medgen 2011 · 23:377–384
DOI 10.1007/s11825-011-0295-7
Online publiziert: 12. Oktober 2011

I.M. Heid[1, 2] · T.W. Winkler[1] · F. Grassmann[3] · B.H.F. Weber[3]
[1] Public Health und Gender Studies, Institut für Epidemiologie und Präventivmedizin, Universitätsklinikum Regensburg
[2] Institut für Epidemiologie I, Helmholtz-Zentrum München, Neuherberg
[3] Institut für Humangenetik, Universität Regensburg

Wie groß sind die kleinen genetischen Risiken?

Genomweite Assoziationsstudien (GWAS) basierend auf nichtverwandten Personen waren in den letzten Jahren sehr erfolgreich, um genetische Risikovarianten für multifaktorielle Erkrankungen und erkrankungsrelevante quantitative Parameter zu identifizieren. Dem gegenüber stehen familienbasierte Studien, mit welchen v. a. die Beschreibung von seltenen Varianten mit großen Effekten gelang. Man spricht oft davon, dass die Risiken von Varianten, die durch genomweite Assoziationsstudien als erkrankungsrelevant identifiziert werden, eher klein sind. Aber wie klein ist klein? Und wie werden diese Risiken berechnet? Im Folgenden wird beschrieben, wie genetische Effekte oder „Risiken" berechnet werden. An 3 Beispielen, Adipositas mit Body-Mass-Index (BMI), Diabetes und altersbedingter Makuladegeneration (AMD), werden diese Berechnungen illustriert.

Abkürzungen	
GWAS	genomweite Assoziationsstudie
BMI	Body-Mass-Index
AMD	altersbedingte Makuladegeneration
T2D	Typ-2-Diabetes
OR	Odds Ratio
PAR	Populationsattributables Risiko
MAF	„minor allele frequency", Häufigkeit des selteneren Allels
EAF	„effect allele frequency", Häufigkeit des Effektallels

Genomweite Assoziationsstudien

Berechnung von SNP-Assoziationen

In genetischen Assoziationsstudien wird untersucht, inwieweit die Gruppen von Personen mit unterschiedlichen Genotypen an einer bestimmten Stelle in der DNA („single nucleotide polymorphisms", SNPs) eine unterschiedliche Ausprägung im Phänotyp zeigen. Solche Phänotypen können z. B. der Body-Mass-Index (BMI, kontinuierlicher Phänotyp), Typ-2-Diabetes (T2D: ja/nein, dichotomer Phänotyp) oder altersbedingte Makuladegeneration (AMD: ja/nein, dichotomer Phänotyp) sein. Dabei wird untersucht, ob die Personen mit unterschiedlichen Genotypen entsprechend unterschiedliche Mittelwerte im BMI oder unterschiedliche Erkrankungswahrscheinlichkeiten für T2D oder AMD aufweisen. Wenn Assoziation nicht nur für einen SNP, sondern für ein dichtes, genomweit verteiltes SNP Panel, z. B. 0,5–2,5 Mio. SNPs, getestet wird, spricht man von einer *genomweiten Assoziationsstudie* (GWAS).

Als Assoziationsstudien eignen sich *Querschnittstudien*, in welchen quantitative Phänotypen (z. B. BMI) oder der Erkrankungsstatus (z. B. T2D oder AMD) bei einer Stichprobe der Allgemeinbevölkerung erhoben werden. Oft sind allerdings die Erkrankungen zu selten, sodass sich nicht genügend Erkrankte in einer Stichprobe der Allgemeinbevölkerung befinden. Dann eignen sich eher *Fall-Kontroll-Studien*, in welchen die Patienten in den Krankenhäusern oder über Arztpraxen rekrutiert werden; geeignete Kontrollgruppen sind dann Gesunde, die aus der selben Region kommen und sich außer durch den Erkrankungsstatus möglichst nicht von den Fällen unterscheiden. In einer solchen Fall-Kontroll-Studie kann der Erkrankungsstatus als dichotomer Phänotyp (erkrankt/nichterkrankt) in einer GWAS ausgewertet werden.

Identifizierung der genetischen Risikofaktoren durch GWAS

In einer GWAS wird jede genetische Variante (SNP) einzeln im Modell auf Assoziation getestet. Dabei wird für einen quantitativen Phänotyp Y (z. B. BMI in kg/m^2) ein *lineares Regressionsmodell*

$$Erwartungswert\ [Y \mid SNP] = a + b\ SNP$$

verwendet, wobei ein einzelner SNP die Werte 0, 1, 2 als mögliche Genotypwerte annimmt und damit ein additiver genetischer Effekt modelliert wird. Für einen T > C-Polymorphismus könnte z. B. C als Effektallel verwendet werden und die Genotypen T/T mit 0, T/C mit 1 und C/C mit 2 kodiert werden. Während a als Interzept der Regressionslinie den Mittelwert des Phänotyps in der Gruppe der Personen mit Genotyp 0 beschreibt, beschreibt b die Erhöhung des Mittelwerts von Y je Effektallel (d. h. je eine Einheit Erhöhung im Genotyp). Wenn also z. B. der mittlere BMI unter allen Personen mit Genotyp T/T (SNP = 0) eines T > C-Polymorphismus 23,0 kg/m^2 beträgt, und b für die-

sen SNP 0,5 kg/m^2 je Effektallel C betrüge, dann würden die Personen mit Genotyp C/T (SNP = 1) im Mittel einen BMI von 23,5 kg/m^2 und diejenigen mit Genotyp C/C (SNP = 2) einen BMI von 24,0 kg/m^2 aufweisen.

