Class 3 Outline

1. Review of generalized linear models and regression
2. Review of linear regression
3. Review of logistic regression
4. Log-linear regression
5. Cox proportional hazards regression
6. Summary

0. Learning Objectives

- Review generalized linear models
- Review the assumptions, data structure and model for linear regression and logistic regression
- Introduce log-linear (Poisson) regression and Cox proportional hazards regression
- Critique and interpret regression analysis results in the published literature

1. Review of Generalized Linear Models

Generalized Linear Models (GLMs) provide a way to express the relationship between the response (Y) and explanatory variables (X’s):

Function of expected Y = $\beta_0 + \beta_1 * X_1 + \ldots + \beta_p * X_p$

The β’s, the “regression coefficients” express the relationships.

1.2 Review of Simple vs. Multiple GLMs

- *Simple:* One predictor variable (X) used to predict response (Y)
  Function of expected Y = $\beta_0 + \beta_1 * X_1$

- *Multiple:* Two or more predictor variables (X’s) used to predict response (Y)
  Function of expected Y = $\beta_0 + \beta_1 * X_1 + \ldots + \beta_p * X_p$

Predictors can be any quantifiable variables

2. Simple Linear Regression (SLR)

- Y = continuous response
- Distribution of Y: Gaussian (normal)

- Model: E[Y] = $\beta_0 + \beta_1 X$ where β_0, β_1 are unknown

$$Y = E[Y] + \text{deviation} = \beta_0 + \beta_1 X + \varepsilon$$

- Data: $(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)$
- β_0 = Expected or average Y value when X = 0
- β_1 = Expected or average difference in Y per unit increase in X

2.1 Multiple Linear Regression (MLR)

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots . . + \beta_p X_p$$

$$Y = E[Y] + \varepsilon$$

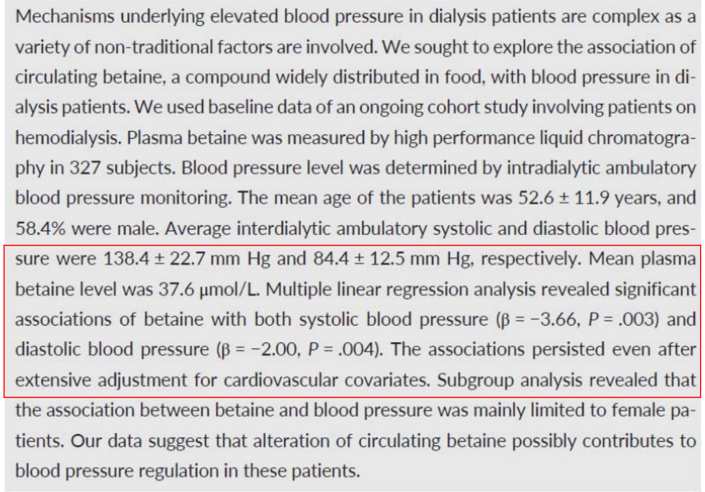
β_1 - Expected (average) change in Y when X_1 increases by one unit and X_2 , X_3 ….X_p are unchanged

β_2 – Expected (average) change in Y when X_2 increases by one unit and X_1 ,X_3 ….X_p are unchanged

β_0 = Expected or average Y value when *all X's* = 0

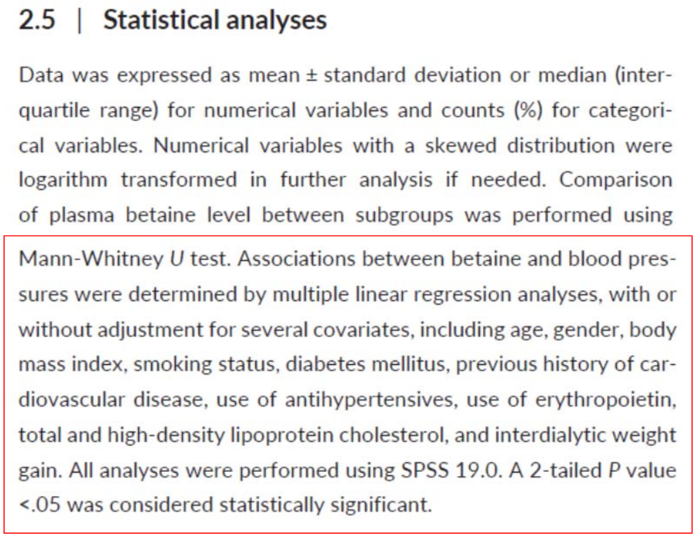
2.2a Example: MLR

Wang L, Zhao M, Liu W et al (2018). Association of betaine with blood pressure in dialysis patients. *J Clin Hypertension* 1-6.



Mechanisms underlying elevated blood pressure in dialysis patients are complex as a variety of non-traditional factors are involved. We sought to explore the association of circulating betaine, a compound widely distributed in food, with blood pressure in dialysis patients. We used baseline data of an ongoing cohort study involving patients on hemodialysis. Plasma betaine was measured by high performance liquid chromatography in 327 subjects. Blood pressure level was determined by intradialytic ambulatory blood pressure monitoring. The mean age of the patients was 52.6 ± 11.9 years, and 58.4% were male. Average interdialytic ambulatory systolic and diastolic blood pressure were 138.4 ± 22.7 mm Hg and 84.4 ± 12.5 mm Hg, respectively. Mean plasma betaine level was 37.6 μmol/L. Multiple linear regression analysis revealed significant associations of betaine with both systolic blood pressure (β = −3.66, *P* = .003) and diastolic blood pressure (β = −2.00, *P* = .004). The associations persisted even after extensive adjustment for cardiovascular covariates. Subgroup analysis revealed that the association between betaine and blood pressure was mainly limited to female patients. Our data suggest that alteration of circulating betaine possibly contributes to blood pressure regulation in these patients.

