

Medizinische Informatik Klinische Versuchsplanung



1

Definition: Klinische Studie

Eine klinische Studie untersucht die Wirkung (und ihr(e) Maß/Größe) einer (oder mehrerer) Intervention(en) an Menschen.

Intervention (z.B.):

- Behandlung (Vergleich zw. Behandlungsmethoden)
- Anwesenheit eines Faktors, Faktoren bei Krankheiten (z.B.: Übergewicht als Risiko für Herz-Kreislauf-Erkrankungen)



2

(Einige) Gebiete der klinischen (statistischen) Studien

- **Fallzahl-schätzung, Konfidenzintervalle**
- Survival-Analyse
- Äquivalenznachweise
- Randomisation und Verblindung
- Phase I-IV - Studien



Entscheidung	Wirklichkeit	
	$H_0: \pi_1 = \pi_2$	$H_0: \pi_1 \neq \pi_2$
Nicht-Ablehnung von H_0 (Annahme von H_0)	Richtige Nicht-Ablehnung	Falsche Nicht-Ablehnung (β -, Typ II-Fehler)
Ablehnung von H_0	Falsche Ablehnung (α -, Typ I-Fehler)	Richtige Ablehnung

3

Asymptotischer 4-Felder Test zum Vergleich von 2 **Raten**

		Erfolg			Rate — equiv. relative Häufigkeit
		+	-		
Gruppe	1	a	b	N_1	$N = N_1 + N_2;$
	2	c	d	N_2	$N_1 = a + b;$
				N	$N_2 = c + d;$

$$r_1 = a / (a + b) = a / N_1; \quad r_2 = c / (c + d) = c / N_2$$

für genügend große N-werte:
 $r_1 \sim \pi_1; r_2 \sim \pi_2$

Testhypothesen:

$$H_0: \pi_1 = \pi_2 \quad \text{vs.} \quad H_1: \pi_1 \neq \pi_2$$

Transformation der Raten zur Prüfgröße:

$$Z = \frac{(r_1 - r_2)}{\sqrt{\frac{r_1(1-r_1)}{N_1} + \frac{r_2(1-r_2)}{N_2}}}$$

4

$$Z = \frac{(r_1 - r_2)}{\sqrt{\frac{r_1(1-r_1)}{N_1} + \frac{r_2(1-r_2)}{N_2}}}$$

Entscheidungsregel über Normal-Approximation:

$H_0: Z \rightarrow \mathcal{N}(0,1)$ Standardnormalverteilung

$$|Z| \geq Z_{(1-\frac{\alpha}{2})}: H_0: \text{ablehnen}$$

einseitig oder zweiseitig?

Warum $1-\alpha/2$?

Z-Werte: -Z und +Z symmetrisch

Beispiele:

80 — 80%

	Erfolg	kein Erfolg	
Intervention	80	20	100
Kontrolle	70	30	100

$$Z_{\text{berechnet}} = 1.644$$

800 — 800%

	Erfolg	kein Erfolg	
Intervention	800	200	1000
Kontrolle	700	300	1000

$$Z_{\text{berechnet}} = 5.199$$

$$\alpha = 0.05 \quad 1-\alpha/2 = 0.975 \quad |Z| \geq Z_{(1-\frac{\alpha}{2})}: Z_{0.975} = 1.960$$

5

$$\text{wenn } |Z| \geq Z_{(1-\frac{\alpha}{2})}: H_0: \text{ablehnen}$$

mit Excel:

Erfolg	kein Erfolg		
80	20	100	NI
70	30	100	NK
150	50	200	
$r_{\text{Intervention}}=r_1=0.8$			
$r_{\text{Kontrolle}}=r_2=0.7$			
$\alpha=0.05$			
$1-\alpha/2=0.975$			
$ Z _{\text{berechnet}}=1.6440$			
$Z_{(1-\alpha/2)=0.975}=1.9600$			

$$Z = \frac{(r_1 - r_2)}{\sqrt{\frac{r_1(1-r_1)}{N_1} + \frac{r_2(1-r_2)}{N_2}}}$$

$$r_1=80/100; r_2=70/100; N_1=NI; N_2=NK$$

$$Z_{\text{berechnet}} < Z_{(1-\alpha/2)}: H_0 \text{ zu behalten}$$

6

Fallzahlsschätzung für Differenz von 2 (unbekannten) Raten

Raten (d.h. relative Häufigkeiten):

- r_I : Erfolg bei Interventionen/Anzahl-Interventionen
- r_K : Erfolg bei Kontrollen/Anzahl-Kontrollen
- r_u : untere Grenze des Konfidenzintervalls **für Differenz der Raten**
- r_o : obere Grenze des Konfidenzintervalls **für Differenz der Raten**

$$r_u = (r_I - r_K) - z_{(1-\frac{\alpha}{2})} \sqrt{\frac{r_I(1-r_I)}{N_I} + \frac{r_K(1-r_K)}{N_K}}$$

$$r_o = (r_I - r_K) + z_{(1-\frac{\alpha}{2})} \sqrt{\frac{r_I(1-r_I)}{N_I} + \frac{r_K(1-r_K)}{N_K}}$$

$$r_u = (r_I - r_K) - d \quad r_o = (r_I - r_K) + d$$

$$KI_{(1-\alpha)} = (r_I - r_K) \pm d$$

7

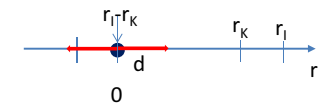
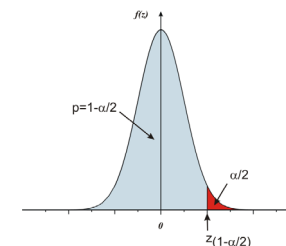
Zweiseitiges Konfidenzintervall für Differenz von 2 (unbekannten) Raten

