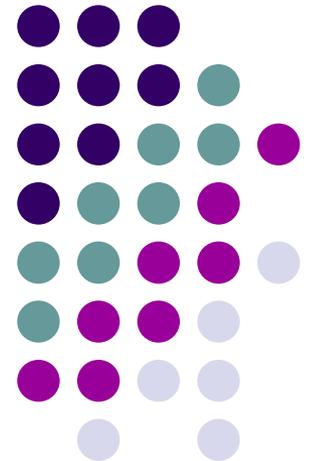


# Statistik in der genetischen Forschung

Anja Victor, Merck KGaA

Disclaimer: Die folgenden Ausführungen stellen die Meinung der Referentin dar und stimmen nicht notwendigerweise mit denen der Merck KGaA überein





# Fokus

- Hilfe zum Verständnis statistischer Ergebnisse
- Hinweise auf Validitätsmerkmale
- Anwendungsbeispiele
  - Genetische Assoziationsstudien
  - Genexpressionsanalysen zur Prognose von
    - Therapieansprechen
    - Überlebenszeiten/Metastasierung bei Krebserkrankungen



# Inhalt

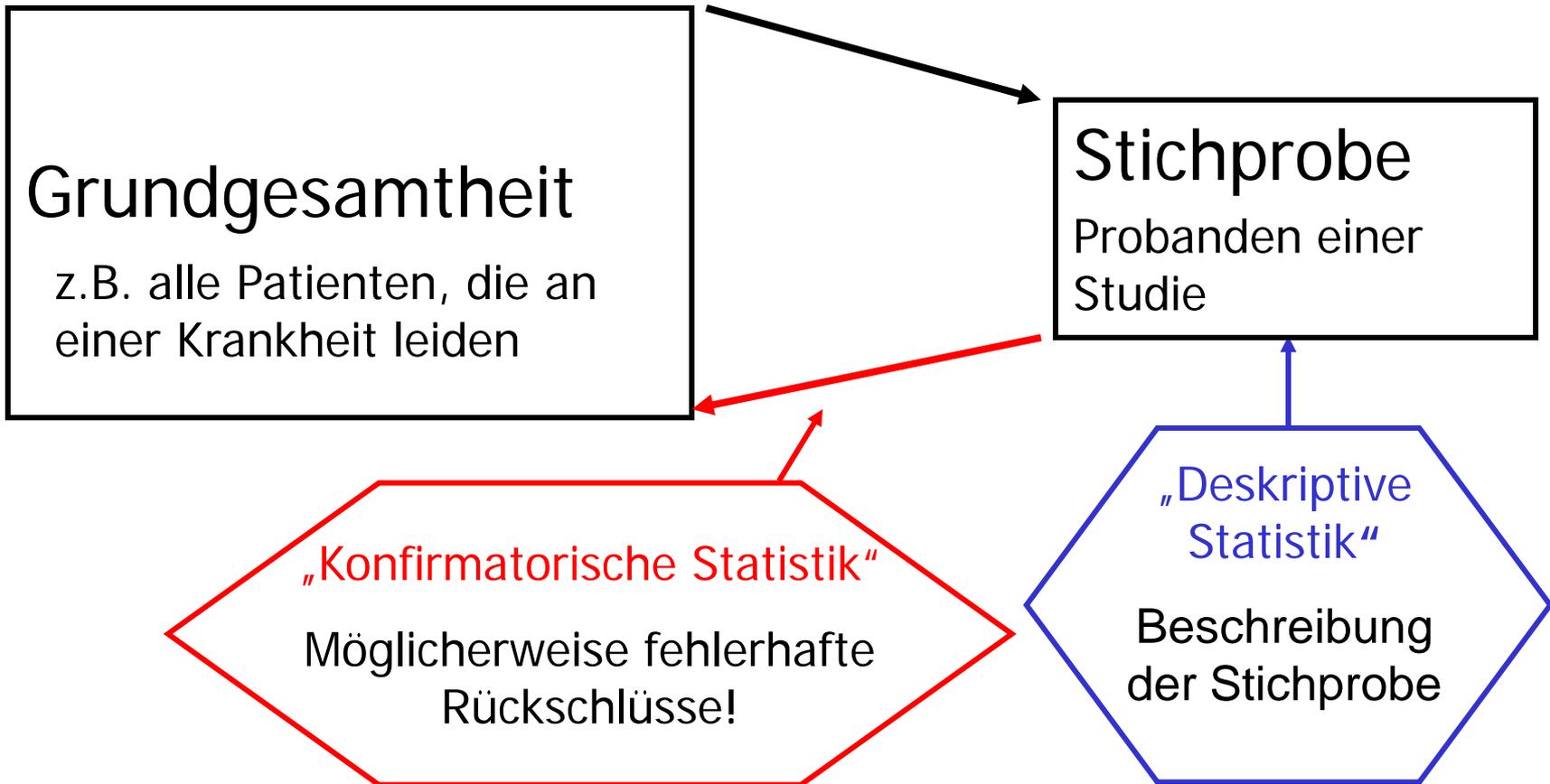
- Grundlagen
  - Was ist Statistik?
  - Genetische Assoziationsstudien
- Statistisches Testen (p-Werte)
- Multiples Testen
- Multivariable Prognosemodelle
- Maßzahlen und Interpretation
- Zusammenfassung