Für einen dichotomen Phänotyp D (z. B. Diabetes: ja/nein, AMD: ja/nein) wird ein *logistisches Regressionsmodell*

$$logit(Erkrankungswahrscheinlichkeit \mid SNP) = a + b\,SNP$$

verwendet, wobei logit für die Funktion $\text{logit}(p) = \ln(p/(1-p))$ steht. Durch die Kodierung des Genotyps für einen SNP durch 0, 1, 2 wird wieder ein additives genetisches Modell angesetzt. Während exp(a) der Erkrankungswahrscheinlichkeit unter den Personen mit Genotyp 0 (SNP = 0) entspricht, welche in Fall-Kontroll-Studien per Design künstlich ist und nicht interpretiert wird, beschreibt exp(b) die Odds Ratio, OR. Die OR ist näherungsweise dem relativen Risiko gleichzusetzen, also dem Faktor, um welchen die Erkrankungswahrscheinlichkeit von Personen je Effektallel (d. h. je Einheit Erhöhung im Genotyp) ansteigt oder fällt. Wenn sich z. B. für b = 1,030 eine OR von exp(1,030) = 2,8 für das Effektallel C eines T > C-Polymorphismus errechnet, dann bedeutet das, dass

- Personen mit C/T (SNP = 1) ein um (2,8–1,0)% = 180% erhöhtes Erkrankungsrisiko im Verhältnis zur Personen mit T/T (SNP = 0) besitzen und dass
- Personen mit C/C (SNP = 2) im Vergleich zu Personen mit T/T eine OR von $\exp(2 * 1{,}030) = OR^2 = 2{,}8^2 = 7{,}8$ und deshalb ein um (7,8–1,0)% = 680% erhöhtes Erkrankungsrisiko aufweisen (also ein über 6-fach erhöhtes Risiko).

Im einfachsten Fall kann die OR auch durch einen Vergleich der Effektallelfrequenzen (EAF) von Fällen und Kontrollen berechnet werden. Wenn z. B. die Häufigkeit des C-Allels bei Fällen 60% und bei Kontrollen 35% beträgt und man 500 Fälle und 1000 Kontrollen unter Beobachtung hat, dann könnte man die OR auch durch $(0{,}60 * 500/0{,}40 * 500)/(0{,}35 * 1000/0{,}65 * 1000) = (0{,}60 * 0{,}65)/(0{,}35 * 0{,}40) = 2{,}8$ errechnen.

Identifizierung von neuen Loci

Was ist ein „signifikanter SNP" und wie wird ein SNP als „neuer Lokus" identifiziert? Da die derzeit verwendeten genomweiten SNP-Panels meist 1 Mio. unabhängige Tests involvieren, hat sich ein *genomweites Signifikanzniveau* von 0,05/1 Mio. $= 5 * 10^{-8}$ durchgesetzt. Wenn also der zu b gehörende p-Wert $< 5 * 10^{-8}$ ist, spricht man von einem „genomweit signifikanten" SNP und „der Identifikation eines Genorts für den Phänotyp".

Die meisten Genorte werden derzeit durch *GWAS-Metaanalysen* im zweistufigen Verfahren identifiziert: Im ersten Schritt („discovery") werden genomweit für jeden einzelnen SNP die b_j je Studie j berechnet. Der Gesamteffekt b wird dann als Summe über alle $w_j\,b_j$ berechnet, wobei w_j gewichtet für die Größe der Studie j oder für das Inverse des Quadrats des Standardfehlers von b_j. Für die besten SNPs werden unabhängige Studien in einem zweiten Schritt („follow-up") erneut ausgewertet. Die genomweite Signifikanz wird im Allgemeinen basierend auf dem gemeinsamen p-Wert von „discovery" und „follow-up" ermittelt. Das „follow-up" wird manchmal verwirrenderweise als Replikation („replication") bezeichnet, obwohl es sich dabei normalerweise nicht um ein unabhängiges Experiment handelt. Für eine genomweit signifikante SNP-Assoziation bezeichnen wir diese gemeinsame Analyse von „discovery" und „follow-up" als „identifizierende Studie".

Quantifizierung des genetischen Effekts für einen SNP

Maße für den genetischen Effekt

Der *genetische Effekt* kann für einen SNP durch b bzw. bei dichotomonen Phänotypen durch exp(b) = OR angegeben werden: Um wie viel ist der BMI im Mittel je Effektallel erhöht oder erniedrigt? Um wie viel ist die Erkrankungswahrscheinlichkeit pro Effektallel erhöht oder erniedrigt?

Für beide, OR oder b, muss die *Effektrichtung* berücksichtigt werden: Oft wird das „Effektallel", also das Allel, in dessen Richtung der Effekt zeigt (in dem Beispiel C), als „Risikoallel" bezeichnet. Dann muss man sich jedoch vergegenwärtigen, dass dieses Risiko „advers" (C erhöht den mittleren BMI oder die Erkrankungswahrscheinlichkeit) oder „protektiv" (C erniedrigt den mittleren BMI oder die Erkrankungswahrscheinlichkeit) sein kann. Oft wird die Effektgröße in der Richtung des selteneren Allels („minor allele") angegeben: Für OR > 1 oder b > 0 geht das seltenere Allel dann mit einer erhöhten Erkrankungswahrscheinlichkeit bzw. höherem mittlerem BMI einher; im anderen Fall ist OR < 1 bzw. b < 0. Dies hat Nachteile, wenn die Häufigkeit des selteneren Alleles („minor allele frequency", MAF) des SNPs nahe 50% liegt, da sich dann die „selteneren" Allele der einzelnen Studien per Zufall unterscheiden können. In jüngerer Zeit hat es sich eingebürgert, den Effektschätzer in die Richtung des „schlechteren" Phänotyps anzugeben (also ein echtes „Risikoallel"); dann ist OR > 1 bzw. b > 0 für das Allel, welches mit erhöhter Erkrankungswahrscheinlichkeit oder mit erhöhtem BMI einhergeht. Dies hat den Vorteil, dass man kumulative Risikoscores (s. unten) leichter berechnen kann, aber den Nachteil, dass die „Risikoallele" für verschiedene Phänotypen unterschiedlich sein können: z. B. kann sich die Diabeteserkrankungswahrscheinlichkeit für das Allel C erhöhen (OR > 1 für das C Allel), während sich C als protektiv für AMD erweisen könnte, dann aber die OR > 1 für das T-Allel beschrieben wird.