2.2b Example: MLR



2.5 | Statistical analyses

Data was expressed as mean ± standard deviation or median (interquartile range) for numerical variables and counts (%) for categorical variables. Numerical variables with a skewed distribution were logarithm transformed in further analysis if needed. Comparison of plasma betaine level between subgroups was performed using Mann-Whitney *U* test. Associations between betaine and blood pressures were determined by multiple linear regression analyses, with or without adjustment for several covariates, including age, gender, body mass index, smoking status, diabetes mellitus, previous history of cardiovascular disease, use of antihypertensives, use of erythropoietin, total and high-density lipoprotein cholesterol, and interdialytic weight gain. All analyses were performed using SPSS 19.0. A 2-tailed *P* value <.05 was considered statistically significant.

2.2c Example: MLR

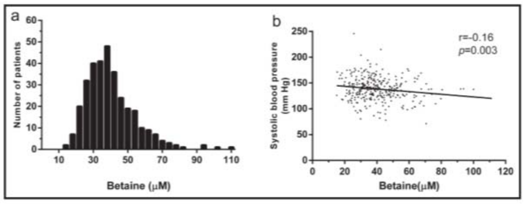


FIGURE Distribution of circulating betaine level (A) and its crude correlation to systolic blood pressure (B)

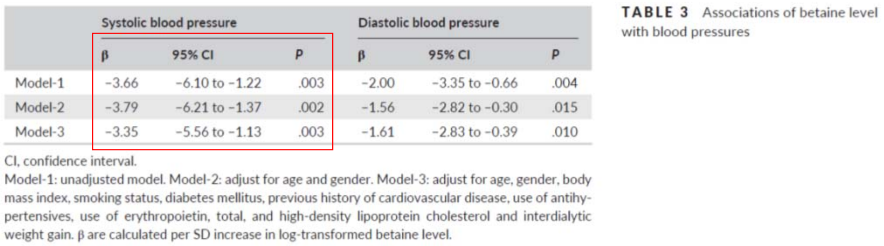


TABLE 3 Associations of betaine level with blood pressures

	Systolic blood pressure			Diastolic blood pressure		
	β	95% CI	P	β	95% CI	P
Model-1	−3.66	−6.10 to −1.22	.003	−2.00	−3.35 to −0.66	.004
Model-2	−3.79	−6.21 to −1.37	.002	−1.56	−2.82 to −0.30	.015
Model-3	−3.35	−5.56 to −1.13	.003	−1.61	−2.83 to −0.39	.010

CI, confidence interval.

Model-1: unadjusted model. Model-2: adjust for age and gender. Model-3: adjust for age, gender, body mass index, smoking status, diabetes mellitus, previous history of cardiovascular disease, use of antihypertensives, use of erythropoietin, total, and high-density lipoprotein cholesterol and interdialytic weight gain. β are calculated per SD increase in log-transformed betaine level.

3. Simple Logistic Regression (SLogR)

- Y = binary or dichotomous response variable
- Distribution of Y: Binomial
- Pr(Y=1) = E[Y]
- Model: $\log_e$ (odds) = $\log_e$ {Pr(Y=1)/Pr(Y=0)} = $\beta_0 + \beta_1 X$
  which implies the model for the probability where β_0, β_1 are unknown
- Data: $(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)$
- β_0 = log odds when X = 0
- β_1 = change in log odds per unit increase in X = log(odds ratio)

3.1 Multiple Logistic Regression (MLogR)

log(odds) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots.. + \beta_p X_p$

β_1 - Change in log(odds) when X_1 increases by one unit and X_2 , X_3 ….X_p are unchanged

β_2 – Change in log(odds) when X_2 increases by one unit and X_1 ,X_3 ….X_p are unchanged

β_0 = log(odds) when *all X's* = 0

3.2 Log Odds and Probability

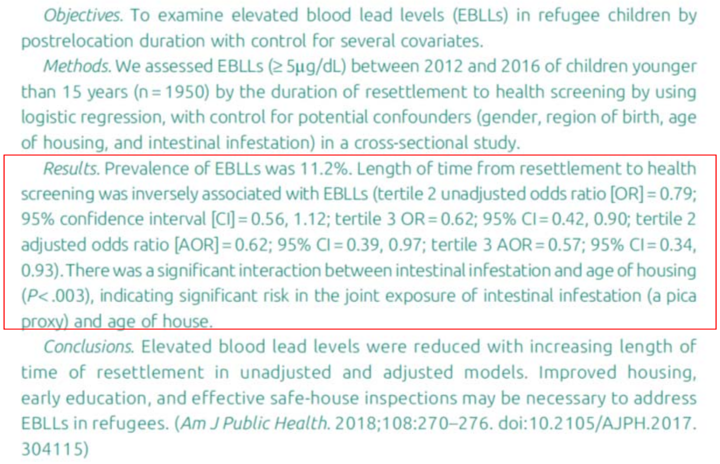
$$\log\left\{\frac{\Pr(Y=1)}{\Pr(Y=0)}\right\} = \log\left\{\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

is equivalent to

$$\Pr(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1+\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

3.3a Example: MLogR

Kotey S, Carrico R, Wiemken TL, et al. (2018). Elevated blood lead levels by length of time from resettlement to health screening in Kentucky refugee children. *Am J Public Health* 108:270-276.