# Grundlagen



# Was ist Statistik?





# Auswahl der Stichprobe

- Zufallsstichprobe?
- Retrospektiv? Prospektiv?
- Repräsentativ? Für was ?
  - andere Kollektive, andere genetische Eigenschaften?

# Genetische Assoziationsstudien



In der Regel: Fall-Kontrollstudien

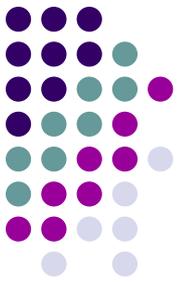
- Fälle:
  - Patienten mit gewisser Erkrankung (z.B. Brustkrebs, Atherosklerose.....)
  - Patienten, die auf gewisse Therapie ansprechen
- Kontrollen:
  - Personen ohne diese Erkrankung
  - Patienten, die auf gewisse Therapie nicht ansprechen
- Zumeist Untersuchung auf Allelverteilung an (vielen) Single Nucleotide Polymorphisms (SNPs)
  - z.B. Genomweite Scans (SNPs über das ganze Genom verteilt werden untersucht)

# Genetische Assoziationsstudien



Variable	Patienten N=50	Kontrollen N=50	p-Wert
Ausprägung des rs12345678 (SNP im XYZ Gen)			0.005
AA	10 (20%)	2 ( 4%)	Konfirmatorische (schliessende) Statistik
AT	25 (50%)	19 (38%)	
TT	15 (30%)	29 (58%)	

Deskriptive  
(beschreibende)  
Statistik



# Statistisches Testen



# Was sind statistische Tests

Beispiel: Bei 10mal würfeln fällt 10mal die 6.

- Ist der Würfel fair? Kann 10 mal die 6 zufällig sein?
- Wäre der Würfel fair, käme dieses Ergebnis äußerst selten vor, nämlich mit einer Wahrscheinlichkeit von  $(1/6)^{10}=0.00000002$  (“p-Wert”).

Dies spricht gegen die (Null)Hypothese, dass der Würfel fair ist



# Statistische Tests

- Alternative:  $H_1$  = Assoziation zwischen SNP XY und Auftreten von Brustkrebs
- Nullhypothese:  $H_0$  = **Keine** Assoziation zwischen SNP XY und Auftreten von Brustkrebs

Die (unbekannte) Wahrheit  
(Gesamtkollektiv)

		$H_0$	$H_1$
Das Fazit aus der Studie (Stichprobe)	$H_0$	✓	Fehler 2. Art
	$H_1$	Fehler 1. Art	✓

# Statistische Tests



- Fehler 1. Art und 2. Art können in Analyse nicht gleichzeitig kontrolliert werden
- Konvention: Fehler 1. Art wird kontrolliert durch Signifikanzniveau  $\alpha$  (maximal tolerable Fehlerwahrscheinlichkeit), zumeist 5%
- Entscheidung für Alternative  $H_1$  kann also zu maximal  $\alpha$  irrtümlich sein

# Statistische Tests



1. Hypothesenformulierung
2. Datensammlung
3. Berechnung p-Wert aus den Daten
4. Vergleich p-Wert mit Signifikanzniveau  $\alpha$
5.  $p \leq \alpha$  Entscheidung für  $H_1$  („signifikant“) oder  $p > \alpha$  Beibehaltung von  $H_0$

# Statistische Tests



Signifikanz  
( $p \leq \alpha$ )

≠

Klinische Relevanz

(z.B. bei großen Fallzahlen  
werden auch kleine Effekte  
signifikant)

Nichtsignifikanz  
( $p > \alpha$ )

≠

Beweis der  
Nullhypothese

(z.B. keine Nicht-Assoziation  
folgerbar)