Signifikanzniveau: α (an beiden Enden der Verteilungsfunktion $\alpha/2$)

Konfidenzniveau: $1-\alpha$

$$KI_{1-(1-\alpha)} = (r_I - r_K) \pm d \quad d = z_{(1-\frac{\alpha}{2})} \sqrt{\frac{r_I(1-r_I)}{N_I} + \frac{r_K(1-r_K)}{N_K}}$$

$z_{(1-\frac{\alpha}{2})}$: z-Wert der Standardnormalverteilung bei $p=1-\alpha/2$



zum Auftreten von „d“ gehört auch eine relative Häufigkeit!

8

Beispiel:

Erfolgsrate bei neuer Intervention: $r_1=0.8$;

Erfolgsrate mit alter Methode: $r_2=0.6$;

In beiden Untersuchungen sei der Stichprobenumfang 100;

Wie groß ist das Konfidenzintervall mit einer $\gamma=0.95$ (d.h. $Kl_{0.95}$) für die Differenz der Raten?

Erfolg Stp1	80		d=	0.1240
Erfolg Stp2	60		$r_{\text{untere}}=$	0.076
N1	100		$r_{\text{obere}}=$	0.324
N2	100			
r1	0.8			
r2	0.6			
r1-r2=	0.2			
alpha=	0.05			
1-alpha/2=	0.975			
z1-alpha/2=	1.960			
$r=(r1+r2)/2=$	0.7			

Problem:

Wir möchten die **Länge** ($2d$) des Konfidenzintervalls verkleinern (d.h. bei dem selben Konfidenzniveau $\gamma=0.95$). Wie viele Fälle soll man beobachten (wie groß muss der Umfang sein)?

9



Fallzahlsschätzung für Differenz von 2 (unbekannten) Raten — bestimmten Fehler erster Art ist zu erreichen

$$d = z_{(1-\frac{\alpha}{2})} \sqrt{\left[\frac{r_I(1-r_I)}{N_I} + \frac{r_K(1-r_K)}{N_K} \right]} \quad \text{mit } N_I=N_K=N$$

$$N = \frac{z_{(1-\alpha/2)}^2}{d^2} \sqrt{[r_I(1-r_I) + r_K(1-r_K)]}$$

Beispiel:

Erfolgsrate bei neuer Intervention: $r_1=0.8$;

Erfolgsrate mit alter Methode: $r_2=0.7$;

$d=0.05$

$N=?$



10

Fallzahlsschätzung für Differenz von 2 (unbekannten) Raten — bestimmter **ersten und zweiten** Fehlerarten sind zu erreichen



4-Felder-Test Fallzahlsschätzung für Differenz von 2 (unbekannten) Raten

	Erfolg	kein Erfolg	
Intervention	80	20	100
Kontrolle	70	30	100

$$r_1=0.8; r_K=0.7; \alpha=0.05; \gamma=0.95; (1-\alpha/2)=0.975; \quad (1-\alpha/2) = \frac{1+\gamma}{2}$$

11

Formulierung der Fragestellung:

- die alte Methode (z.B. Behandlung) ergibt 70% Erfolg
- mit neuer Methode ist 80% Erfolg erwartet
- **wie groß muss der Stichprobenumfang sein, diesen Unterschied statistisch abzusichern?**
- **wir möchten einen α -Fehler [α] und einen β -Fehler [β] erreichen.**

H_0 :

- **Die zwei Raten übereinstimmen, d.h. die Differenz der Raten/Häufigkeiten tritt wegen Zufall auf** (andere Formulierung: die Differenz der Raten/Häufigkeiten weicht nicht signifikant von Null ab.)
 - Diese Differenz kann sowohl kleiner, als auch größer sein.
- Zweiseitige Fragestellung.

12

Formel für asymptotischen Fall:

Bezeichnungen:

N_1, N_2 (z.B.: N_I, N_K) — Vereinfachung: $N_1=N_2=N$;

π_1, π_2 (z.B.: r_I, r_K);

$\pi=(\pi_1+\pi_2)/2$;

α und β als entsprechende Fehler;

zweiseitige Fragestellung:

$$N \approx \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta} \right)^2 \frac{[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}$$

einseitige Fragestellung:

$$N \approx \left(z_{1-\alpha} + z_{1-\beta} \right)^2 \frac{[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}$$

Güte $1-\beta$, aufgrund N

$$z_{1-\beta} = \frac{|\pi_2 - \pi_1| N^{\frac{1}{2}}}{(2\pi(1-\pi))^{1/2}} - z_{1-\frac{\alpha}{2}}$$



13

Beispiel 1.:

$\pi_1=0,8$;

$\pi_2=0,7$;

$\alpha = 0,05$ (2-seitig)

$\beta = 0,05$ ($1-\beta=0,95$)

$p_1 =$	0.8	$z_{1-\alpha/2}$	1.960
$p_2 =$	0.7	$z_{1-\beta}$	1.645
$\alpha =$	0.05	Term 1.	12.99
$\beta =$	0.05	Term 2.	0.37
$1-\beta =$	0.95	Term 3.	0.01
		$N =$	481.00



