# Ein Permutationstest auf Assoziiertheit der Haplotypenverteilung mit einer ordinalen Variable

K. Neumann

Institut für Biometrie und Klinische Epidemiologie
Charité, Berlin

# Einleitung

- Stichprobe: N=573 Patienten, die chronisch an Hepatitis C (HCV) leiden.

- Zwei Polymorphismen des CTLA4-Gens (Cytotoxic T-lymphocyte antigen-4) wurden typisiert:
  - im Promotorbereich (C->T, -318)
  - im Exon1 (A->G, 49)

# Einleitung (2)

- Möglicher Zusammenhang dieser Polymorphismen mit der Immunantwort und damit mit dem Verlauf der Hapatitis C Erkrankung.

- Der histologische Fibrosegrad (ordinal von 0 bis 4) beschreibt die Schädigung der Leber durch das Hepatitis C Virus (0 = keine Fibrose, 4 = Zirrhose)

# Fragestellung

- Gibt es einen Zusammenhang zwischen den Polymorphismen von CTLA4 und dem Fibrosegrad?

Insbesondere:

- Können „Risikohaplotypen“ angegeben werden?

# Haplotyp-Rekonstruktion

- Ist ein Individuum an mehr als einem (typisierten) Genlocus heterozygot, dann kann nicht eindeutig auf das Paar der Haplotypen geschlossen werden:

# Haplotyp-Rekonstruktion (2)

- Bei zwei Genloci ist das die einzige Mehrdeutigkeit.
- Allgemein gilt: Ist ein Individuum an k>1 Loci heterozygot, dann gibt es $2^{k-1}$ verschiedene Auflösung in Paare von Haplotypen.
- Bei k Loci gibt es maximal:
  - $3^k$ Phänotypen (=Genotypen ohne Information über die „Phase“).
  - $2^{k-1}(2^k+1)$ Genotypen (mit Information über die Phase).

# Haplotyp-Rekonstruktion (3)

- Direkte Bestimmung des Genotyps (mit Phase) ist sehr aufwändig (z.B. Genotypisierung naher Verwandter)
- Ausweg: Schätzung der Häufigkeit der Haplotypen in der gesamten Stichprobe unter anderem durch
  - EM Algorithmus
  - Bayesische Methoden (PHASE von M. Stephens)

# EM Algorithmus zur Schätzung der Häufigkeit der Haplotypen (E-Schritt)

Für jeden Genotyp (mit Phase) $H_i H_j$ ermittelt man den zugehörigen Phänotyp S und die Menge $G_S$ aller Auflösungen von $S=S(H_1 H_2)$ in Genotypen.

<u>Beispiel:</u> Für den Genotyp $H_1H_2$ mit

$H_1=(C,G)$ und $H_2=(T,A)$

ist

$G_S=\{(C,G)(T,A),(T,G)(C,A)\}$.

# EM Algorithmus (E-Schritt 2)

| Genotyp | Phänotyp | $G_s$ |
|---|---|---|
| $H_1H_1$ | WW | $\{H_1H_1\}$ |
| $H_1H_2$ | Wh | $\{H_1H_2\}$ |
| $H_1H_3$ | hW | $\{H_1H_3\}$ |
| $H_1H_4$ | hh | $\{H_1H_4, H_2H_3\}$ |
| $H_2H_2$ | MW | $\{H_2H_2\}$ |
| $H_2H_3$ | hh | $\{H_1H_4, H_2H_3\}$ |
| $H_2H_4$ | Mh | $\{H_2H_4\}$ |
| $H_3H_3$ | WM | $\{H_3H_3\}$ |
| $H_3H_4$ | hM | $\{H_3H_4\}$ |
| $H_4H_4$ | MM | $\{H_4H_4\}$ |

| | |
|---|---|
| $H_1$ | (w,w) |
| $H_2$ | (m,w) |
| $H_3$ | (w,m) |
| $H_4$ | (m,m) |

# EM Algorithmus (E-Schritt 3)

Mit den im g-ten Schritt geschätzten Häufigkeiten der Haplotypen $p_1^{(g)},\ldots,p_h^{(g)}$ berechnet man

$$P(H_iH_j)^{(g+1)} = \frac{N_s}{N}\frac{p_i^{(g)}p_j^{(g)}}{\sum_{H_mH_n\in G_s} p_m^{(g)}p_n^{(g)}}.$$

$N_S$: Häufigkeit des zu $H_iH_j$ gehörigen Phänotyps.

N: Stichprobenumfang.