Man kann das Ausmaß der „genetischen Wirkung" auch als den durch den SNP *erklärten Anteil der Varianz* im Phänotyp angeben. Wenn ein SNP 1% des BMI in der Allgemeinbevölkerung erklärt, bedeutet das, dass sich die Varianz des BMI in der Allgemeinbevölkerung bei „Verschwinden des SNPs" um 1% reduzieren würde. Eine 100% durch die Genetik erklärte Varianz im BMI würde bedeuten, dass man den BMI-Wert einer Person genau durch die Genetik bestimmen könnte. Dieser durch einen SNP erklärte Anteil der Varianz eines quantitativen Parameters wird als R^2 durch

$$2EAF(1-EAF)\frac{b^2}{Var(Y)}$$

berechnet, wobei Var(Y) die Varianz des Phänotyps (= Standardabweichung von Y

medgen 2011 · 23:377–384
DOI 10.1007/s11825-011-0295-7

I.M. Heid · T.W. Winkler · F. Grassmann · B.H.F. Weber

Wie groß sind die kleinen genetischen Risiken?

Zusammenfassung

Während Familienstudien sich als sehr geeignet zeigen, um starke genetische Risikovarianten aufzuspüren, sind genomweite Assoziationsstudien mit nichtverwandten Personen besonders effizient in der Identifizierung von moderaten und schwachen genetischen Risiken bei multifaktoriellen Erkrankungen und erkrankungsrelevanten quantitativen Parametern. Hier wird dargestellt, wie das genetische Risiko für solch moderat bis schwach wirkende Varianten berechnet wird. An den Beispielen Adipositas, Diabetes und altersbedingte Makuladegeneration wird gezeigt, welche Modelle Anwendung finden und wie groß diese „kleinen" genetischen Risiken sind.

Schlüsselwörter

Genetische Assoziationsstudien · Risikobestimmung · Lineare Modelle · Adipositas · Altersbedingte Makuladegeneration

How big are the small genetic risks?

Abstract

While family studies are ideal to pinpoint strong genetic risk effects, genome-wide association studies in unrelated individuals are particularly successful in identifying moderate and small genetic risks for multifactorial diseases and disease-relevant quantitative parameters. Here, we present how the genetic risk for such variants is computed and what models are used to derive cumulative genetic risk. Using the examples of obesity, diabetes, and age-related macular degeneration, we illustrate how these risks are computed and tackle the question of how big the small genetic risks are.

Keywords

Genetic association studies · Risk assessment · Linear models · Obesity · Age-related macular degeneration

Tab. 1 Durch einen SNP erklärte Varianz für einen quantitativen Phänotyp, R^2, wobei EAF die Effektallelefrequenz, b die Effektgröße und Var(Y) die Varianz des quantitativen Phänotyps beschreiben

EAF	2 EAF (1−EAF)	b	b^2/Var(Y)	Erklärte Varianz nach 2 EAF(1−EAF)b^2/Var(Y) (%)
0,5	0,5	0,5	0,015625	0,78
0,5	0,5	0,1	0,000625	0,03
0,25	0,375	0,5	0,015625	0,59
0,25	0,375	0,1	0,000625	0,02
0,1	0,18	0,5	0,015625	0,28
0,1	0,18	0,1	0,000625	0,01

Tab. 2 Das populationsattributable Risiko für einen SNP mit gegebener Effektallelfrequenz, EAF, und einem gegebenen relativen Risiko RR (das für Erkrankungen mit weniger als 10% Häufigkeit durch die OR angenähert werden kann)

Relatives Risiko, RR	Effektallelfrequenz, EAF	Populationsattributables Risiko, PAR (%)
1,1	0,5	4,8
1,1	0,1	1,0
1,5	0,5	20,0
1,5	0,1	4,8
2,0	0,5	33,3
2,0	0,1	9,1

im Quadrat) und 2 EAF (1–EAF) die Varianz der Genotypen beschreibt. Ein SNP mit genetischer Effektgröße b = 0,5 kg/m^2 und einer EAF von 50% in der Allgemeinbevölkerung mit BMI-Standardabweichung von 4 kg/m^2 würde 0,78% des BMI erklären (**Tab. 1**).

Ein häufiger SNP (z. B. EAF = 0,5) mit kleiner Effektgröße (z. B. b = 0,1 kg/m^2) kann eine vergleichbare „Wirkung" haben wie ein seltener SNP (z. B. EAF = 0,1) mit großer Effektgröße (z. B. b = 0,5). Das R^2 berücksichtigt also nicht nur die genetische Effektgröße, sondern auch die Häufigkeit der Variante in der Bevölkerung.

Analog für dichotome Phänotypen gibt es den Begriff des populationsattributablen Risikos (PAR), welches sich aus der genetische Effektgröße OR (in der Annahme, dass es das relative Risiko RR annähert) und der Häufigkeit des Effektallels („effect allele frequency", EAF) in der Allgemeinbevölkerung durch

$$\frac{RR-1}{RR-1+\frac{1}{EAF}}$$

berechnet. Würde ein SNP mit EAF von 10% also einen RR von 1,50 für eine Erkrankung wie T2D aufweisen, würde man ein PAR von 4,8% errechnen; für einen SNP mit EAF von 50% wäre das PAR 20% (**Tab. 2**).

Studiendesignaspekte

Im Idealfall wird für die Erhebung der genetischen Effektgröße eine *effektbestimmende Studie* (oder Metaanalyse von Studien) verwendet, welche unabhängig von der Studie sein sollte, welche die SNP-Assoziation identifizierte (*identifizierende Studie*). Bei der ersten identifizierenden GWAS tendieren Effektschätzer dazu, größer als der wirkliche Effekt zu sein („winner's curse"). Es entspricht auch „der guten epidemiologischen Praxis", die Hypothese in einem ersten Datensatz zu generieren (d. h. einen SNP als relevant zu identifizieren) und in einem unabhängigen zweiten Datensatz die Hypothese zu testen bzw. die Effekgröße zu berechnen. Nicht ideal ist es daher, die genetische Effektgröße in der identifizierenden GWAS zu ermitteln. Ein Problem kann entstehen, wenn in der effektbestimmenden Studie ein identifizierter SNP nicht signifikant ist, was verschiedene Gründe haben kann:

- falsch-positiver SNP durch „Pech" in der identifizierenden Studie,
- zu kleine Stichprobe oder „Pech" in der effektbestimmenden Studie oder

- Heterogenität des genetischen Effekts zwischen den Studien.