# Multiple Testen

# Problem des multiplen Testens

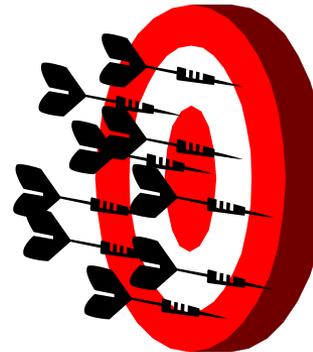


Fragestellung unklar;  
Durchführung vieler  
Analysen



1

Ergebnisse



2

Selektive Darstellung  
der Ergebnisse;  
Überinterpretation



3

# Problem des multiplen Testens



- „ein regelmäßiges **Frühstück mit Getreideflocken** scheint eher die Geburt eines Jungen zu begünstigen.“ (*Spiegel, 2008*)
- „**Müsli macht Männer**“ (*Süddeutsche Zeitung, 2008*)
- Die Studie
  - Mathews F, Johnson JJ, Neil A (2008) *Proceedings of the Royal Society B Vol 275:1661-1668*
  - Befragung von 740 Frauen bzgl. 133 Lebensmitteln
  - U.a. 266 statistische Tests bezüglich der einzelnen Lebensmittel
  - Ohne Korrektur für multiples Testen:
    - Wahrscheinlichkeit für mindestens ein fälschlich signifikantes Ergebnis >99%\*
    - sind  $266 \cdot 5\% = 13$  zufällige Signifikanzen zu erwarten
- Was könnte zur Klärung beitragen?
  - Z.B. weitere, prospektive Studie zur Replikation der Ergebnisse

\* bei Unabhängigkeit

# Beispiel aus genetischen Assoziationsstudien



- Replikationsstudien zeigen geringere Risikoschätzer als Originalstudie

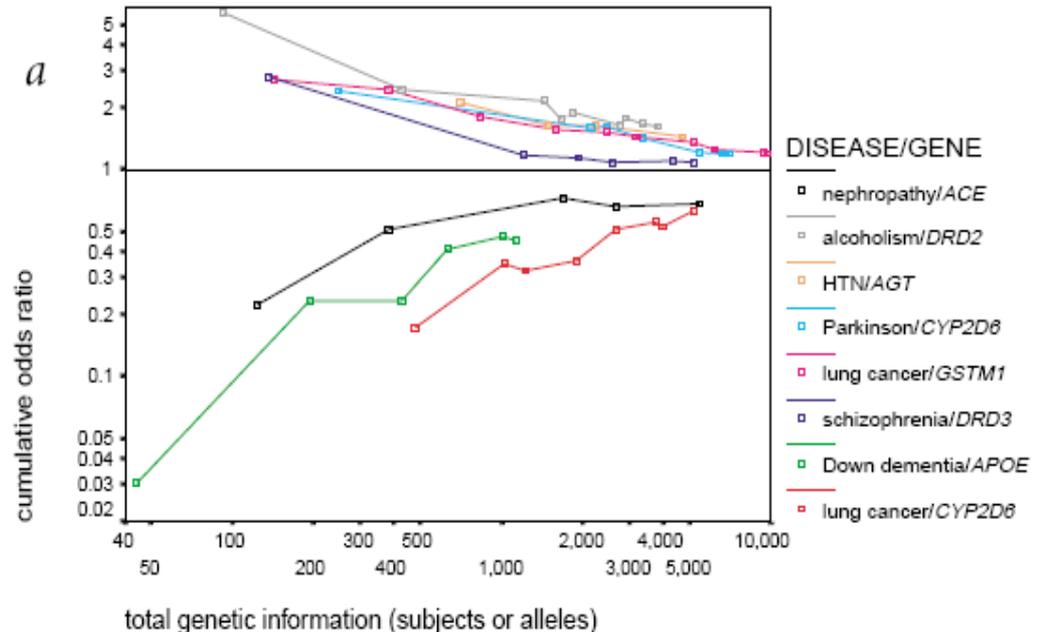
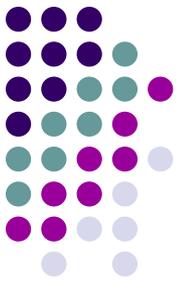


Abbildung aus: Ioannidis JPA, Ntzani EE, Trikalinos TA, et al. (2001), Nature Genetics Vol. 29: 306-309



# Multivariable Prognosemodelle (z.B. in Genexpressionsanalysen)

# Multivariate Prognosemodelle



Alle Bildnachweise am Ende des Vortrags

## Giftige Blumen

- Alpenveilchen:
  - Rosa Blüten
  - herzförmige Blätter
  - Knolle
- Schachblume
  - Rosa/lila Blüte
  - Glatte, längliche Blätter
  - Zwiebel