# EM Algorithmus zur Ermittlung der Häufigkeiten der Haplotypen (M-Schritt)

Aus den $P(H_iH_j)^{(g+1)}$ wird die Schätzung der Häufigkeit der Haplotypen für die g+1–te Iteration gewonnen:

$$p_i^{(g+1)} = \frac{1}{2}\left( \sum_{j \neq i} P(H_iH_j)^{(g+1)} + 2P(H_iH_i)^{(g+1)} \right)$$

# EM Algorithmus zur Ermittlung der Haplotypen

- Konvergiert der Algorithmus für $g \to \infty$, dann hat man die Häufigkeiten

$$p_1, \ldots, p_h$$

der Haplotypen $H_1, \ldots, H_h$ geschätzt.

- Aus $p_1, \ldots, p_h$ kann für jedes Individuum der Erwartungswert für die Anzahl von Haplotyp $H_i$ (i=1,…,h) angegeben werden (0 bis 2).

# Test auf Zusammenhang der Verteilung der Haplotypen mit dem Fibrosegrad

- Mit diesen Erwartungswerten kann für jeden Haplotyp eine Rangsumme $R_i, i = 1, \ldots, h$ bezüglich des Fibrosegrades bestimmt werden.
- Als Teststatistik wird berechnet:

$$T = \sum_{i=1}^{h} \frac{R_i^2}{N_i} \quad \text{mit} \quad N_i = 2 p_i N.$$

# Test auf Zusammenhang der Verteilung der Haplotypen mit dem Fibrosegrad (2)

Die Statistik für

$$T = \sum_{i=1}^{h} \frac{R_i^2}{N_i}$$

wird durch Permutation der Fibrosegrade ermittelt. Für jede Permutation σ wird der zugehörige Wert der Statistik $T_\sigma$ berechnet.

# Test auf Zusammenhang der Verteilung der Haplotypen mit dem Fibrosegrad (3)

- Der P-Wert ergibt sich als

$$P = \frac{\#\left\{\sigma \text{ Perm. vom Grad N} \mid T \le T_\sigma\right\}}{N!}$$