Im dritten Fall würden mehrere effektbestimmende Studien benötigt, um dieses Puzzle zu lösen; der erste Fall sollte durch rigides Design der identifizierenden Studie minimiert sein, der zweite Fall durch eine genügend große risikobestimmende Studie vermieden werden. Zum Beispiel wäre die Power für den BMI-SNP mit dem stärksten genetischen Effekt (rs1558902, *FTO*, b = 0,39, EAF = 0,42) 99% in einer Studie mit 5000 Personen (5%-Signifikanzniveau, Standardabweichung von 4 kg/m^2); für den schwächsten BMI-SNP (rs206936, *NUDT3*, b = 0,06, EAF = 0,20) wäre die Power aber dann nur 7% [1].

Bei dichotomen Phänotypen könnte man *prävalente Fallgruppen* (d. h. Erkrankte, die man innerhalb eines engen Zeitrahmens als krank identifiziert, z. B. alle Diabetesfälle, die sich bei einer Befragung im Jahr 2007 als Diabetiker bezeichnen) von *inzidenten Fallgruppen* (d. h. Erkrankte, die innerhalb eines Zeitraums neu erkranken und vorher bekanntermaßen gesund waren, z. B. alle AMD-Fälle, die sich in den Jahren 2006–2009 in der Ophthalmologie erstmals vorstellen) unterscheiden. Risikoberechnungen in der nichtgenetischen Epidemiologie sollten sich auf inzidente Fälle beziehen und die Exposition (z. B. Gewicht) sollte vor dem Fallerhebungszeitraum gemessen werden, da man bei gleichzeitiger Erhebung von Exposition und Erkrankung (wie in Querschnittstudien) nicht sagen kann, was zuerst war (z. B. ob das höhere Gewicht die Erkrankungswahrscheinlichkeit erhöhte oder ob die Erkrankung das Gewicht erhöhte). Bei genetischen Assoziationen ist dies weniger von Belang, da die Genetik – ohne Berücksichtigung von etwaigen epigenetischen Effekten – von der Zeugung an festgelegt ist. Bei genetischen Assoziationen können prävalente und inzidente Fälle nur mit unterschiedlichen Risiken einhergehen, wenn

- die Genotypen mit unterschiedlichem Schweregrad der Erkrankung assoziiert sind und schwerere Fälle eher nicht in die Studien eingeschlossen werden,
- Patienten mit dem einen Genotyp schneller versterben als die anderen (d. h. die Wahrscheinlichkeit, dass diese Patienten als prävalente Fälle erfasst werden, ist geringer als für Patienten die länger im Krankenstatus verweilen).

In beiden Fällen würde eine Risikoschätzung basierend auf inzidenten Fällen vorzuziehen sein.

Quantifizierung des kumulativen Risikos

SNP-Selektion zur Berechnung des kumulativen Risikos

Als kumulativer Effekt wird der Effekt bezeichnet, welcher nicht nur durch einen einzelnen SNP, sondern durch eine ganze Reihe von SNPs entsteht. Man kann den kumulativen Effekt einer Genregion oder aller für den Phänotyp identifizierter Genregionen berücksichtigen. Die SNPs müssen also irgendwie ausgewählt werden, aber wie?

Um den *kumulativen Effekt einer Genregion* zu beschreiben, stehen verschiedene SNP-Selektionsansätze zur Verfügung:

- Man kann den einen SNP, der in der GWAS den stärksten Effekt für die Region aufweist, als „Top-SNP" auswählen („top SNP selection"). Dies wird aber etwaige unabhängige zweite SNPs in der gleichen Genregion unberücksichtigt lassen und das Risiko unterschätzen.
- Alternativ kann man alle signifikanten und „unabhängigen" SNPs auswählen („independent SNP selection"). Solche unabhängigen SNPs können z. B. solche sein, welche eine paarweise Korrelation < 0,2 oder/und einen Abstand von >1 Mb zueinander aufweisen.
- Manchem erscheint es unangebracht, nichtsignifikante SNPs für den Risikoscore zu verwenden – aus statistischer Sicht spräche hier aber nichts dagegen: Also könnten auch alle zur Verfügung stehenden SNPs aus der Genregion verwendet werden („comprehensive SNP selection").
- Möglich ist auch die Auswahl von Tagging-SNPs, also z. B. SNPs, welche die gesamte Variation an Genetik in der Stichprobe am besten beschreiben („tagging SNP selection").
- Man könnte sich auch auf den Standpunkt stellen, dass nur die als funktionell nachgewiesene SNPs verwendet werden sollten; diese sind aber für die wenigsten Genorte derzeit bekannt.

Um den *kumulativen Effekt aller Genregionen* für einen bestimmten Phänotyp zu berechnen, kann man analog zu den obigen Möglichkeiten alle Top-SNPs, alle unabhängigen, alle SNPs oder alle Tagging-SNPs aus allen identifizierten Genregionen verwenden. Man kann sich aber auch auf den Standpunkt stellen, dass sich unter den SNPs mit p-Werten $<1*10^{-5}$ sehr viele echt assoziierte SNPs befinden. Daher ist es auch möglich, alle SNPs mit p-Werten $<1*10^{-5}$ für die Berechnung des kumulativen Risikos zu verwenden. Man könnte alle SNPs aus den GWAS nehmen, das wird allerdings aus Praktikabilitätsgründen nicht gemacht.

Berechnung des kumulativen Risikos

Nehmen wir also an, dass eine bestimmte Liste von k SNPs ausgewählt wurde, für welche wir den genetischen kumulativen Effekt berechnen wollen. Dann schätzen wir die Einzel-SNP-Effekte (z. B. b_1) adjustiert für alle anderen SNP-Effekte (z. B. b_2, …, b_k) durch das lineare Regressionsmodell für quantitative Phänotypen

$$\textit{Erwartungswert}\ [Y \mid SNP] = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Die genetischen Effekte b_1, …, b_k sollten dabei alle in die Richtung des „schlechteren" Phänotyps zeigen (d. h. alle $b_i > 0$). Damit lässt sich das R^2 *für das gesamte Modell* als Maß für die erklärte Varianz verwenden, wobei man das R^2 heranziehen sollte, das für die Anzahl an Parametern im Modell korrigiert ist. Dieses ist verfügbar in den gängigen Softwarepaketen. Für logistische Regression ist diese Modellierung prinzipiell auch möglich, die Interpretation des R^2 ist allerdings nicht so einfach und wird stark debatiert. Für dichotome Phänotypen kann alternativ die „area under the curve" (AUC) der

„ROC (receiver operating characteristic) curve“ ermittelt werden.