## Ungiftige Blume

- Löwenzahn:
  - Gelbe Blüte
  - Gezackte Blätter
  - Keine Zwiebel
- Fingerkraut
  - Gelbe Blüten
  - Gefiederte Blätter
  - Keine Zwiebel



## → Zuordnungsschema:

Gelbe Blüten, gezackte/gefiederte Blätter, keine Zwiebel/Knolle: **Ungiftig**

Rosa/lila Blüten, Blätter mit glattem Rand, Zwiebel/Knolle: **Giftig**

# Multivariable Prognosemodelle



- Ranunkel:
  - Gelbe Blüten,
  - gezackte Blätter,
  - keine Zwiebel/Knolle



→ Ungiftig ? (nach obigem Zuordnungsschema)

- Schnittlauch:
  - Rosa Blüten,
  - glatte Blätter,
  - Zwiebel



→ Giftig ? (nach obigem Zuordnungsschema)

Alle Bildnachweise am Ende des Vortrags

# Multivariable Prognosemodelle



- Mögliche Probleme:
  - Mehr Zuordnungsvariablen als beobachtete Objekte vorhanden
    - Immer eine Zuteilung der beobachteten Objekte möglich
    - Aber nicht notwendigerweise verallgemeinerbar auf andere Objekte
  - Zuordnung auf Kollektiv optimiert
    - Risikoschätzer verzerrt
    - Trennschärfe bei Verallgemeinerung auf andere Kollektive?

# Genexpressionsanalysen



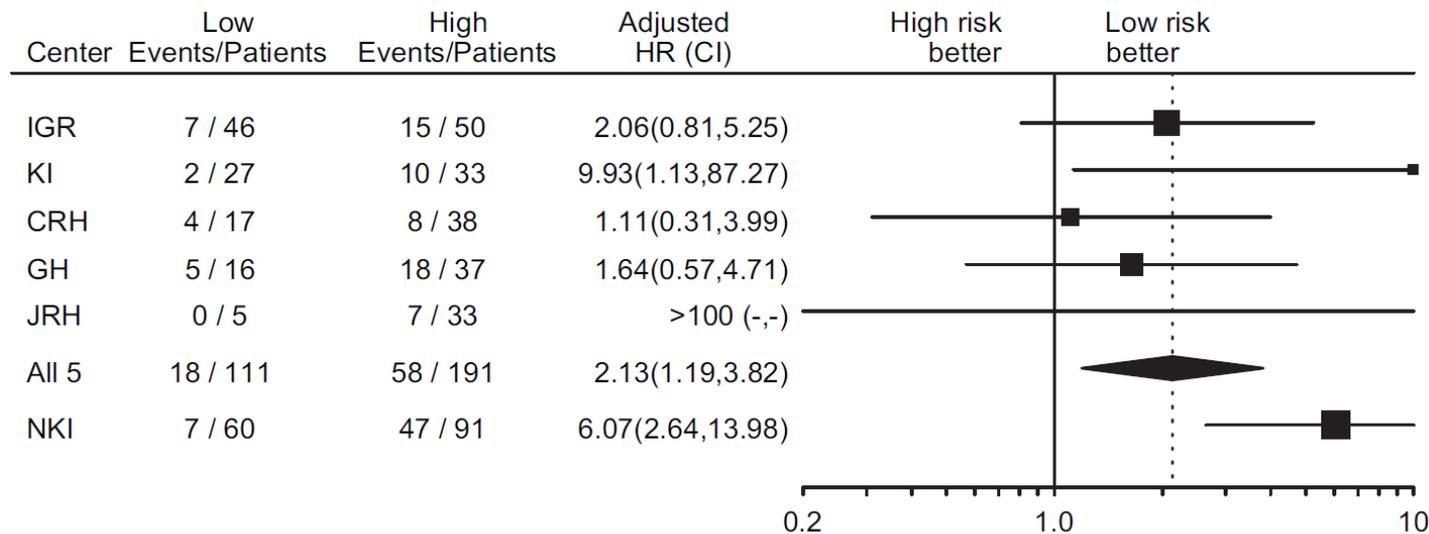
- z.B. zur Prognose von
  - Metastasierungswahrscheinlichkeit
  - Therapieansprechen
- Viele Tausend Gene werden auf Expression untersucht, um Patienten zuzuordnen
- Beispiel: Mammaprint®-Entwicklung
  - Vergleich der Genexpression weniger Hundert Patientinnen eines Behandlungszentrums zwischen Patientinnen (mit nodal-negativem Brustkrebs)
    - mit Metastasen nach 5 Jahren und
    - ohne Metastase nach 5 Jahren
  - Auswahl von 70 Gensequenzen
  - Entwicklung eines Algorithmus zur Zuordnung zu einer Hochrisikogruppe oder einer Niedrigrisikogruppe