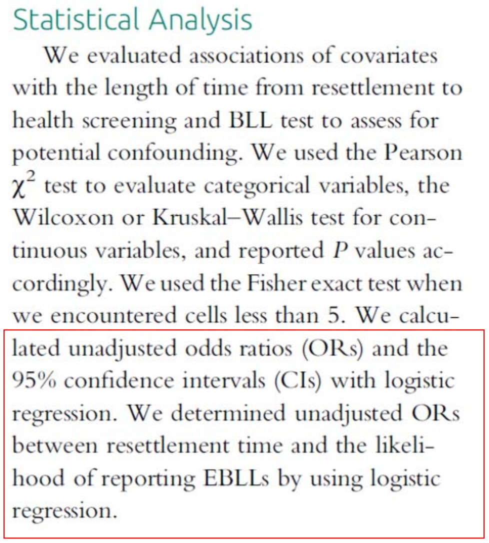
Objectives. To examine elevated blood lead levels (EBLLs) in refugee children by postrelocation duration with control for several covariates.

Methods. We assessed EBLLs (≥ 5µg/dL) between 2012 and 2016 of children younger than 15 years (n = 1950) by the duration of resettlement to health screening by using logistic regression, with control for potential confounders (gender, region of birth, age of housing, and intestinal infestation) in a cross-sectional study.

Results. Prevalence of EBLLs was 11.2%. Length of time from resettlement to health screening was inversely associated with EBLLs (tertile 2 unadjusted odds ratio [OR] = 0.79; 95% confidence interval [CI] = 0.56, 1.12; tertile 3 OR = 0.62; 95% CI = 0.42, 0.90; tertile 2 adjusted odds ratio [AOR] = 0.62; 95% CI = 0.39, 0.97; tertile 3 AOR = 0.57; 95% CI = 0.34, 0.93). There was a significant interaction between intestinal infestation and age of housing (*P* < .003), indicating significant risk in the joint exposure of intestinal infestation (a pica proxy) and age of house.

Conclusions. Elevated blood lead levels were reduced with increasing length of time of resettlement in unadjusted and adjusted models. Improved housing, early education, and effective safe-house inspections may be necessary to address EBLLs in refugees. (*Am J Public Health.* 2018;108:270–276. doi:10.2105/AJPH.2017.304115)

3.3b Example: MLogR



Statistical Analysis

We evaluated associations of covariates with the length of time from resettlement to health screening and BLL test to assess for potential confounding. We used the Pearson χ^2 test to evaluate categorical variables, the Wilcoxon or Kruskal–Wallis test for continuous variables, and reported *P* values accordingly. We used the Fisher exact test when we encountered cells less than 5. We calculated unadjusted odds ratios (ORs) and the 95% confidence intervals (CIs) with logistic regression. We determined unadjusted ORs between resettlement time and the likelihood of reporting EBLLs by using logistic regression.

3.3c Example: MLogR

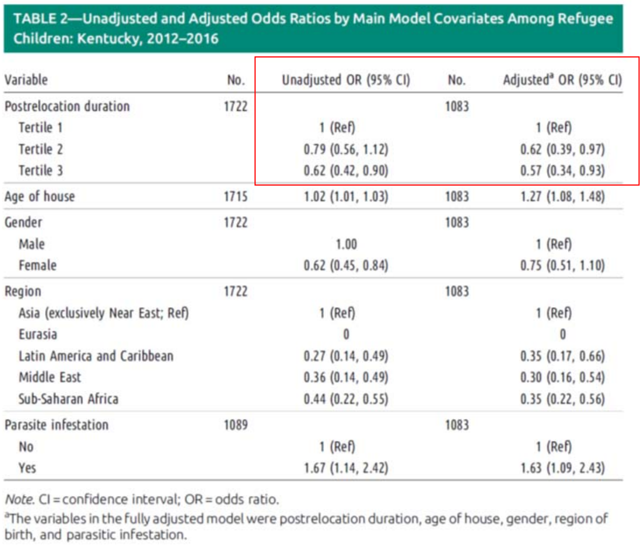


TABLE 2—Unadjusted and Adjusted Odds Ratios by Main Model Covariates Among Refugee Children: Kentucky, 2012–2016

Variable	No.	Unadjusted OR (95% CI)	No.	Adjusted[a] OR (95% CI)
Postrelocation duration	1722		1083	
Tertile 1		1 (Ref)		1 (Ref)
Tertile 2		0.79 (0.56, 1.12)		0.62 (0.39, 0.97)
Tertile 3		0.62 (0.42, 0.90)		0.57 (0.34, 0.93)
Age of house	1715	1.02 (1.01, 1.03)	1083	1.27 (1.08, 1.48)
Gender	1722		1083	
Male		1.00		1 (Ref)
Female		0.62 (0.45, 0.84)		0.75 (0.51, 1.10)
Region	1722		1083	
Asia (exclusively Near East; Ref)		1 (Ref)		1 (Ref)
Eurasia		0		0
Latin America and Caribbean		0.27 (0.14, 0.49)		0.35 (0.17, 0.66)
Middle East		0.36 (0.14, 0.49)		0.30 (0.16, 0.54)
Sub-Saharan Africa		0.44 (0.22, 0.55)		0.35 (0.22, 0.56)
Parasite infestation	1089		1083	
No		1 (Ref)		1 (Ref)
Yes		1.67 (1.14, 2.42)		1.63 (1.09, 2.43)

Note. CI = confidence interval; OR = odds ratio.

[a] The variables in the fully adjusted model were postrelocation duration, age of house, gender, region of birth, and parasitic infestation.