Man kann einen *genetischen Risikoscore* basierend auf den selektierten SNPs für jede Person berechnen: Für eine Person mit den Genotypen $x_1, x_2, \ldots, x_k$ für die k SNPs wird die gewichtete Summe

$$Score = w_1\ x_1 + w_2\ x_2 + \ldots + w_k\ x_k$$

berechnet. Für die Wahl der Gewichte gibt es verschiedene Ansätze:

1. Ungewichtet, also $w_1 = 1, \ldots, w_k = 1$ („unweighted risk score“): Hierbei werden einfach alle Risikoallele gezählt; der Score kann Werte von 0 bis 2k erreichen. Unberücksichtigt bleibt hierbei, dass die Effekte der einzelnen SNPs unterschiedlich sind. Im Beispiel des BMI hätte das Risikoallel im *FTO*-Gen eine schwerwiegendere Auswirkung als ein Risikoallel im *NUDT3*.
2. Effektgrößen als Gewichte, also $w_1 = b_1, \ldots, w_k = b_k$ („effect size weighted risk score“): Dabei werden die geschätzten Einzel-SNP-Effektgrößen eingesetzt und ein Risikoallel mit einem starken Effekt geht also entsprechend stärker ein. Der Risikoscore selbst ist dann etwas schwer interpretierbar, da eine Einheit Erhöhung im Risikoscore sowohl durch einen SNP mit Effektgröße 1 oder durch 10 SNPs mit Effektgröße 0,1 entstehen kann.
3. Relative Effektgrößen als Gewichte, also $w_1 = b_1/mean\_b, \ldots, w_k = b_k/mean\_b$ („relative effect size weighted risk score“): Hier stellt mean_b den Mittelwert aus allen $b_1, \ldots, b_k$ dar. Nun berücksichtigt der Risikoscore die unterschiedlichen Effektgrößen und der Score rangiert wieder zwischen 0 und 2k. Allerdings ist auch hier eine Einheit Erhöhung im Risikoscore schwer zu interpretieren; am ehesten kann man sie sich als eine Einheit Erhöhung eines der k SNPs mit einem mittelgroßen Effekt vorstellen.

Mithilfe des Risikoscores kann man graphisch die Verteilung des kumulativen Risikos in der Bevölkerung beschreiben (s. Abschnitt „Beispiele“). Außerdem kann man den Gesamteffekt durch logistische Regression

$$logit\,(Erkrankungswahrscheinlichkeit \mid SNP) = a + b\ Score$$

oder lineare Regression

$$Erwartungswert\ [Score \mid SNP] = a + b\ Score$$

auf Signifikanz testen und den Effekt pro Risikoscore-Einheit berechnen.

Modellierungsaspekte

Bei der Quantifizierung des kumulativen Risikos sind einige Modellierungsaspekte zu beachten:

1. Die SNPs im Modell umfassen nicht die „gesamte Genetik“: Die derzeitigen genomweiten SNP Panels sind auf häufige SNPs reduziert (MAF > 5%) und beinhalten noch relativ wenige seltenere Polymorphismen (MAF < 5%). Außerdem können komplexere genetische Strukturen bislang nur unzureichend abgebildet werden (z. B. CNVs). Wenn Gesamteffektberechnungen basierend auf diesen SNPs also zu 5% erklärter Varianz führen, dann kann der echte „durch die Genetik“ erklärte Anteil deutlich höher liegen. Tatsächlich beobachtet man, dass die derzeit durch SNPs erklärte Varianz deutlich geringer ist als die in Zwillingsstudien ermittelte Heritabilität, ein dem R^2 verwandtes Maß, was als „missing heritability“ diskutiert wird.
2. Der echte Effekt könnte größer sein, wenn Gen-Umwelt-Interaktionen existierten. Diese könnten durch geeignete Interaktionsmodelle berücksichtigt werden, was derzeit wenig Anwendung findet, da die Datenlage zu den Gen-Umwelt-Interaktionen noch unzureichend ist:

$$Score = (b_1 SNP) + (b_2 Umweltfaktor) + (b_3 SNP * Umweltfaktor)$$

3. In den bisher genannten Modellen ist unberücksichtigt, dass sich die SNPs gegenseitig beeinflussen können. Dies könnte durch

$$Score = (b_1 SNP_1) + (b_2 SNP_2) + (b_3 SNP_1 * SNP_2)$$

modelliert werden, was derzeit noch eher unüblich ist.

Beispiele

Body-Mass-Index

Ein erhöhter BMI ist von Interesse als Risikofaktor für schwere chronische Erkrankungen wie Diabetes, Herzinfarkt und Mortalität. Der genetische Anteil an Adipositas wird auf 40–70% durch Zwillingsstudien geschätzt. Wie groß sind die genetischen Effekte, die derzeit bekannt sind, und welchen Anteil des BMI erklären sie?

Das GIANT-Konsortium beschrieb jüngst 32 Genorte für den BMI aus der derzeit größten GWAS-Metaanalyse [1]. Die einzelnen SNPs erklären je 0,01–0,34% der Varianz des BMI; die einzelnen genetischen Effekte betragen 0,06–0,39 kg/m^2 pro Effektallel. Wenn man den Phänotyp dichotomisiert (Adipositas: ja/nein, BMI ≥ 30 kg/m^2 oder < 30 kg/m^2) erhält man ORs von 1,016–1,203. Alle 32 SNPs gemeinsam erklären 1,45% der Varianz im BMI in der Allgemeinbevölkerung.