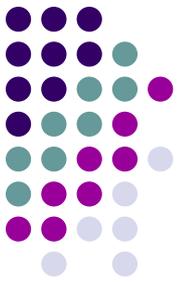
# Genexpressionsanalysen

- Beispiel: Mammaprint® Validierung
  - Validierung in anderen Patientinnenkollektiven
  - Prospektive Studie

**A**



Aus: Buyse M, Loi S, van't Veer L, et al. (2006) Journal of the National Cancer Institute 98:1183-1192



# Maßzahlen und Interpretation

# Maßzahlen und Interpretation



- Z.B. Genetische Assoziationsstudie
  - Tragen eines Risikoallels, Vergleich zwischen Fällen und Kontrollen ergibt für einen SNP:
    - P-Wert (Ergebnis zufällig?) 0.00000005
    - Odds Ratio („Risikoerhöhung“) 1.32
  - Und für weitere 7 SNPs ähnliche Ergebnisse
- Interpretation ?
- Berücksichtigung von Interaktionen?
- Konsequenz?



# Zusammenfassung

# Zusammenfassung



- Beachtet werden sollte
  - Auswahl der Stichprobe (häufig retrospektiv)
  - multiples Testen
  - Hochdimensionalität der Messungen
  - Qualität
- Auch, wenn für multiples Testen korrigiert wurde, sind statistische Tests immer
  - Schlüsse von einer Stichprobe auf größere Gesamtheit
    - Verallgemeinerung ?
    - Irrtumswahrscheinlichkeit in Höhe des Signifikanzniveaus
- Validierung/Replikation unabdingbar
- Interpretation der Ergebnisse nicht einfach

# Zusammenfassung



- Statistik entwickelt Methoden mit denen man auch sehr komplexe Daten analysieren kann
- Die korrekte Anwendung der Statistik ist unabdingbare Voraussetzung für die korrekte Interpretation der Ergebnisse
- Statistik alleine kann aber keine Entscheidung liefern, ob Ergebnisse klinisch relevant sind oder nicht !

# Literatur, Bildnachweise



- Mathews F, Johnson PJ, Neil A (2008) „You are what your mother eaty: evidence for maternal preconception diet influencing foetal sex in humans“ Proceedings of the Royal Society B Vol. 275: 1661-1668.
- Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) „Replication validity of genetic association studies“. Nature Genetics, Vol 29: 306-309
- Buyse, M., Loi S., van't Veer, L. et al. (2006) „Validation and Clinical Utility of a 70-gene prognostic signature for women with node-negative breast cancer“. Journal of the National Cancer Institute, Vol 98: 1183-1192.
- Alpenveilchen: Alpenveilchen Cyclamen 1.jpg, Urheber: Darkone, Wikimedia Commons, lizenziert unter CreativeCommons-Lizenz by-sa-2.0-de, URL: <http://creativecommons.org/licenses/by-sa/2.0/de/legalcode>
- Schachblume: Schachblume 3.jpg, Urheber: Marion Schneider & Christoph Aistleitner, Wikimedia Commons, lizenziert als gemeinfrei
- Löwenzahn: Gewöhnlicher Löwenzahn Taraxacum officinale agg. big.jpg Urheber: Darkone, Wikimedia Commons, lizenziert unter [Creative Commons-Lizenz Namensnennung-Weitergabe unter gleichen Bedingungen 2.5 US-amerikanisch \(nicht portiert\)](http://creativecommons.org/licenses/by-sa/2.5/deed.de), URL: <http://creativecommons.org/licenses/by-sa/2.5/deed.de>
- Fingerkraut: Potentilla erecta.jpeg, Urheber: Kristian Peters, Wikimedia Commons, lizenziert unter CreativeCommons-Lizenz by-sa-3.0-de, URL: <http://creativecommons.org/licenses/by-sa/3.0/de/legalcode>
- Ranunkel: Ranunculus carinthiacus 2.jpg, Urheber Franz Xaver, Wikimedia Commons, lizenziert unter CreativeCommons-Lizenz by-sa-3.0-de, URL: <http://creativecommons.org/licenses/by-sa/3.0/de/legalcode>
- Schnittlauch Allium schoenoprasum var alpinum.jpg, Urheber: Tigerente, Wikimedia Commons, lizenziert unter CreativeCommons-Lizenz by-sa-3.0-de, URL: <http://creativecommons.org/licenses/by-sa/3.0/de/legalcode>