$\pi=(\pi_1+\pi_2)/2$

Nerrechnet=	481	Zaehler=	2.19317122
p1=	0.8	Nenner=	0.612372436
p2=	0.7	$Z_{(1-\beta)}$ berechnet, N=	1.621
alpha=	0.05	beta(berechnet, N)=	0.052
beta (gezielt)=	0.05		
$\pi =$	0.75		
$z_{1-\alpha/2}$	1.960		
$z_{1-\beta}$ (gezielt)	1.645		

Literatur:

<http://imsiweb.uni-koeln.de/lehre/q1/Q1-06-Konfidenzintervalle.pdf>

<http://imsiweb.uni-koeln.de/lehre/klinstud/KlinStud07-Fallzahlschaetzungen.pdf>

http://de.wikipedia.org/wiki/Konfidenzintervall_einer_unbekannten_Wahrscheinlichkeit

14

Überlebenszeitanalyse: Der Log-Rang-Test

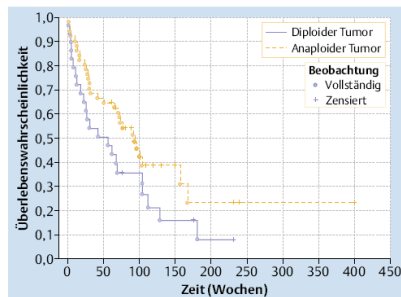


Abb. 1 Kaplan-Meier Kurven für die Überlebenszeit der 80 Zungenkrebspatienten. Es wird in orange/blau die Wahrscheinlichkeit gezeigt, dass ein Patient mit aploidem/diploidem Tumor eine Zeit (in Wochen) überlebt.

- Fallzahlschätzung, Konfidenzintervalle
- **Survival-Analyse**
- Äquivalenznachweise
- Randomisation und Verblindung
- Phase I-IV - Studien

Dtsch. Med. Wochenschr. 2007;
132: e39–e41 · A. Ziegler et al.,
Überlebenszeitanalyse: Der Log-
Rang-Test



15

Überlebens-Daten:

sind Zeiten bis zum Auftreten eines Ereignisses, Lebensdauern/Zeitdauern (Zeitintervalle) von Start- bis Ziel- bzw. Endereignis

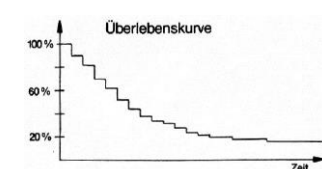
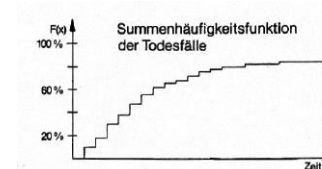
Ist das Endereignis „Tod“ aufgetreten, → „Lebensdauern/Überlebenszeiten“

Definitionen:

$F(t)$ gibt den Anteil der Individuen an, die zum Zeitpunkt $t, t \geq 0$, bereits gestorben sind.

$S(t) = 1 - F(t)$ ist die Überlebenskurve. Sie gibt den Anteil der Individuen zum Zeitpunkt t an, die noch leben.

Aufgrund der Definitionen, die $F(t)$ ist eine Summenhäufigkeitskurve.



16

Kaplan-Meier-Kurve

Begriffe/Definitionen:

Ereigniszeiten t_j : $0 <= t_1 < t_2 < \dots < t_j$ Zeitpunkte der Beobachtungen;

n_j : Anzahl der Individuen unter Risiko zum Zeitpunkt t_j (einschließlich später zensierter);

d_j : Anzahl der Ereignisse (Todesfälle) zum Zeitpunkt t_j ;

$r_j = d_j / n_j$ ist der Anteil der zum Zeitpunkt t_j Gestorbenen unter denjenigen, die t_{j-1} überlebt haben (bzw. die zu t_j unter Risiko stehen);

$s_j = (n_j - d_j) / n_j = (1 - d_j / n_j)$ ist der Anteil der den Zeitpunkt t_j Überlebenden unter denjenigen, die t_{j-1} überlebt haben (bzw. zu t_j unter Risiko stehen);

$S(t_1) = s_1, S(t_2) = S(t_1) \cdot s_2 = s_1 \cdot s_2, S(t_j) = S(t_{j-1}) \cdot s_j = s_1 \cdot s_2 \cdot \dots \cdot s_j = \prod_{i=1}^j s_i$ für $t_j \leq t$

$S(t_j) = \prod_{i=1}^j s_i$ ist einer Punkt der Überlebenskurve — K-M-Kurve $S(t_j) = \prod_{i=1}^j s_i$

Mediane Überlebenszeit: der Zeitpunkt t_M , an dem die Hälfte der Patienten/Individuen noch lebt.

17

Logrank-Test (Mantel-Haenszel)

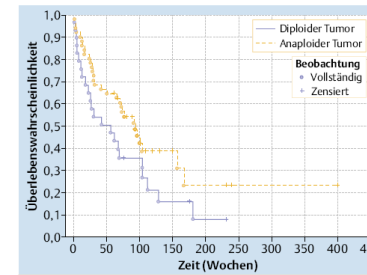


Abb. 1 Kaplan-Meier Kurven für die Überlebenszeit der 80 Zungenkrebspatienten. Es wird in orange/blau die Wahrscheinlichkeit gezeigt, dass ein Patient mit aploidem/diploidem Tumor eine Zeit (in Wochen) überlebt.

Der Log-Rang-Test ist das Standardverfahren in der Überlebenszeitanalyse für einfache Gruppenvergleiche in klinisch-therapeutischen Studien. Mit diesem nichtparametrischen Test lässt sich statistisch überprüfen, ob das Mortalitätsrisiko in zwei oder mehr Gruppen verschieden ist.