4. Simple Log-Linear Regression

- Models for “count” data -- e.g., number of events per unit time, number of events per number at risk
- Distribution: Poisson
- Model: log(risk of event) =$\beta_0 + \beta_1 X$
- β_0 = log(risk) when $X = 0$
- β_1 = change in log(risk) per unit increase in X
  = log(risk ratio) = log(“rate ratio”)

4.1 Multiple Log-Linear Regression

$$\log(\text{risk}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots.. + \beta_p X_p$$

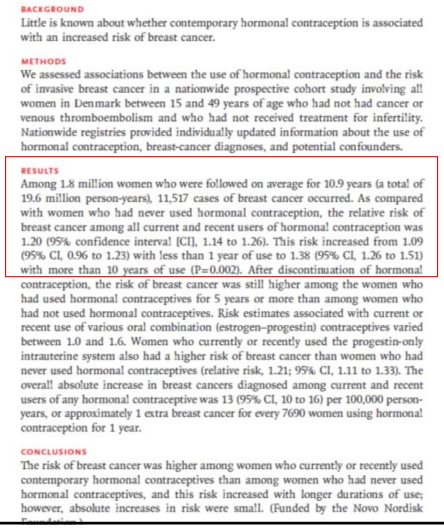
β_1 - Change in log(risk) when X_1 increases by one unit and X_2 , X_3 ….X_p are unchanged

β_2 – Change in log(risk) when X_2 increases by one unit and X_1 ,X_3 ….X_p are unchanged

β_0 = log(risk) when *all X's* = 0

4.2a Example: Multiple Poisson Regression

Morch LS, Skovlund CW, Hannaford PC, et al. (2012) Contemporary hormonal contraception and the risk of breast cancer.NEJM 337: 2228-39



BACKGROUND

Little is known about whether contemporary hormonal contraception is associated with an increased risk of breast cancer.

METHODS

We assessed associations between the use of hormonal contraception and the risk of invasive breast cancer in a nationwide prospective cohort study involving all women in Denmark between 15 and 49 years of age who had not had cancer or venous thromboembolism and who had not received treatment for infertility. Nationwide registries provided individually updated information about the use of hormonal contraception, breast-cancer diagnoses, and potential confounders.

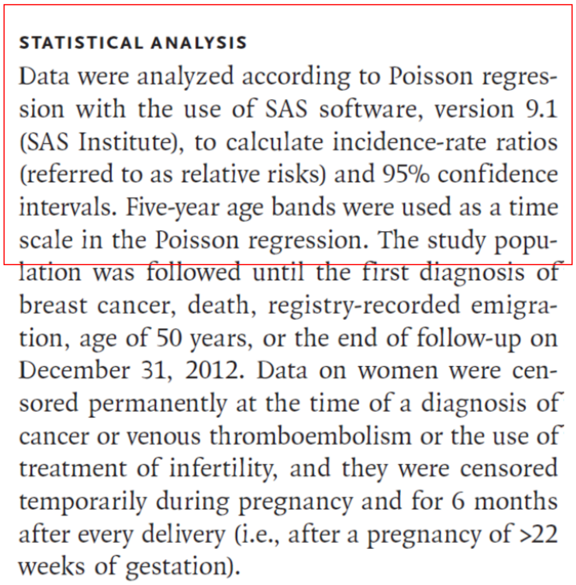
RESULTS

Among 1.8 million women who were followed on average for 10.9 years (a total of 19.6 million person-years), 11,517 cases of breast cancer occurred. As compared with women who had never used hormonal contraception, the relative risk of breast cancer among all current and recent users of hormonal contraception was 1.20 (95% confidence interval [CI], 1.14 to 1.26). This risk increased from 1.09 (95% CI, 0.96 to 1.23) with less than 1 year of use to 1.38 (95% CI, 1.26 to 1.51) with more than 10 years of use (P=0.002). After discontinuation of hormonal contraception, the risk of breast cancer was still higher among the women who had used hormonal contraceptives for 5 years or more than among women who had not used hormonal contraceptives. Risk estimates associated with current or recent use of various oral combination (estrogen–progestin) contraceptives varied between 1.0 and 1.6. Women who currently or recently used the progestin-only intrauterine system also had a higher risk of breast cancer than women who had never used hormonal contraceptives (relative risk, 1.21; 95% CI, 1.11 to 1.33). The overall absolute increase in breast cancers diagnosed among current and recent users of any hormonal contraceptive was 13 (95% CI, 10 to 16) per 100,000 person-years, or approximately 1 extra breast cancer for every 7690 women using hormonal contraception for 1 year.

CONCLUSIONS

The risk of breast cancer was higher among women who currently or recently used contemporary hormonal contraceptives than among women who had never used hormonal contraceptives, and this risk increased with longer durations of use; however, absolute increases in risk were small. (Funded by the Novo Nordisk Foundation.)

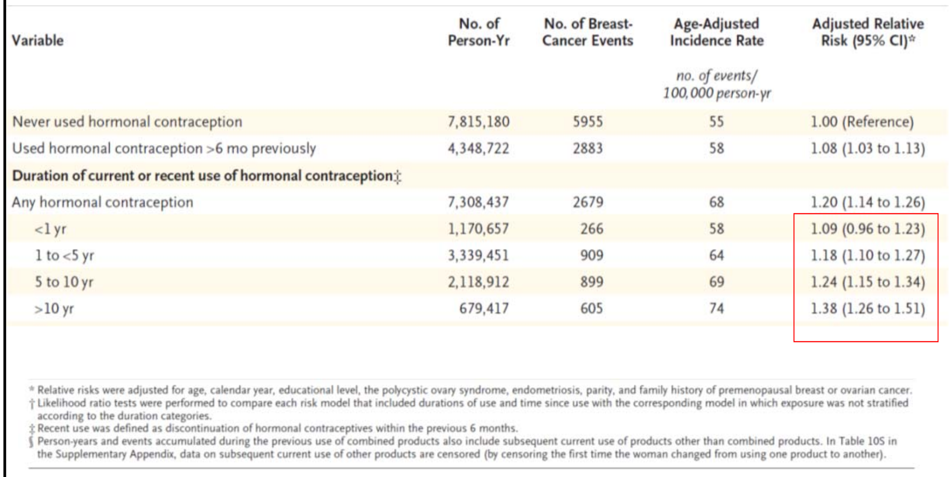
4.2b Example: Multiple Poisson Regression