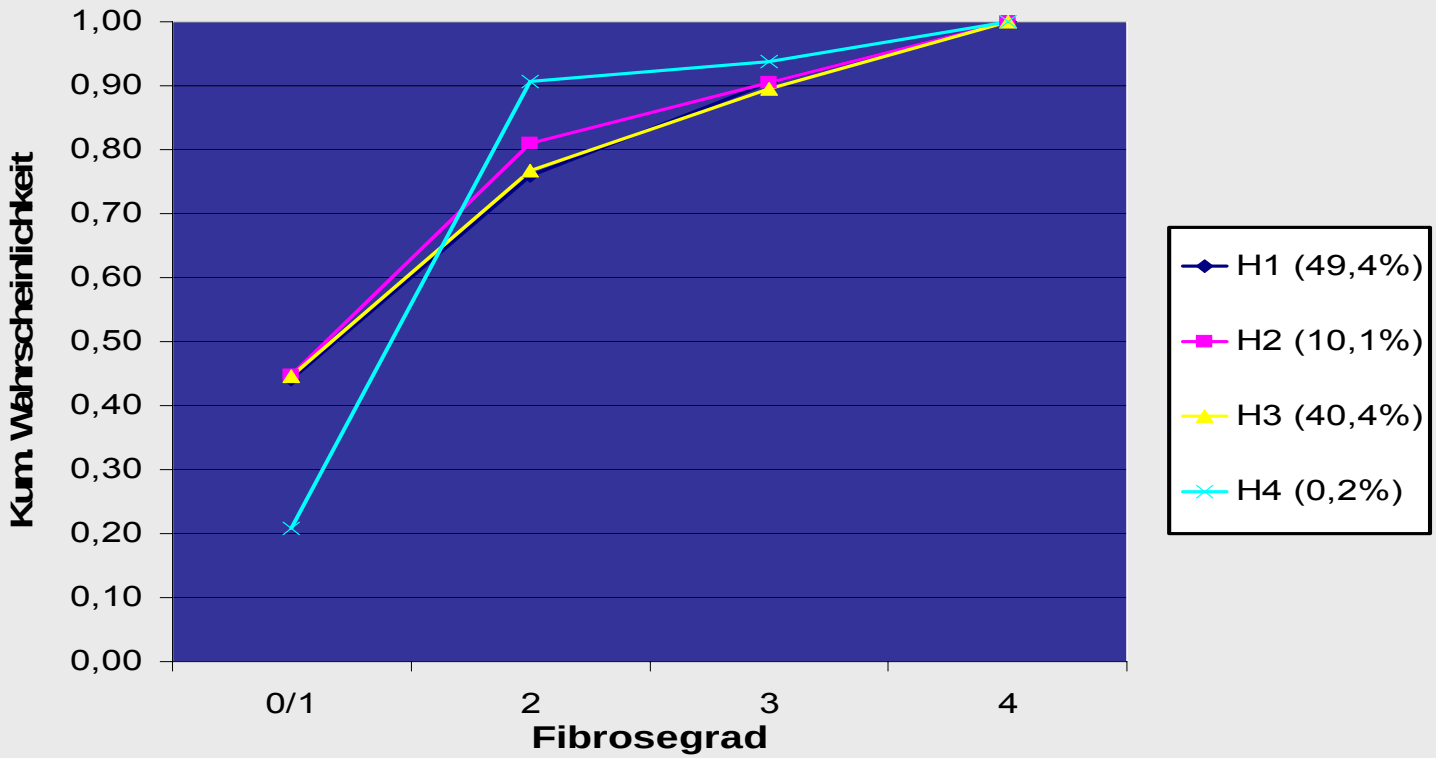
oder bei großem N näherungsweise durch

$$P_{MC} = \frac{\#\left\{i \mid T \le T_{\sigma_i}, i = 1,\ldots,N_{MC}\right\}}{N_{MC}}$$

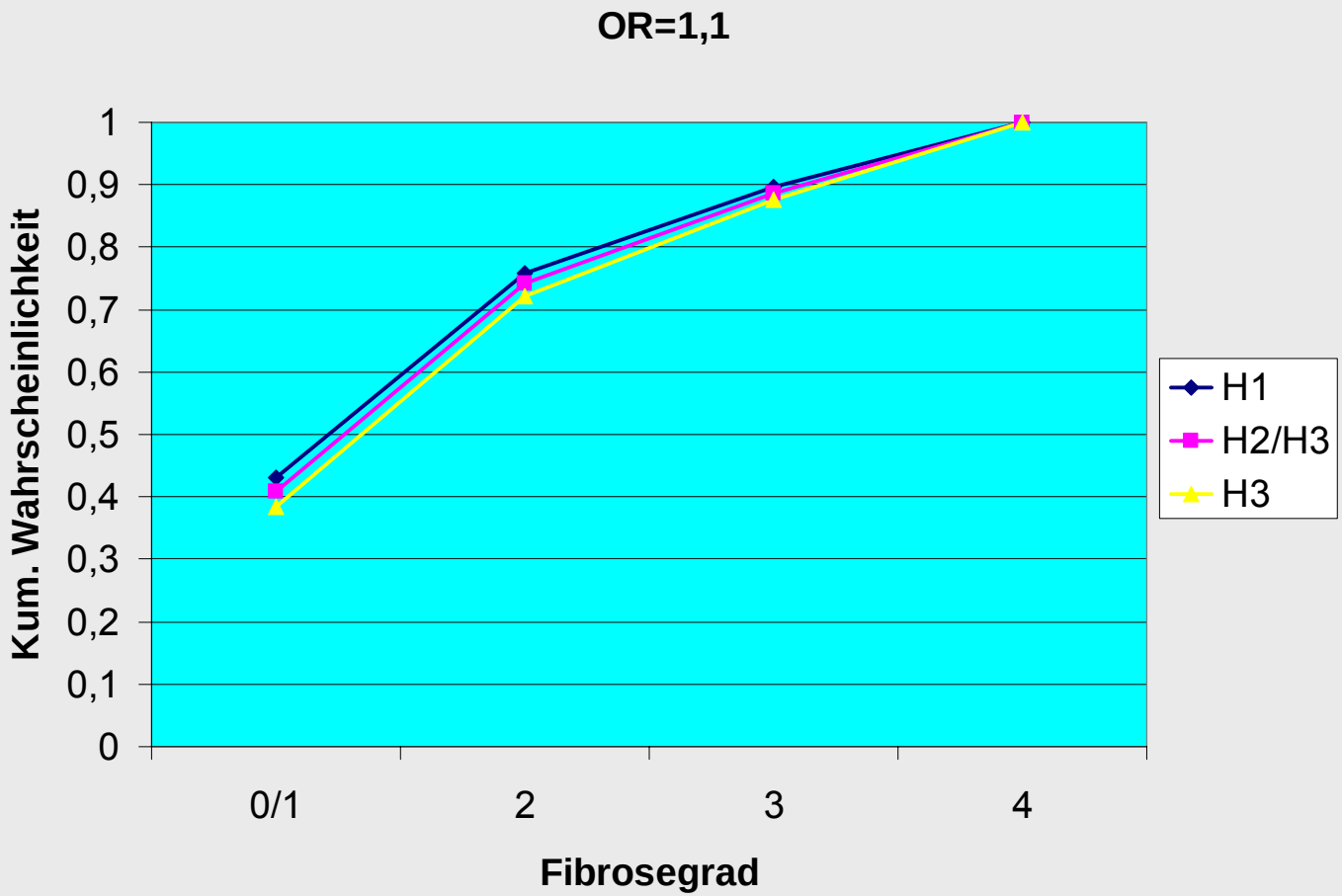
$\sigma_i$ (i=1,…,$N_{MC}$) zufällig ausgewählte Permutationen vom Grad N.