Die Autoren hatten großes Interesse daran, einen genetischen Risikoscore zu berechnen. Deshalb wurde ein genetischer Risikoscore in einer der größten GIANT-Studien, der „Atherosclerosis-Risk-in-Communities(ARIC)-Studie“ (n = 8120) berechnet: Die Anzahl der BMI-erhöhenden Allele pro Person wurden für alle 32 identifizierten Top-SNPs (ein SNP je Genregion) aufsummiert und gewichtet mit dem relativen genetischen Effekt (d. h. gewichtet mit dem genetischen SNP-Effekt geteilt durch den Mittelwert der genetischen Effekte, „relative effect size weighted“). Für diese Gewichtung wurden die genetischen Einzel-SNP-Effekte aus dem „follow-up“ der GWAS-Metaanalyse verwendet, da die ARIC-Studie für die Berechnung der Einzel-SNP-Effekte wiederum zu klein war: Die Power für ein Modell mit Risikoscores ist deutlich besser als die Power von Einzel-SNP-Effekten, da die Risikoscores die Information aus vielen SNPs in einer einzelnen Variablen vereinen und diese Variable ein breites Wertespektrum annimmt. Die Autoren ermittelten, dass mit jeder Erhöhung des genetischen Risikoscores um eine

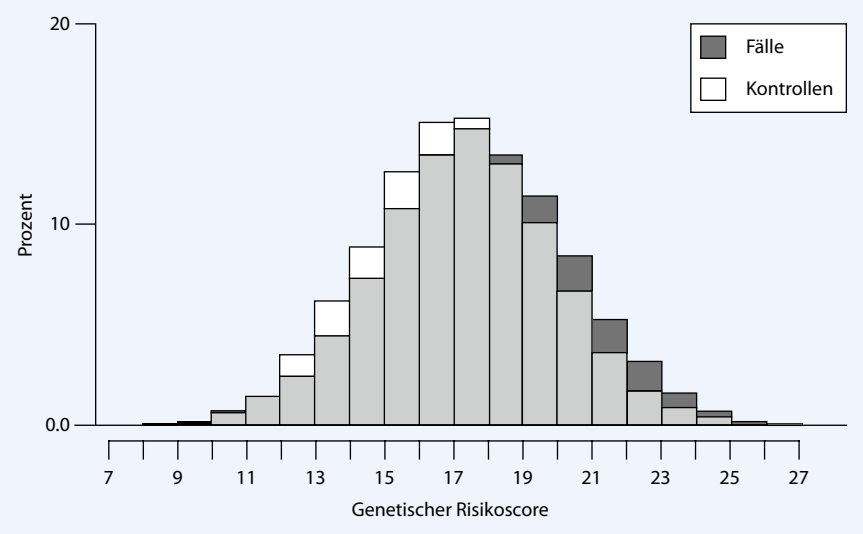

Abb. 1 ▲ Häufigkeitsverteilung eines genetischen Risikoscores (als ungewichtete Anzahl der Risikoallele), wie man es bei Personen mit und ohne Typ-2-Diabetes erwarten würde. (In Anlehnung an [2])

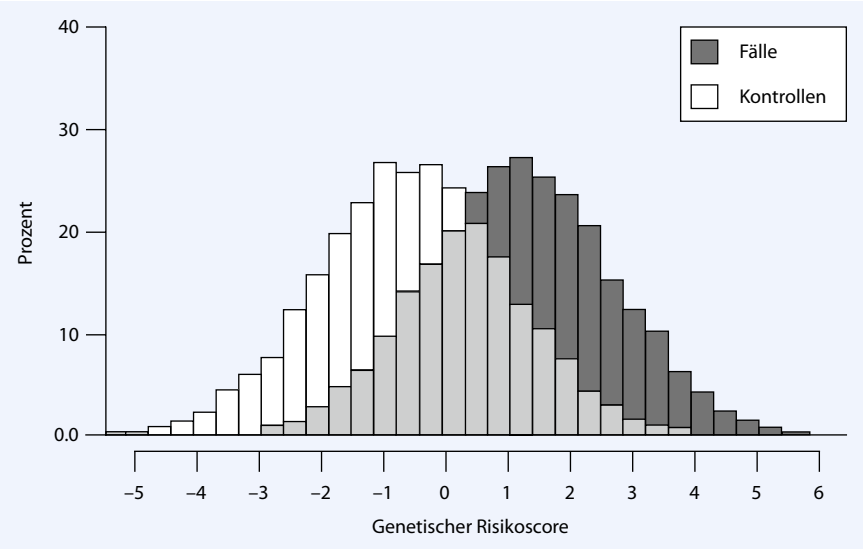

Abb. 2 ▲ Häufigkeitsverteilung eines genetischen Risikoscores (als Anzahl der Risikoallele gewichtet für die Effektgröße), wie man es bei Personen mit und ohne alterbedingte Makuladegeneration erwarten würde. (In Anlehnung an [6])

Einheit der BMI um 0,17 kg/m^2 ansteigt, was einer Erhöhung des Körpergewichts um 435–551 g je Risikoscore-Einheit bei 160–180 cm großen Erwachsenen entspräche. Der BMI von Personen mit hohem genetischen Risikoscore (≥38 Scorepunkte) ist dann im Mittel um 2,73 kg/m^2 höher als bei Personen mit geringem genetischem Risikoscore (<21 Scorepunkte), also 6,99–8,85 kg bei 160–180 cm großen Erwachsenen.

Die durch die 32 Genorte erklärte Varianz im BMI von 1,45% liegt deutlich unter der geschätzten Heritabilität von 40–70%. Die Gesamtheit aller SNPs, die in der Metaanalyse p-Werte kleiner als 0,05 aufgewiesen hatten, erklärten in einer von der Metaanalyse unabhängigen Studie 2,5% der BMI-Varianz. Es bleibt also nach wie vor ein großer Teil der Heritabilität unerklärt. Diese Diskrepanz zwischen der durch die identifizierten Genorte erklärten phänotypischen Varianz und der geschätzten Heritabilität zeigt sich für die meisten Phänotypen und wird derzeit sehr heftig debatiert. Unter den Erklärungen für diese „missing heritability" sind

- weitere Polymorphismen, die durch noch größere GWAS mit Assoziation für den Phänotyp identifiziert werden können,
- viele seltene Polymorphismen oder Mutationen, welche für den Phänotyp relevant sind und evtl. durch Sequenzierung detektiert werden können,
- andere genetische Konstrukte, welche durch die SNP-Panels nicht genügend beschrieben sind, z. B. „copy number variations",
- Gen-Gen- oder Gen-Umwelt-Interaktionen oder
- eine Überschätzung der Heritabilität.