STATISTICAL ANALYSIS

Data were analyzed according to Poisson regression with the use of SAS software, version 9.1 (SAS Institute), to calculate incidence-rate ratios (referred to as relative risks) and 95% confidence intervals. Five-year age bands were used as a time scale in the Poisson regression. The study population was followed until the first diagnosis of breast cancer, death, registry-recorded emigration, age of 50 years, or the end of follow-up on December 31, 2012. Data on women were censored permanently at the time of a diagnosis of cancer or venous thromboembolism or the use of treatment of infertility, and they were censored temporarily during pregnancy and for 6 months after every delivery (i.e., after a pregnancy of >22 weeks of gestation).

4.2c Example: Multiple Poisson Regression



Variable	No. of Person-Yr	No. of Breast-Cancer Events	Age-Adjusted Incidence Rate	Adjusted Relative Risk (95% CI)*
			no. of events/ 100,000 person-yr	
Never used hormonal contraception	7,815,180	5955	55	1.00 (Reference)
Used hormonal contraception >6 mo previously	4,348,722	2883	58	1.08 (1.03 to 1.13)
Duration of current or recent use of hormonal contraception‡				
Any hormonal contraception	7,308,437	2679	68	1.20 (1.14 to 1.26)
<1 yr	1,170,657	266	58	1.09 (0.96 to 1.23)
1 to <5 yr	3,339,451	909	64	1.18 (1.10 to 1.27)
5 to 10 yr	2,118,912	899	69	1.24 (1.15 to 1.34)
>10 yr	679,417	605	74	1.38 (1.26 to 1.51)

* Relative risks were adjusted for age, calendar year, educational level, the polycystic ovary syndrome, endometriosis, parity, and family history of premenopausal breast or ovarian cancer.

† Likelihood ratio tests were performed to compare each risk model that included durations of use and time since use with the corresponding model in which exposure was not stratified according to the duration categories.

‡ Recent use was defined as discontinuation of hormonal contraceptives within the previous 6 months.

§ Person-years and events accumulated during the previous use of combined products also include subsequent current use of products other than combined products. In Table 10S in the Supplementary Appendix, data on subsequent current use of other products are censored (by censoring the first time the woman changed from using one product to another).

5. Survival Analysis and Cox Regression

- Most important special case of log-linear regression
- Model:

  log(hazard or “risk”) of event at time t =

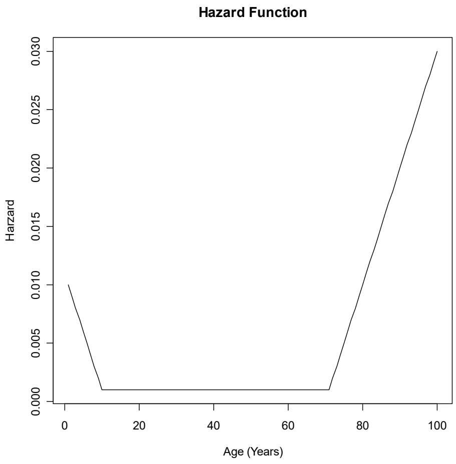
  $$f(t)\ + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$$

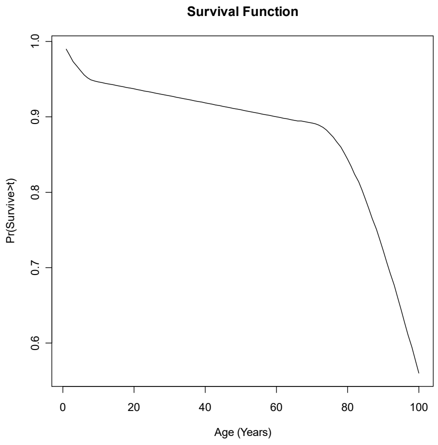
- β_1 is the log(hazard) of the event associated with X_1 at any given time *t*, and X_2 ,X_3 ….X_p are unchanged
- Same as hazard of event = $e^{f(t)} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots e^{\beta_p X_p}$
- “Multiplicative” or “proportional hazards” model (Cox, 1972)

5.1 Hazard and Survival Functions

- Survival function S(t) = Prob(event occurs after t)
  - Prob(“survive beyond t”)
  - S(t) in (0,1)
  - S(t) non-increasing function since S(t) is greater than or equal to S(t+r), r>0
- Hazard function h(t) = risk of event in small interval (t,t+dt) given person survived to t per unit time
- h(t) is the slope of the S(t) as a fraction of S(t)

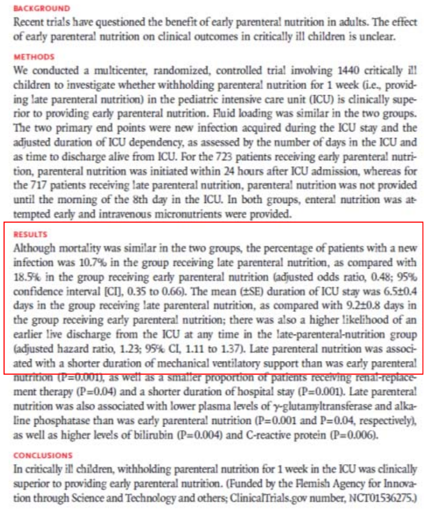
5.2 Hypothetical Hazard/Survival Functions





5.3a Example: Survival/Cox Reg

Fivez T, Kerklaan D, Mesotten D, et al. (2016). Early versus late parenteral nutrition in critically ill c*hildren. NEJM* 374: 1111-22.



