomes
- Logistic regression for binary outcomes
- Cox proportional hazards model for time-to-event outcomes
- Poisson regression for count data

The Poisson regression model

Count data given by a rate:

$$\log(\mu/t) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

- μ : expected number of events
- t : length of time at risk

$$\log(\mu) - \log(t) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \implies$$

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \log(t)$$

- Modelling the number of events
- $\log(t)$: offset. Allows for different time at risk for each subject

Effect of drug treatment on exacerbation rate in patients with COPD

Objective:

- Investigate the effect of drug treatment on exacerbation rate and pulmonary function in patients with chronic obstructive pulmonary disease (COPD)

Methodology:

- Primary endpoint: Mean rate of moderate to severe COPD exacerbations per patient per year
- Secondary endpoint: Time to first COPD exacerbation
- Active drug versus placebo
- Analysis of primary endpoint: Poisson regression model with adjustment for treatment, age, sex, smoking status, concomitant treatment with LABA, country and baseline post-bronchodilator FEV1
- Analysis of secondary endpoint: Cox proportional hazards model with adjustment for treatment, age, sex, smoking status, concomitant treatment with LABA and country

Results

Primary endpoint:

N	Active Rate [95% CI]	N	Placebo Rate [95% CI]	Risk ratio [95 % CI]	Ratio Active/placebo Change (%)	p-value
772	1.210 [1.074 ; 1.364]	796	1.485 [1.333 ; 1.655]	0.815 [0.710 ; 0.935]	-18.5	0.0035

Secondary endpoint:

Active n	Placebo n	Active	Patients with event		Cox PH model		
			%	Placebo	%	HR [95 % CI]	p-value
772	796	373	48.3	432	54.3	0.894 [0.778 ; 1.027]	0.1132

Poisson regression versus the Cox proportional hazards model

- For time to first event a Poisson model can be formulated to give similar results as the Cox proportional hazards model (the situation in which each cell of the cross-classification of person-time and events includes a single event)
- The Poisson model gives interpretable rate estimates by group
- No proportional hazards assumption for the Poisson model
- The Poisson model can handle subjects switching from exposed to not-exposed and vice versa several times in the study

Poisson regression versus the Cox proportional hazards model

Table 3 Comparison of estimated regression beta coefficients* and standard errors for brain cancer and cumulative magnetic field exposure in the electrical worker cohort

	Ungrouped Poisson	Proportional hazards	Grouped Poisson	
Model	β (SE)	β (SE)	β_{\dagger} (SE)	β_{\ddagger} (SE)
Model 1: age, exposure	0.0842 (0.0375)	0.0842 (0.0375)	0.1483 (0.0503)	0.1123 (0.0392)
Model 2: age, calendar time, race, exposure	0.0910 (0.0380)	0.0910 (0.0380)	0.1479 (0.0508)	0.1183 (0.0396)