Diabetes

Auch zu Typ-2-Diabetes (T2D) sind durch GWA-Metaanalysen mittlerweile zahlreiche Genorte identifiziert worden. Man geht davon aus, dass ein genetischer Risikoscore gezielte Vorhersagen hinsichtlich des Auftretens von Erkrankungen in der Allgemeinbevölkerung erlaubt. Meigs et al. haben dies untersucht [2]. Sie genotypisierten SNPs in 18 Genorten für T2D bei 2377 Probanden der Framingham Offspring Study, berechneten einen ungewichteten Risikoscore durch reines Addieren der Allele (Spannweite der Werte: 0–36), und erhielten eine OR von 1,12 je Erhöhung um eine Einheit im Risikoscore (95%-Konfidenzintervall: 1,07–1,17), was einer Erhöhung der Diabeteswahrscheinlichkeit um 12% je Risikoscore-Einheit entspricht. Ist das viel? Erlaubt also der Risikoscore eine Vorhersage, ob eine Person im Lauf ihres Lebens Diabetes entwickeln wird oder nicht?

Die Autoren Meigs et al. beschrieben weiter, dass unter den Personen mit <15, 16–20, oder >21 Risikoscore-Punkten sich 7%, 11%, und 17% mit Diabetes befänden. Das klingt nach einer deutlichen Erhöhung. Dabei ist aber die Häufigkeit, mit welcher diese Risikoscores auftauchen, noch nicht berücksichtigt: Da die Risikoscoregruppen je mit 25%, 64%, und 11% auftreten, erhält man ein populationsattributables Risiko von $0{,}64 * (0{,}11–0{,}07) + 0{,}11 * (0{,}17–0{,}07) \sim 0{,}04$.

Also sind 4% der Diabetesfälle den Genotypen zuzuschreiben; d. h. 4% können aufgrund ihres Genotyp-Scores eindeutig klassifiziert werden. Ist dies genug diskriminierende Kraft, um einen Patienten da-

Tab. 3 Einige Beispiele von AMD-Effektgrößen als Odds Ratio (OR) und populationsattributables Risiko (PAR). (Basierend auf [6])

Top-SNP	Genregion	Effektallelhäufigkeit nach HapMap	OR	PAR (%)
rs10490924	*HTRA1*	0,19	2,80	25,64
rs1061170	*CFH*	0,38	2,21	31,68
rs2230199	*C3*	0,21	1,43	8,45
rs10468017	*LIPC*	0,73	1,19	12,23

hingehend aufzuklären? Ein Test, der 4% Sensitivität aufweist?

Das Versagen des Risikoscores für Diabetes, eine Trennung der zukünftigen Patienten von den Gesundbleibenden vorzunehmen, zeigt auch das Folgende: Der Risikoscore lag bei Personen, die Diabetes entwickelten, im Mittel bei 17,7 (Standardabweichung: ±2,7) und bei Personen, die keinen Diabetes entwickelten, bei 17,1 (±2,6; $p < 0{,}001$). Die **Abb. 1** zeigt diese Verteilung, wobei deutlich wird, dass eine Trennung von Patienten und Gesunden durch den Risikoscore nicht möglich ist. Ohne Zweifel hat die Aufklärung eines Teils der genetischen Basis des T2D einen großen Beitrag zum Verständnis der Mechanismen geleistet. Die Auswirkungen auf das Auffinden neuer Biomarker oder die Entwicklung innovativer Therapiemöglichkeiten wird die Zukunft zeigen. Eine sinnvolle Risikovorhersage basierend auf der jetzt bekannten Genetik erscheint jedoch sehr zweifelhaft.

Altersabhängige Makuladegeneration

In einem der ersten publizierten GWAS überhaupt wurde ein genomweit signifikantes Assoziationssignal im Komplementfaktor-H(*CFH*)-Gen bei 96 AMD-Fällen und 50 Kontrollen beschrieben [3]. Diese Region war zuvor grob bereits durch Kopplungsanalysen beschrieben worden, konnte damals aber nicht näher eingegrenzt werden. In der identifizierenden Studie von Klein et al. weist die Variante rs1061170 (*CFH*: Y402H) einen OR von 4,6 (95%-Konfidenzintervall: 2,0–11,0) pro Effektallel auf. In nachfolgenden Studien wurde eine OR von 2,8 für diese Variante beschrieben, was auf einen „winner's curse", also auf einen durch Zufall überhöhten Effekt in der identifizierenden Studie, deutet. Diese Variante findet sich mit einer EAF von 60% in Fällen und 35% in Kontrollen. Nachfolgend wurde dann ein zweiter Genort, *ARMS2/HTRA1*, mit einer OR von 2,8 pro Effektallel beschrieben [4]. Eine Gen-Gen-Interaktion zwischen diesen beiden Genorten wurde bisher nicht gefunden. Schätzungen gehen davon aus, dass diese beiden Genorte für bis zu 50% der Krankheitsfälle verantwortlich sind. Solch starke genetische Effekte verleihen der AMD eine Sonderstellung bei den komplexen Erkrankungen.

Seddon et al. berechneten einen nach der Effektgröße gewichteten Risikoscore [5]. Auch wenn hier der Risikoscore anders gebildet wurde („effect size weighted") als der Risikoscore für Diabetes in dem obigen Beispiel („unweighted"), zeigt die Verteilung der Werte des Risikoscores deutlich die starke Aufteilung in AMD-Fälle einerseits und AMD-Kontrollen andererseits (**Abb. 2**), während für Diabetes fast keine Trennung zu erkennen war. Zu bemerken ist allerdings auch, dass für den mittleren Bereich des Risikoscores eine Unterscheidung zwischen AMD-Fällen und AMD-Kontrollen nicht möglich ist.

Inzwischen konnten 9 Genorte für AMD identifiziert werden, die ORs zwischen 1,1 und 2,9 aufweisen [6]. Das daraus mithilfe der Effektallelhäufigkeit in der Allgemeinbevölkerung berechnete populationsattributable Risiko ist für einige Beispiele in **Tab. 3** gegeben.

Ausblick

Die genetischen Effekte für einzelne mit komplexen Erkrankungen assoziierte SNPs sind generell niedrig mit ORs je Effektallel selten über 2,0 und eher im Bereiche 1,05–1,50. Aber das Beispiel der altersabhängigen Makuladegeneration zeigt, dass es auch Gene mit starken Effekten gibt, die durch Assoziationsstudien mit nichtverwandten Personen identifiziert wurden und von den zuvor durchgeführten Familienstudien nicht genau zu bestimmen waren. Insgesamt wurden sowohl beim BMI, bei Diabetes als auch für die altersabhängige Makuladegeneration die stärksten genetischen Assoziationen durch GWAS aufgedeckt (*FTO*, *TCF7L2* bzw. *CFH*).