BACKGROUND

Recent trials have questioned the benefit of early parenteral nutrition in adults. The effect of early parenteral nutrition on clinical outcomes in critically ill children is unclear.

METHODS

We conducted a multicenter, randomized, controlled trial involving 1440 critically ill children to investigate whether withholding parenteral nutrition for 1 week (i.e., providing late parenteral nutrition) in the pediatric intensive care unit (ICU) is clinically superior to providing early parenteral nutrition. Fluid loading was similar in the two groups. The two primary end points were new infection acquired during the ICU stay and the adjusted duration of ICU dependency, as assessed by the number of days in the ICU and as time to discharge alive from ICU. For the 723 patients receiving early parenteral nutrition, parenteral nutrition was initiated within 24 hours after ICU admission, whereas for the 717 patients receiving late parenteral nutrition, parenteral nutrition was not provided until the morning of the 8th day in the ICU. In both groups, enteral nutrition was attempted early and intravenous micronutrients were provided.

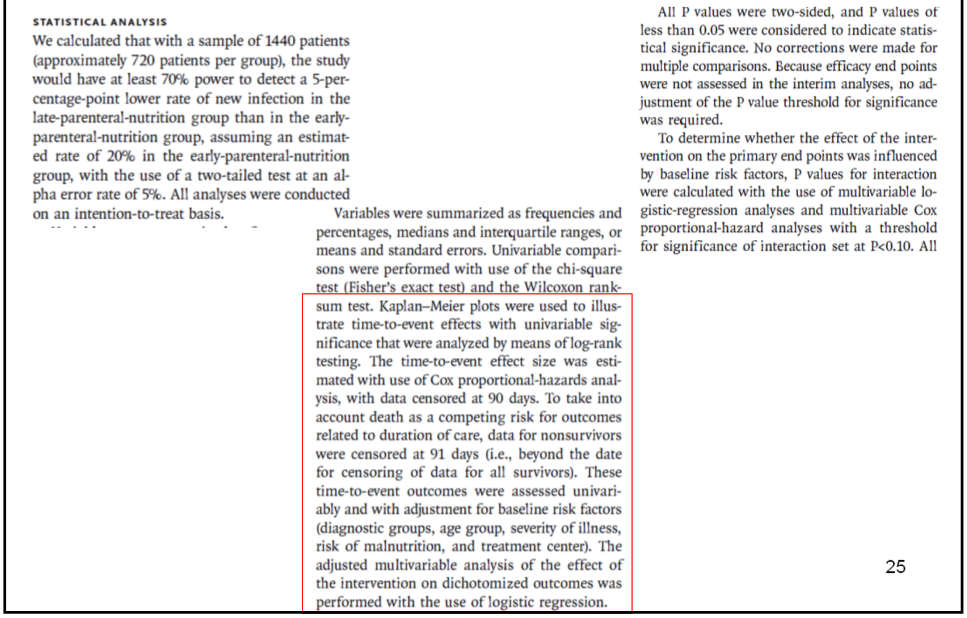
RESULTS

Although mortality was similar in the two groups, the percentage of patients with a new infection was 10.7% in the group receiving late parenteral nutrition, as compared with 18.5% in the group receiving early parenteral nutrition (adjusted odds ratio, 0.48; 95% confidence interval [CI], 0.35 to 0.66). The mean (±SE) duration of ICU stay was 6.5±0.4 days in the group receiving late parenteral nutrition, as compared with 9.2±0.8 days in the group receiving early parenteral nutrition; there was also a higher likelihood of an earlier live discharge from the ICU at any time in the late-parenteral-nutrition group (adjusted hazard ratio, 1.23; 95% CI, 1.11 to 1.37). Late parenteral nutrition was associated with a shorter duration of mechanical ventilatory support than was early parenteral nutrition (P=0.001), as well as a smaller proportion of patients receiving renal-replacement therapy (P=0.04) and a shorter duration of hospital stay (P=0.001). Late parenteral nutrition was also associated with lower plasma levels of γ-glutamyltransferase and alkaline phosphatase than was early parenteral nutrition (P=0.001 and P=0.04, respectively), as well as higher levels of bilirubin (P=0.004) and C-reactive protein (P=0.006).

CONCLUSIONS

In critically ill children, withholding parenteral nutrition for 1 week in the ICU was clinically superior to providing early parenteral nutrition. (Funded by the Flemish Agency for Innovation through Science and Technology and others; ClinicalTrials.gov number, NCT01536275.)

5.3b Example: Survival/ Cox Reg



STATISTICAL ANALYSIS

We calculated that with a sample of 1440 patients (approximately 720 patients per group), the study would have at least 70% power to detect a 5-percentage-point lower rate of new infection in the late-parenteral-nutrition group than in the early-parenteral-nutrition group, assuming an estimated rate of 20% in the early-parenteral-nutrition group, with the use of a two-tailed test at an alpha error rate of 5%. All analyses were conducted on an intention-to-treat basis.