Der Logrank-Test ist ein (nichtparametrischer) Signifikanztest, der 2 (unabhängige) Survivalfunktionen auf Gleichheit überprüft, unter Berücksichtigung von zensierten Daten.

Zensierte Daten (auch trunkierte (gestutzte) Daten) sind Daten, bei denen nicht alle Werte einer Statistische Variablen bekannt sind; z. B.: das Ereignis ist noch nicht eingetreten/wenn Kontakt zum Patienten abgebrochen ist.

Literatur:

Dtsch Med Wochenschr 2007; 132: e39–e41 · A. Ziegler et al., Überlebenszeitanalyse:

Der Log-Rang-Test;

<http://imsiweb.uni-koeln.de/lehre/klinstud/KlinStud04-Survival.pdf>

18

Äquivalenznachweise

- Fallzahlschätzung, Konfidenzintervalle
- Survival-Analyse
- **Äquivalenznachweise**
- Randomisation und Verblindung
- Phase I-IV - Studien



Problem/Frage:

Ist ein statistisch signifikanter Unterschied auch klinisch relevant?

$\delta = \Delta = \mu_1 - \mu_2$: ist die kleinste Überlegenheit von μ_1 vs. μ_2 , die klinisch relevant ist.

$A = (-\epsilon - +\delta)$ ist Äquivalenzbereich, z.B.: $A = (-\delta - +\delta)$

μ : Mittelwert, Häufigkeit,...

$KI_{(1-2\alpha)}$ -Konfidenzintervall für $\Delta = \mu_T - \mu_R$ (T: treatment, R-Referenz)

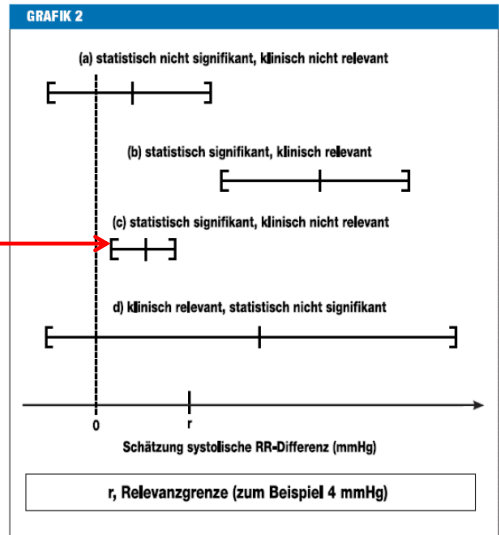
obere- und untere einseitige $(1-\alpha)$ -KI

Behandlungen sind „äquivalent“ mit Sicherheitswahrscheinlichkeit $(1-\alpha)$, wenn A beinhaltet das ganze $(1-2\alpha)$ -KI

19

du Prel J-B, Hommel G, Röhrig B, Blettner M, 2009: Konfidenzintervall oder p-Wert? Teil 4 der Serie zur Bewertung wissenschaftlicher Publikationen. Dtsch Arztebl 106, 335–9

KI enthält nicht "die Null" als Unterschied, erreicht aber die Relevanzgrenze nicht.



Statistische Signifikanz und klinische Relevanz

20

Randomisation (Randomisierung/Zufallszuteilung)

Sie ist eine Technik für jeden Patienten zufällig eine Zuteilung zu einer Behandlungsgruppe zu realisieren, wobei alle Gruppen mit gleicher Wahrscheinlichkeit ausgewählt werden können.

- Fallzahl-schätzung, Konfidenzintervalle
- Survival-Analyse
- Äquivalenznachweise
- **Randomisation und Verblindung**
- Phase I-IV - Studien



Ziel:

- bewussten und unbewussten Störgrößen auszuschalten;
- zu verhindern den Prüfer-Bias (z.B.: abhängig von weiteren Umständen teilt der Prüfer einen Patienten zu einer oder anderer Gruppe zu: Verzerrungen (Bias) können in Behandlungseffekten auftreten);
- zu garantieren, dass statistische Tests gültige/valide Signifikanz-Niveaus ergeben.

22

Randomisierungstechniken

- **einfache Randomisierung:** Zuteilung aufgrund erzeugten Zufallszahlen
 - Münzwurf, Würfeln, Zufallszahlentabellen,....;
 - Nachteil (z.B.):
 - Umfänge der Behandlungsgruppen kann unterschiedlich sein — Verschlechterung der Teststärke (Güte 1-β)
 - Ungleichverteilung der Störfaktoren in den Gruppen
- **Blockrandomisation:** in einem Block bestimmter Länge (z.B. 2,4,6...) sind die Patienten zufällig einer Behandlungsart zugeteilt (AABB, ABAB...);
 - Balanciertheit der Gruppen/Zuteilungen ist möglich
 - Nachteil (z.B.):
 - zufällige Blocklänge
 - bei bekannten Blocklänge kann Code aufgebrochen werden
- **Stratifizierte (geschichtete) Randomisierung:**
Ursache: bekannte/unbekannte Risiko-Faktoren, Unbalanciertheiten können die Auswertungen/Ergebnisse verfälschen.
 - Schichten auf Grund z.B.: Alter, Geschlecht, Stadium, Zentren, Ausgangslage,...
 - Randomisierung innerhalb einer Schicht (Strata)
 - Vorteil (z.B.): Varianz und die Einflüsse der Störfaktoren nehmen ab.