Die Genetik des BMI zeigt auch, dass sich Genorte der seltenen starken Effekte (wie das *MC4R*-Gen, das durch Mutationen in Familienstudien mit extremer Adipositas gefunden wurde), mit den Genorten der schwachen häufigen Effekte aus GWAS-Signalen überlappen [7]. Bei wie vielen der häufigen schwachen GWA-Signale wird man zugrunde liegende seltene Polymorphismen finden? Werden die derzeit anlaufenden Sequenzierungen hier weiterhelfen?

Das Berechnen von Risikoscores und des attributablen Risikos (für dichotome Phänotypen) oder des Anteils der erklärten Varianz (R^2, für quantitative Phänotypen) bietet eine gute Möglichkeit, den Gesamteffekt von mehreren SNPs bzw. mehreren Genorten zu ermitteln. Inwieweit solche Risikoscores dazu dienen können, Voraussagen über Erkrankungswahrscheinlichkeiten für einzelne Personen sinnvoll treffen zu können, muss kritisch hinterfragt werden. Die meisten genetischen Effekte für komplexe Erkrankungen wie Diabetes sind zu klein und erklären insgesamt zu wenig, sodass Vorhersagen durch diese Varianten derzeit als nicht sinnvoll anzusehen sind. Aber es spielt nicht nur die Stärke der Voraussage eine Rolle, sondern auch die Therapiemöglichkeiten. Hat eine Person einen Nutzen davon, im Alter von 20 Jahren zu erfahren, dass sie mit einer Wahrscheinlichkeit von 50% im Alter von 80 Jahren an AMD erkranken wird? Wird diese Person durch das Wissen in den Nutzen von frühen Therapiemöglichkeiten kommen (falls diese in den nächsten 60 Jahren entwickelt werden sollten) oder wird die Person durch Sorge über diese Prognose über Gebühr belastet sein? Würde die Kenntnis von BMI-Risikovarianten eine Person zu einem besseren Lebensstil hinsichtlich Ernährung und Bewegung lenken, oder würde ein Ohnmachtsgefühl wenig oder

sogar gegenteilige Reaktionen hervorrufen?

Während diese Genvarianten hinsichtlich individueller Vorhersage für die meisten komplexen Erkrankungen eher wenig klinische Relevanz aufweisen, ist zu betonen, dass kleine genetische Risiken durchaus das Potenzial für hohe klinische Relevanz haben, indem sie auf potente Mechanismen für Medikamente weisen können. Als Beispiel sei das *HMGCR* erwähnt, das erst in einer der jüngsten GWAS-Metaanalysen mit über 100.000 Personen eine eher kleine Assoziation mit Gesamtcholesterin und LDL-Cholesterin zeigte [8], obwohl das Genprodukt, die Hydroxy-3-Methylglutargyl-Coenzym-A-Reduktase, den Mechanismus der Statine beschreibt, die als Medikament zur Senkung von Cholesterin von herausragender klinischer Bedeutung sind.

Wichtig wird zunächst sein, die genetische Basis von komplexen Erkrankungen noch besser zu verstehen. Vor allem wird interessant sein, ob weniger häufige Polymorphismen (1000-G-Imputationen) oder viele seltene Mutationen (Resequenzierung), Gen-Gen- oder Gen-Umwelt-Interaktionen noch mehr von der postulierten Heritabilität erklären. Dazu benötigt man v. a. große Studien, in welchen – unabhängig von den identifizierenden Studien – die Effektgrößen berechnet werden können.

Korrespondenzadresse

Prof. Dr. I.M. Heid
Public Health und Gender Studies
Institut für Epidemiologie und Präventivmedizin
Universitätsklinikum Regensburg
Josef-Strauss-Allee 11
93053 Regensburg

Interessenkonflikt. Der korrespondierende Autor gibt an, dass kein Interessenkonflikt besteht.

Literatur

1. Speliotes EK et al (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42(11):949–960
2. Meigs JB et al (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359(21):2208–2219
3. Klein RJ et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308(5720):385–389
4. Rivera A et al (2005) Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. Hum Mol Genet 14(21):3227–3236
5. Seddon JM et al (2009) Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. Invest Ophthalmol Vis Sci 50(5):2044–2053
6. Yu Y et al (2011) Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration. Hum Mol Genet (Epub ahead of print)
7. Loos RJ et al (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40(6):768–775
8. Teslovich TM et al (2010) Biological, clinical, and population relevance of 95 loci for blood lipids. Nature 466(7307):707–713

Genuntersuchungen zu erblich bedingtem Bluthochdruck veröffentlicht

Das International Consortium for Blood Pressure Genome-wide Association Studies (ICBP-GWAS) hat unter Beteiligung von 200 Zentren aus 24 Ländern die Erbinformationen von über 200.000 Probanden analysiert, um den genetischen Auslösern des Bluthochdrucks auf die Spur zu kommen. Unter den Probanden waren auch 4.000 Menschen aus Vorpommern, die sich für die Bevölkerungsstudie SHIP (Study of Health in Pomerania) zur Verfügung gestellt hatten.
Den Forschern gelang es, 28 Genloci zu identifizieren, die an der Regulation von systolischem und diastolischem Druck beteiligt sind. Bei 16 dieser Regionen war der Zusammenhang mit dem Blutdruck bisher noch nicht bekannt.
In einer weiteren Forschungsarbeit wurden die Blutdruckamplitude und der mittlere arterielle Blutdruck untersucht. Hier konnten die Forscher 4 neue Loci identifizieren, deren Gene sich auf die Blutdruckamplitude auswirken. Weitere 2 neu-identifizierte Genloci werden mit dem mittleren Blutdruck in Zusammenhang gebracht.
Wie die neu-identifizierten Regionen an der Regulation des Blutdrucks beteiligt sind, wird Gegenstand von Folgestudien sein.

Literatur:
The International Consortium for Blood Pressure Genome-Wide Association Studies (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature; doi:10.1038/nature10405

Wain LV, Verwoert GC, O'Reilly PF et al (2011) Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nature Genetics; doi:10.1038/ng.922

Quelle: Universitätsmedizin Greifswald, www.medizin.uni-greifswald.de