Variables were summarized as frequencies and percentages, medians and interquartile ranges, or means and standard errors. Univariable comparisons were performed with use of the chi-square test (Fisher's exact test) and the Wilcoxon rank-sum test. Kaplan–Meier plots were used to illustrate time-to-event effects with univariable significance that were analyzed by means of log-rank testing. The time-to-event effect size was estimated with use of Cox proportional-hazards analysis, with data censored at 90 days. To take into account death as a competing risk for outcomes related to duration of care, data for nonsurvivors were censored at 91 days (i.e., beyond the date for censoring of data for all survivors). These time-to-event outcomes were assessed univariably and with adjustment for baseline risk factors (diagnostic groups, age group, severity of illness, risk of malnutrition, and treatment center). The adjusted multivariable analysis of the effect of the intervention on dichotomized outcomes was performed with the use of logistic regression.

All P values were two-sided, and P values of less than 0.05 were considered to indicate statistical significance. No corrections were made for multiple comparisons. Because efficacy end points were not assessed in the interim analyses, no adjustment of the P value threshold for significance was required.

To determine whether the effect of the intervention on the primary end points was influenced by baseline risk factors, P values for interaction were calculated with the use of multivariable logistic-regression analyses and multivariable Cox proportional-hazard analyses with a threshold for significance of interaction set at P<0.10. All

5.3c Example: Survival/ Cox Reg

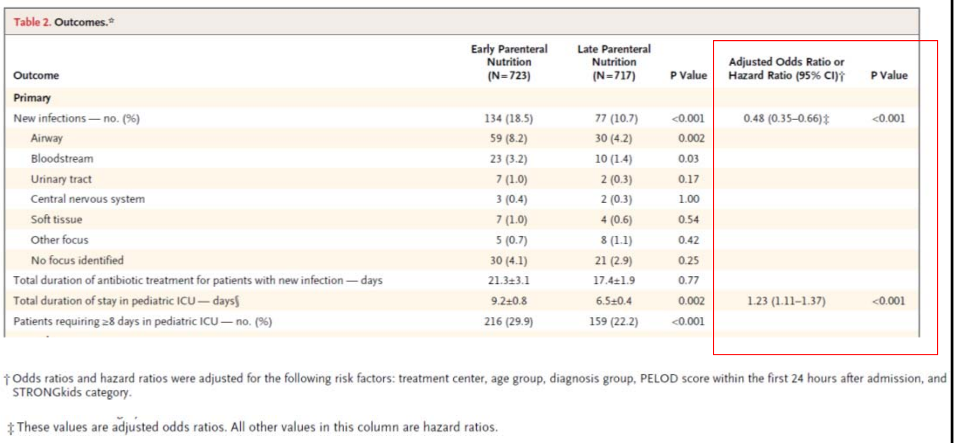


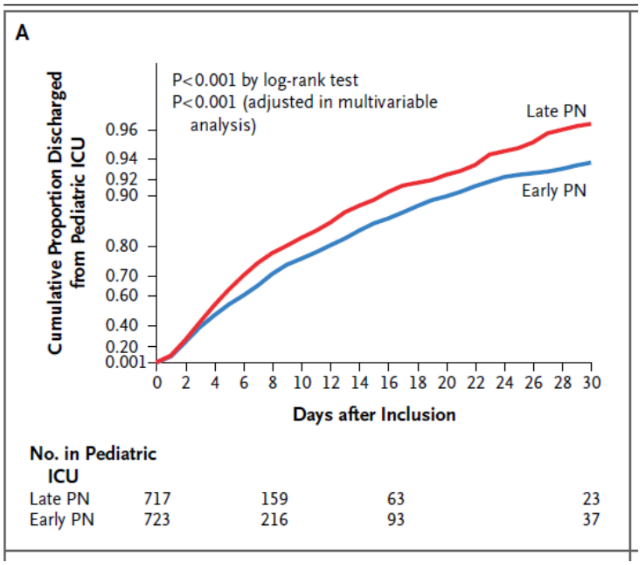
Table 2. Outcomes.*

Outcome	Early Parenteral Nutrition (N=723)	Late Parenteral Nutrition (N=717)	P Value	Adjusted Odds Ratio or Hazard Ratio (95% CI)†	P Value
Primary					
New infections — no. (%)	134 (18.5)	77 (10.7)	<0.001	0.48 (0.35–0.66)‡	<0.001
Airway	59 (8.2)	30 (4.2)	0.002		
Bloodstream	23 (3.2)	10 (1.4)	0.03		
Urinary tract	7 (1.0)	2 (0.3)	0.17		
Central nervous system	3 (0.4)	2 (0.3)	1.00		
Soft tissue	7 (1.0)	4 (0.6)	0.54		
Other focus	5 (0.7)	8 (1.1)	0.42		
No focus identified	30 (4.1)	21 (2.9)	0.25		
Total duration of antibiotic treatment for patients with new infection — days	21.3±3.1	17.4±1.9	0.77		
Total duration of stay in pediatric ICU — days§	9.2±0.8	6.5±0.4	0.002	1.23 (1.11–1.37)	<0.001
Patients requiring ≥8 days in pediatric ICU — no. (%)	216 (29.9)	159 (22.2)	<0.001		

† Odds ratios and hazard ratios were adjusted for the following risk factors: treatment center, age group, diagnosis group, PELOD score within the first 24 hours after admission, and STRONGkids category.