23



Verblindung

In klinischen Studien spricht man von **Verblindung**, wenn

- die Prüfarzte, das Pflegepersonal,
- die teilnehmenden Patienten und auch
- die Personen, die mit dem Monitoring, dem Datenmanagement und der Auswertung der Studie betraut sind,

nicht über die individuelle Behandlungszuteilung der Patienten **informiert sind** und ihre Handlungen somit nicht durch dieses Wissen beeinflusst sein können.

Die Verblindung hat das Ziel, systematischen Unterschieden in der Behandlung der Patienten oder der Bewertung des Therapieerfolgs vorzubeugen.

24

Verblindungsniveaus:

unverblindet (offen)

- Chirurgie, Diät etc.
- Teilnehmer und Untersucher kennen Therapiegruppe
- Bias in beiden Richtungen

einfach-blind

- Teilnehmer blind

doppel-blind

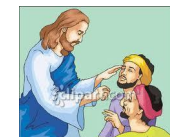
- Teilnehmer und Untersucher blind

dreifach-blind

- Teilnehmer, Untersucher, Auswerter etc. blind



Notfälle - zum Beispiel schwere Nebenwirkungen - können jedoch bei einzelnen Probanden die verfrühte Aufdeckung der Zuordnung zu den Untersuchungsgruppen notwendig machen (Entblindung)



25

Phasenmodell klinischer Studien

Phase 0: Präklinische Entwicklung

Ziele:

1. Abklärung möglicher **toxischer Effekte**, wie Einfluss auf zahlreiche in Laboruntersuchungen bestimmte Größen (Klinische Chemie, Hämatologie), Fertilität, Embryotoxizität/Teratogenität, Cancerogenität
2. Abklärung **sicherheitspharmakologischer Aspekte**, wie Beeinträchtigung von Herz/Kreislauf, Einfluss auf Körpergewicht
3. Hinweise auf **erwünschte pharmakologische Effekte** in vitro/in vivo



- Fallzahlsschätzung, Konfidenzintervalle
- Survival-Analyse
- Äquivalenznachweise
- Randomisation und Verblindung
- **Phase I-IV - Studien**

Phase I: Erstanwendung am Menschen

Meist gesunde Freiwillige („Probanden“), gegebenenfalls besondere Patientengruppe (z. B. bei Studien mit Zytostatika)

Ziele:

- Verträglichkeit, Pharmakokinetik/-dynamik
- Hinweis auf wirksame Dosis (eventuell)/ Arzneimittelinteraktionen

Phase II: Einstieg in die therapeutische Anwendung am Patienten

Begrenzte Zahl von Patienten der anvisierten Indikation

Ziele:

- Verträglichkeit und Dosisfindung
- Wirkung (pharmakologische Effekte)/Wirksamkeit (Heilerfolg)
- Pharmakokinetik in Spezialfällen (z. B. Leber-, Nierenerkrankung)

Phase III: Breite Anwendung im anvisierten Indikationsgebiet, Beleg für die

Einsetzbarkeit als Arzneimittel [Zulassung] Patienten der anvisierten Indikation in Klinik/Praxis

Ziele:

- Beleg der Wirksamkeit an Patienten in unterschiedlichen Populationen
- Ausreichende Beurteilung der Verträglichkeit, besondere Patientengruppen
- Verhalten unter Langzeitbehandlung, Vergleich mit etablierter Therapie

Phase IV: Klinische Prüfung nach der Zulassung: Erkenntniserweiterung über die Substanz, Einsatz unter Praxisbedingungen

Einsatz an großer Zahl von Patienten entsprechend den Vorgaben der Zulassungsbehörden (unter Praxisbedingungen)

Ziele:

- Quantifizierung seltener Nebenwirkungen
- Detailuntersuchungen in bestimmten Patientengruppen
- Einfluss auf Spätfolgen einer Erkrankung (Folgemorbidität, Letalität)
- Tatsächlicher Einsatz des Präparates (→ „Anwendungsbeobachtungen“)
- Hinweis auf weitere Indikationen, zu modifizierende Dosis (→ Phase II)

Grundbegriffe der Informatik
Rolle/Funktion der medizinischen
Datenbanken in Praxis und Forschung

**Medizinische-/Bio-Informatik als inter-
(multi-)disziplinäres Gebiet**

- ✓ **Biologie**
- ✓ **Biotechnologie**
- ✓ **Entwicklungslehre**
- ✓ **Physiomek* (ab Niveau der Gene bis Zusammenfunktion von Geweben, Organen)**
- ✓ **Genomik***
- ✓ **Informationstechnologie**
- ✓ **Mathematik**
- ✓ **Molekülmodellierung**
- ✓ **Proteomik**
- ✓ **Statistik**

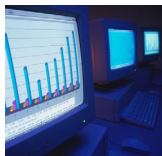
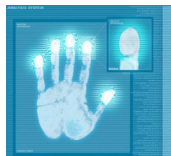
*Empfohlen als wertvolles zu lesen

<http://www.uni-heidelberg.de/presse/news/2106bartram.html>

Themen:

- I. Begriff und Maß der Information
- II. Codierung. Wirkungsgrad
- III. Genetischer Code, sein Informationsgehalt
- IV. Bio-Datenbanken

Begriff und Maß der Information



Ein Zitat von Augustinus (*354 — †430):

„Was ist also **die Zeit**? Wenn mich niemand danach fragt, weiß ich es, wenn ich es aber einem, der mich fragt, erklären sollte, weiß ich es nicht.“



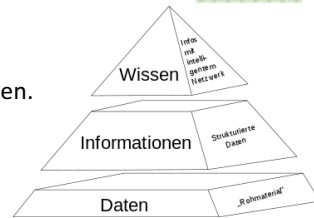
„Was ist also **die Information**? Wenn mich niemand danach fragt, weiß ich es, wenn ich es aber einem, der mich fragt, erklären sollte, weiß ich es nicht.“

Was ist die Information?