# Ergebnis

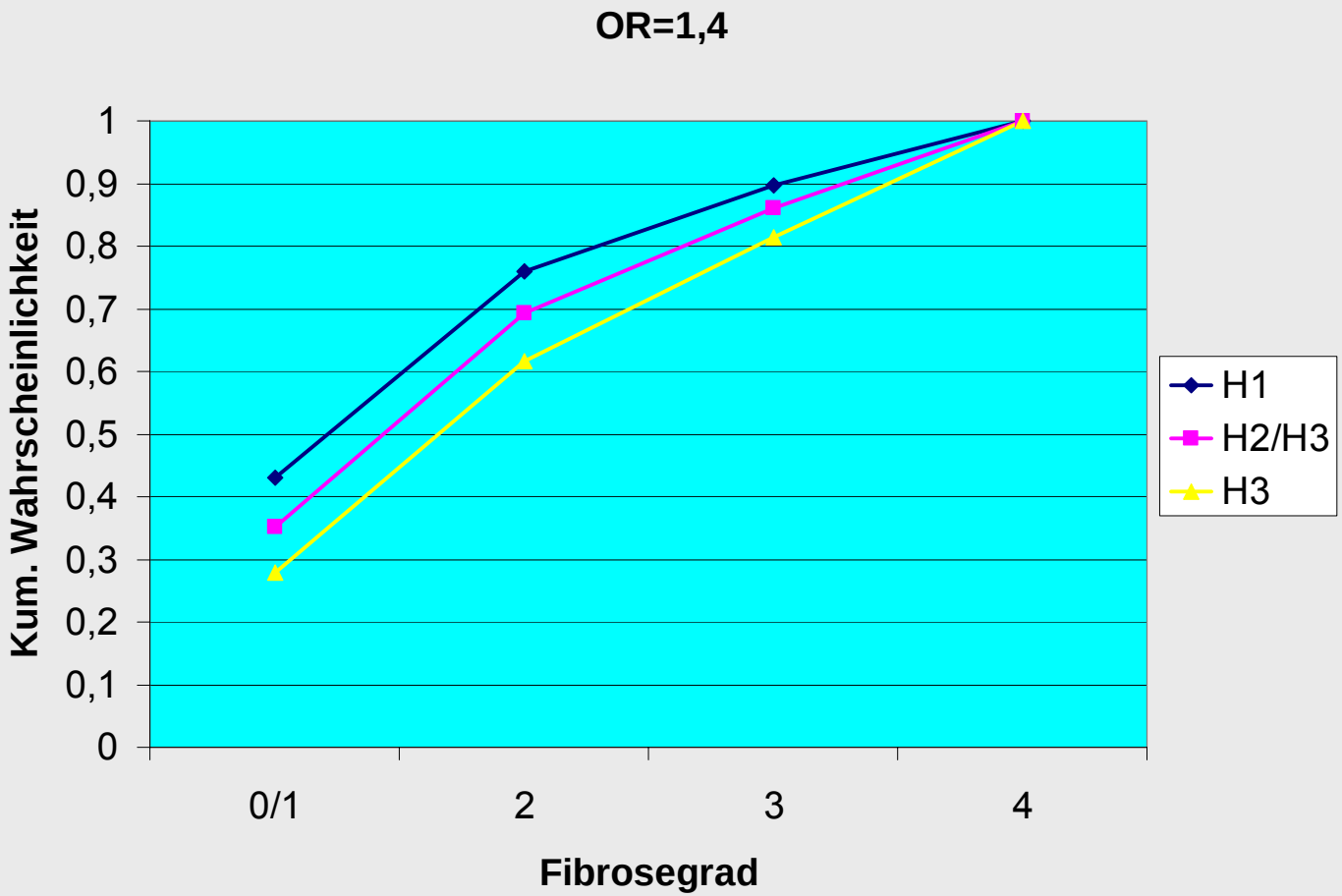
$$P = 0{,}96\ (N_{MC} = 10000)$$



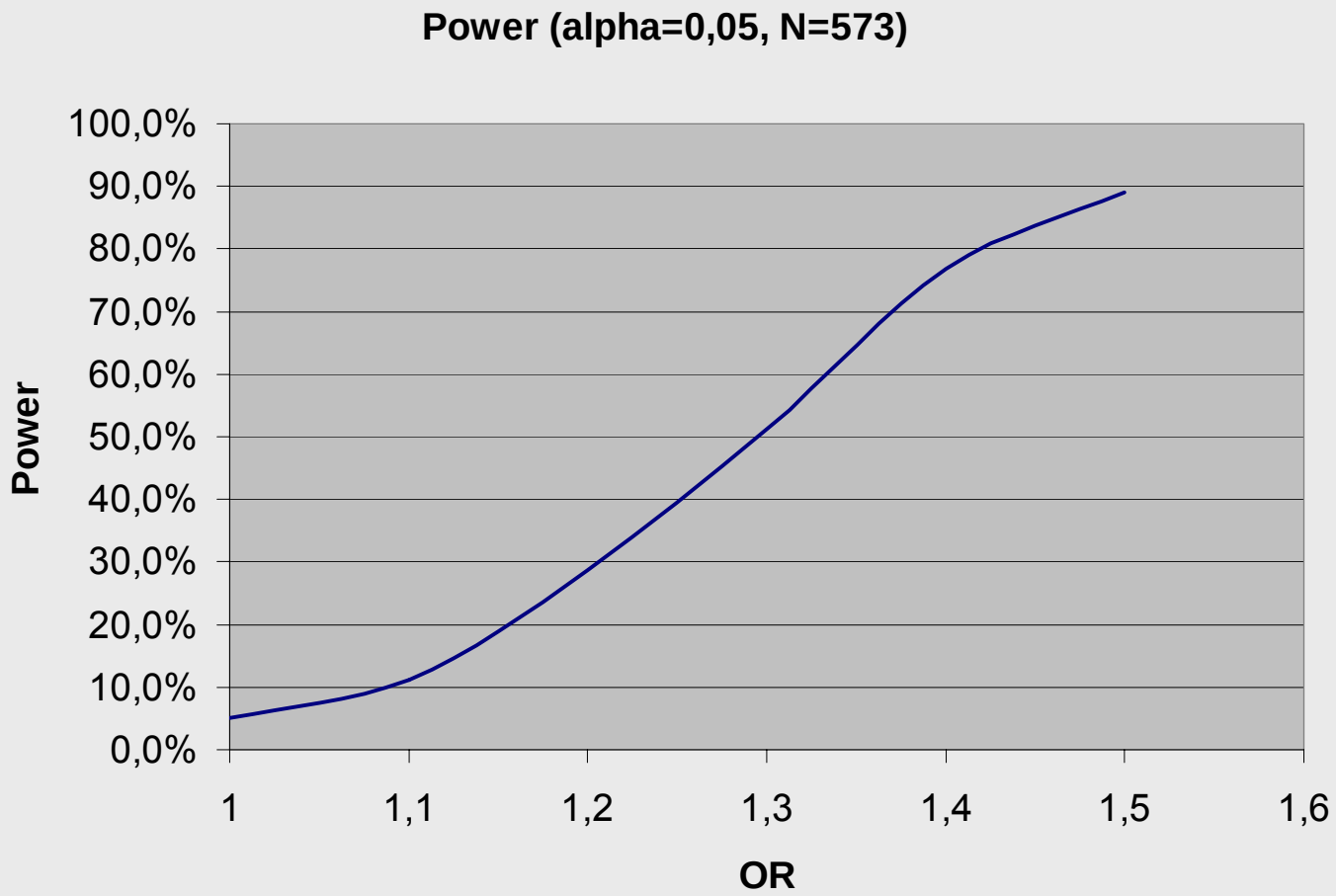
# Monte-Carlo Simulationen



# Monte-Carlo Simulationen (2)



# Monte-Carlo Simulationen (3)



# Monte-Carlo Simulationen (4)