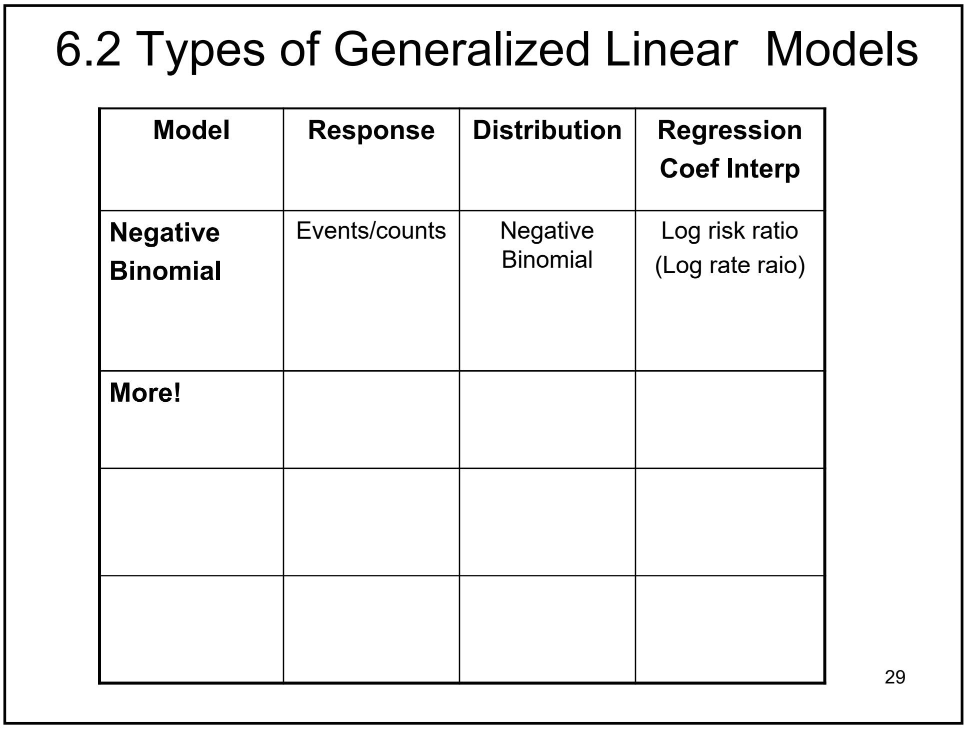
‡ These values are adjusted odds ratios. All other values in this column are hazard ratios.

5.3d Example: Survival/ Cox Reg



6.1 Types of Generalized Linear Models

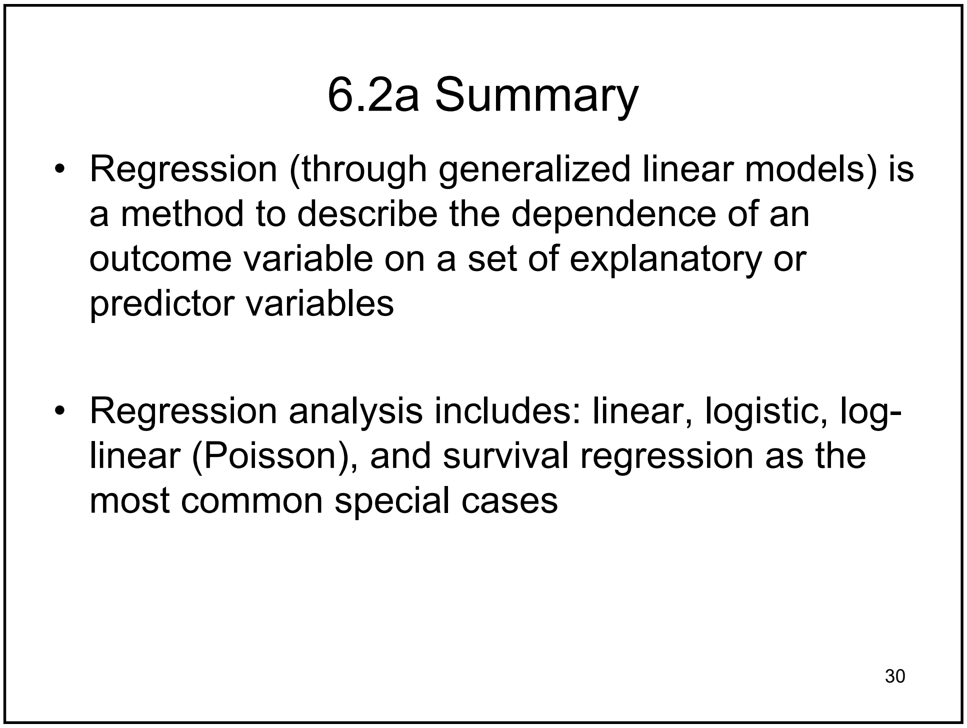
Model	Response	Distribution	Regression Coef Interp
Linear	Continuous	Gaussian	Change in ave(Y) per unit change in X
Logistic	Binary	Binomial	Log odds ratio
Log-linear	Events/counts	Poisson	Log risk ratio (Log rate raio)
Proportional hazards	Times to events	Semi-parametric	Log hazard ratio



6.2 Types of Generalized Linear Models

Model	Response	Distribution	Regression Coef Interp
Negative Binomial	Events/counts	Negative Binomial	Log risk ratio (Log rate raio)
More!			

29



6.2a Summary

- Regression (through generalized linear models) is a method to describe the dependence of an outcome variable on a set of explanatory or predictor variables
- Regression analysis includes: linear, logistic, log-linear (Poisson), and survival regression as the most common special cases

30

6.2b Summary

- The regression coefficients represent characteristics of scientific interest: change in average Y per unit increase in a particular X odds ratio, risk ratio (rate ratio), hazard ratio
- Estimating coefficients from data is the process of using evidence to address a scientific question

6.2c Summary

- Linear regression provides a model for a continuous outcome
- Logistic regression provides a model for a dichotomous outcome
- Log-linear (Poisson) regression provides a model for “count” data; negative binomial regression may also be used
- Survival regression provides a model for the hazard of an event at time t