"informatio":

lat.: „bilden“, „eine Form, Gestalt, Auskunft geben.



Information:

- Wissen (s. Wikipedia: von althochdeutsch *wizzan*; zur indogermanischen Perfektform *woida* "ich habe gesehen"), Kenntnis von jemandem/etwas;
- Kenntnis auf Grund Nachrichten;
- Kenntnis/Wissen von einem gegebenen Umstand/Prozess;

Information als Begriff der Informatik:

Information ist diejenige Bedeutung, welche durch eine Nachricht getragen ist.



weitere Definitionen:

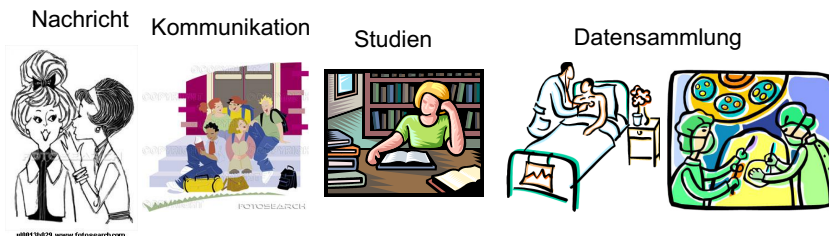
die Information:

- eine neue Kenntnis, die die Ungewissheit/Unbestimmtheit vermindert.
- **Reihenfolge/Struktur der Zeichen**, worin die Zeichen mit bestimmten Wahrscheinlichkeiten auftreten;
- ihre Bedeutung beigemessen werden kann – besitzt Bedeutungsgehalt;
- treibt den Empfänger/Adressat zu einem bestimmten Verhalten, einer Bewegung an

Information enthält – auf Grund einer anderen Definition - "**sinnvolle Daten**";

- besitzt mehreren Formen;
- kann auf unterschiedlichen Datenträgern gespeichert werden, vorhanden sein.

Wie kann man Informationen erwerben?



Reihenfolge/Struktur der Zeichen, worin die **Zeichen** mit bestimmten Wahrscheinlichkeiten auftreten

Zeichen (entsprechend der Bilder);

- ✓ Stimme, Worte(Wörter), Klang/Intonation;
- ✓ Buchstaben, Worte, Sätze, Kontext;
- ✓ die den physiologischen Zustand beschreibenden Eigenschaften, Charakteristiken



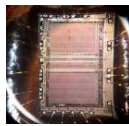
Quelle und Speicherung der Informationen

Als "sinnvolle Daten", kann die Information in unterschiedlichen Formen, auf verschiedenen Datenträgern gespeichert werden.

Speicherung (Z.B.):

bei Computern:

- ✓ magnetisch,
- ✓ optisch,
- ✓ integrierte Schaltkreise (ROM, RAM,...)
- ✓ usw.:



in medizinischer Praxis:

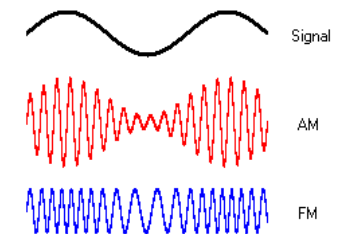
- die primäre Quelle ist der Patient;
- Speicherung der gewonnenen diagnostischen Testwerte;

$$2008_{(10)} = 11111011000_{(2)}$$

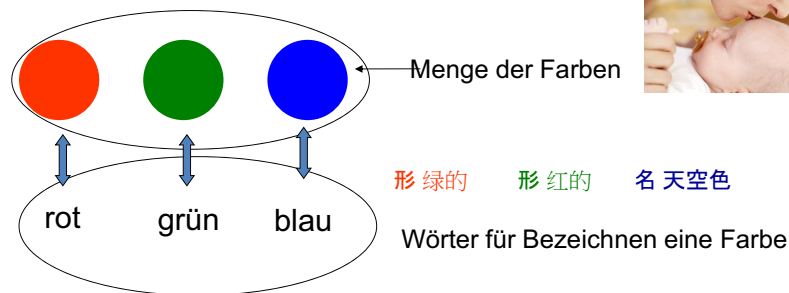
1 bit: eine einzige Stelle für Datenspeicherung in Computern (Taschenrechnern);
1 byte: acht bit

SI: 1kbit=10³ bit; (meistens) in Informatik: 1kbit = 1024 bit

Code



Codierung – Decodierung (Ver/Ent-schlüsselung)



gegenseitig-eindeutige Zuordnung zwischen den Elementen zwei Mengen

Sender: sendet/speichert /codiert Informationen in verschlüsselter Form;
Empfänger: empfängt und entschlüsselt die übertragenen Informationen

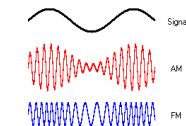
Die Rolle und Funktion der Codierung

✓ Speicherung und Übertragung der Informationen durch Anwendung ein bestimmtes Zeichensystem

z.B.: Morse-Code
Pheromone
DNS-Sequenzen
Hologramm

Zeichensysteme:

Daten;
Zahlen;
Zeichen (z.B: Piktogramme, Hieroglyphen);
Buchstaben;
Aminosäuren (im Aufbau von Proteinen);



Bedingungen:

- ✓ Vereinbarung zwischen dem Sender und Empfänger in den Formulierungen und Regeln (eingeschlossen die Übertragungsmethode) der Informationen z.B.: die Zeichenfolge "blau" muss zweiseitig das selbe bedeuten;
- ✓ die Zeichensätze müssen für den Sender und den Empfänger bekannt sein;



Zusammenfassung I.

Information — Codierung

- ✓ Beschreibung, Speicherung und Übertragung einer Erscheinung, Eigenschaft mit Hilfe eines Zeichensystems (Codierung);
- ✓ angenommen, dass der "Sender" und der "Empfänger" gleichzeitig oder nacheinander anwesend sind (Informations-übertragung/fluss)
↔ die Information existiert nicht allein/selbstständig

Fragen

- Wie groß ist der Informationsgehalt einer Information?
- Wie kann man effizient Codieren?
- Wie wäre es möglich den Informationsübertrag im Allgemeinen zu beschreiben?



Informationsgehalt

- der Patient hat einen lockeren Zahn
- alle Zähne eines Patienten sind locker

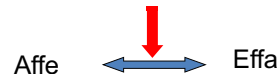
Welche dieser Informationen weist größeren Gehalt auf?

Auf Grund Gefühle, Eingebung:

eine Information mit geringerer Wahrscheinlichkeit weist größeren Informationsgehalt auf.

Auf Grund einer Definition der Information:

- Reihenfolge/Struktur der **Zeichen**, worin die **Zeichen** mit bestimmten Wahrscheinlichkeiten auftreten



derselbe Informationsgehalt

Informationsgehalt der statistisch unabhängigen Ereignissen

Bezeichne **p** die Wahrscheinlichkeit eines gegebenen Ereignisses (d.h. jetzt Zeichen)

Der Informationsgehalt, $I(p)$, dieses Zeichens ist:

Definition 1.:

$$I(p) = \log_2 \left(\frac{1}{p} \right) = -\log_2(p) \quad [I] = \text{bit oder } sh$$

sh: nach dem Namen Claude Shannon, der Begründer der Informationstheorie

Definition 2.:

Der Informationsgehalt ist gleich der minimalen Anzahl der Bits (in sh-Einheiten), die benötigt sind um ein Zeichen der Wahrscheinlichkeit p , verschlüsselt mit minimalen Zeichensatz, effizient zu übertragen:

$$I(p) = -\log_2(p) \quad [I] = \text{bit v. sh}$$

Beispiel:

- $p = 0,5 \quad I(p) = -\log_2(0,5) = -\log_2(1/2) = \log_2(2) = 1;$
- $p = 0,25 \quad I(p) = -\log_2(0,25) = -\log_2(1/2^2) = 2 \cdot \log_2(2) = 2$

Je kleiner die Auftretswahrscheinlichkeit eines Zeichens ist, desto größer ist sein Informationsgehalt.

$I(p=1) = -\log_2(1) = 0$ → besteht die "Botschaft" nur aus einem einzigen Zeichen, ist sein Informationsgehalt gleich Null

Quiz: Wenn die relative Häufigkeit eines Zeichens $p=0,0625$ ist, wie viele Bits sind nötig für die maximal-effiziente Übertragung?

$$I = -\log_2(0,0625) = -\lg(0,0625)/\lg(2) = 4 \text{ bit}$$

Informationsgehalt einer Zeichenfolge

m : Anzahl der unterschiedlichen Ereignisse (m -Ausgänge eines Versuches; d. Anzahl der Buchstaben, usw...);

p_k : relative Häufigkeit/Wahrscheinlichkeit eines Ereignisses;

N : Anzahl der allen Ereignissen ($= n_1 + n_2 + \dots + n_m$; Häufigkeiten)

Definition 3.:

$$I = \sum_{k=1}^m n_k I_k = -\sum_{k=1}^m [n_k \cdot \log_2(p_k)]$$

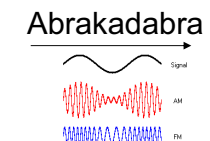
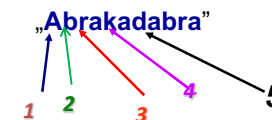
zu übertragen:

„abrakadabra“

$I=?$; wie viele Bits ist, minimal, benötigt zu einer Übertragung?

Wie kann man am effizientesten verschlüsseln, speichern?

$N=1$
 $m=5$



Zeichen	n_k (Häufigkeit)	f_k (rel. Häufigkeit)	$I_k = -n_k \cdot \log_2(f_k)$	
a	5	0.4545	5.688	
b	2	0.1818	4.919	
r	2	0.1818	4.919	
k	1	0.0909	3.459	
d	1	0.0909	3.459	
N=	11			
		$\Sigma I = 22.444$		

Eine effiziente Übertragung benötigt 23 bit für "Abrakadabra"

Wie kann man effizient kodieren?

Ziel:

der minimale Aufwand an Zeit, Energie,...

- ✓ Speicherung
- ✓ Übertragung

Lösung:

1. gemäß dem Informationsgehalt (d.h. mit minimaler Anzahl der benötigten Bits)
2. Zuteilung der kürzesten Codes zu Zeichen mit höchster Wahrscheinlichkeit.



der durchschnittliche Informationsgehalt – Entropie einer Information



$$\bar{x} = \frac{\sum x_i}{N}$$

Definition 4.:

$$\bar{I} = \frac{\sum_{k=1}^m n_k \cdot I_k}{N} = - \sum_{k=1}^m \left[\frac{n_k}{N} \cdot \log_2(p_k) \right] = H$$

H: Entropie einer Versuchsserie/Zeichenfolge; Einheit: bit

$$H = \bar{I} = - \sum_{k=1}^m [p_k \cdot \log_2(p_k)]$$

Informationsgehalt der genetischen Codes

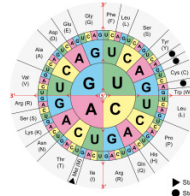
Fragen:

- 1. wie groß sei der minimale Zeichensatz (Anzahl der unterschiedlichen Nukleotiden Codon) für Kodierung der ~ 20 Aminosäuren?
- 2. Wie groß ist der Informationsgehalt einer DNS-Sequenz?



Antworten 1.: (wie es bekannt ist 4)

- ✓ Nukleotidpaare ermöglichen nur einen Variationssatz $4^2=16$;
- ✓ Nukleotidtriplets (Codons) – $4^3=64$;
- Kodierung über Triplets ist minimal und ausreichend;
- einige Aminosäuren sind durch mehreren Codons kodiert;
- einige Triplets besitzen andere Funktionen



Antwort 2:

- ✓ sei angenommen, dass die Nukleotidbasen treten gleichwahrscheinlich auf
- ✓ — $p_k=p=0,25$; $I_1=I_2=I_3=I_4=I_b$
- ✓ Wenn die Länge der Sequenz N ist, dann ist $n_k=N/4=n$

$$I = \sum_{k=1}^4 n_k I_k = nI_1 + nI_2 + nI_3 + nI_4 = 4 \cdot n \cdot I_b$$

$$I = 4 \cdot N / 4 \cdot I_b = N \cdot I_b = -N \cdot \log_2(p)$$

$$I = -N \cdot \log_2(0,25) = N \cdot 1,6021 \text{ bit}$$

$N=10 \rightarrow \sim 16 \text{ bit}$

$N=10^6 \rightarrow \sim 1,6 \cdot 10^6 \text{ bit}$

Zusammenfassung II.



- der Informationsgehalt einer Information kann mit der Informationsentropie beschrieben werden;
- der maximale Wirkungsgrad, die größte Effizienz, einer Kodierung wird mit einer dem Informationsgehalt entsprechenden minimalen Zeichensatz erreicht.
- Die durch DNS- oder Proteinmoleküle getragenen Informationen kann auf Grund der Häufigkeiten der monomer Nukleotiden oder Aminosäuren berechnet werden.

Bioinformatische Datenbanken



Zielsetzung, Aufgabe:

- Speicherung,
- Organisierung,
- Qualitätskontrolle,
- Analyse,
- der Öffentlichkeit zugänglich machen

der Daten, Wissen, Kenntnisse der biologischen, medizinischen Wissenschaften

Anforderungen:

- ✓ schnell und effizient Zugriff;
- ✓ Auffinden nur für die Benutzer wichtige, wesentliche Informationen.

spezialisierte Datenbanken:

Vorteil: kürzer Zugriffszeit, mehr detaillierte Daten

Nachteil: Mangel an Zusammenhängen

weniger spezialisierten Datenbanken:

Vorteil: auch die Zusammenhänge zwischen den Daten/Erscheinungen sind durchsuchbar.

Nachteil: mehrere Gesichtspunkte sind gebraucht für Auffinden einer Kenntnis

Z.B.: cholesterol — 182358
cholesterol transport — 9055
cholesterol transport pediatrics — 128

cholesterol transport pediatrics Chan T. — 2

Jelinek D, Patrick SM, Kitt KN, **Chan T**, Francis GA, Garver WS.: Physiological and coordinate downregulation of the NPC1 and NPC2 genes are associated with the sequestration of LDL-derived cholesterol within endocytic compartments. J Cell Biochem. 2009 Sep 10. [Epub ahead of print]

PMID: 19746448

Sahoo D, Trischuk TC, **Chan T**, Drover VA, Ho S, Chimini G, Agellon LB, Agnihotri R, Francis GA, Lehner R. ABCA1-dependent lipid efflux to apolipoprotein A-I mediates HDL particle formation and decreases VLDL secretion from murine hepatocytes. J Lipid Res. 2004 Jun;45(6):1122-31. Epub 2004 Mar 1.

GenBank from NCBI (National Center for Biotechnology Information) Genetic Sequence Databank;
EMBL Nucleotide Sequence Database (European Molecular Biology Laboratory);
SwissProt és **PROSITE** (protein sequence database);
EC-ENZYME ;
RCSB PDB (3-D makromolekularer Aufbau);
MEDLINE: Medizin, Zahnmedizin, Veterinärmedizin, forschungsmedizinische Informationen,...
PUBMed (<http://www.ncbi.nlm.nih.gov/sites/entrez>): Medizin, Biologie, Biochemie,...

Innerhalb der Universität:

<http://www.lib.sote.hu/>

Quellen

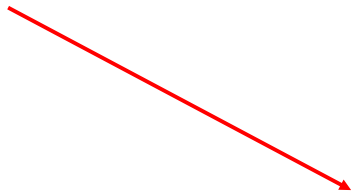
Datenbanken

Wissenschaftliche Artikel

Bücher

szientometrische/bibliometrische Datenbanken

pharmazeutische Datenbanken



C. Shannon (1916-2001)

