Von der Trennschärfe einer statistischen Entscheidung und dem Testen auf Gleichheit

Entscheidungen im zahnmedizinischen Alltag haben in der Regel langfristig Konsequenzen. Dabei sollte die Wahl neben den persönlichen Ansprüchen und Wünschen der Patienten sowie der individuellen klinischen Erfahrung des Zahnarztes (interne Evidenz) stets auch die neuesten Erkenntnisse der Forschung berücksichtigen (externe Evidenz) [3]. In diesem Artikel werden wir folgende, aus praktischer Sicht wichtige Entscheidungsproblematik beschreiben:

Es gibt eine klinisch wirksame Behandlungsmethode. Diese bereitet jedoch kleinere Unannehmlichkeiten, etwa ästhetischer Natur. Aus diesem Grund wurde eine verbesserte Methode entwickelt, die auf ihre klinische Wirksamkeit getestet werden soll. Diese Problematik hatten wir bereits am Ende des EbM-Splitters „Über die Signifikanz eines statistischen Tests und die zugehörigen Fehlentscheidungen“ [1] angesprochen, am Beispiel der fraglichen Klebeeigenschaften eines farblich verbesserten Zements. Eine weitere, zusätzliche Komponente ist die Zeit, die eine Patientin im Behandlungsstuhl verbringt, oder die der Zahntechniker benötigt, um eine Krone anzufertigen. Wie so oft in unserer schnelllebigen Zeit gilt das Prinzip: Je schneller, desto besser. Gerade deshalb sollte genau geprüft werden, ob gut Ding nicht doch lieber Weile haben will.

Wir betrachten hier die Problemstellung einer authentischen In-vitro-Studie zur Aufpassung von maschinell vorgefertigten Titan-Kappen [5]. Ein weiteres Beispiel für den Zwiespalt zwischen Zeit und Passgenauigkeit wird bei *Shodadai* et al. [2] beschrieben.

Problemstellung

Titan-Kappen für die Verblendtechnik sind geeignet, um Kronen im Mund zu befestigen; sie werden nach einer Abformungdruck im Patientenmund maschinell, mittels speziell entwickelter Verfahren, hergestellt. Vor dem Eingliedern werden die Kappen von einer Zahntechnikerin manuell nachgearbeitet; diesen Prozess nennt man Feinaufpassung. Die Passgenauigkeit der Kappen wird in Form der Ausdehnung des Randspalts zwischen Kappe und Zahnstumpf in µm gemessen; sie symbolisiert die klinische Wirksamkeit. Zudem wird die für die Feinaufpassung notwendige Arbeitszeit ermittelt.

Für unsere Zwecke genügt es, zwei Herstellungsverfahren zu betrachten: Verfahren 1 und Verfahren 2. Nehmen wir an, die statistische Analyse der Aufpassungszeit ergibt, dass mit Verfahren 1 gefertigte Kappen deutlich (statistisch signifikant) schneller aufgepasst werden können. Wäre also der erzielte Randspalt der beiden Verfahren gleich, dann würden wir aufgrund der Zeitersparnis Verfahren 1 empfehlen. Die zuletzt genannte Bedingung bedeutet genau Folgendes: Verfahren 2 ist ein klinisch wirksames Verfahren (z.B. ein Goldstandard), in dem Sinne, dass der erzielte Randspalt hinreichend klein ist. Die Grenze für „hinreichend klein” kann in diesem Zusammenhang bei ungefähr 120 µm angenommen werden. Nun ist Verfahren 1 genau dann noch empfehlenswert, wenn der erzielte Randspalt nur um eine klinisch bedeutungslose Größenordnung größer ist als bei Verfahren 2.

Tabelle 1 enthält hypothetische Daten von Randspalten für jeweils 20 Kappen.

Kappe	Verfahren 1	Verfahren 2
1	100	101
2	83	133
3	125	103
4	93	75
5	112	108
6	97	99
7	114	67
8	94	55
9	82	75
10	111	81
11	67	96
12	98	113
13	62	88
14	52	85
15	80	91
16	85	84
17	117	89
18	91	91
19	87	79
20	103	79

Tabelle 1 Hypothetische Daten von Randspalten (in µm) für 20 Titankappen.

In Anlehnung an die oben erwähnte In-vitro-Studie [5] gehen wir von paarigen Beobachtungen aus: Zu jedem von 20 verschiedenen Abdrücken werden jeweils zwei Kappen mit wechselndem Verfahren hergestellt. Der arithmetrische Mittelwert der paarweisen Differenzen liegt bei 3,05 µm, das zugehörige 95 %-Konfidenzintevall reicht von –8,8 µm bis 14,9 µm.

Im Weiteren werden die möglichen statistischen Hypothesen und die zugehörigen Testergebnisse anhand der Beispieldaten diskutiert. Wir gehen von einer Normalverteilung der Randspalten mit unbekannten Mittelwerten (M1 für Verfahren 1; M2 für Verfahren 2) und gleicher Varianz aus. Unter dieser Annahme ist der paarige t-Test die Methode der Wahl, auf die wir uns auch beschränken werden. Des Weiteren setzen wir den klinisch relevanten Unterschied (Toleranz) bei D=10 µm fest. Dies bedeutet z. B., dass eine mittlere Abweichung von 10 µm zwischen den beiden Verfahren als gleichwertig angesehen werden kann.

Möglichkeit 1: Einseitiger Test (Abb. 1)

Hypothese: $M1 - M2 < D$

Alternative: $M1 - M2 > D$

Die Alternative bedeutet, dass der erzielte Randspalt von Verfahren 1 mindestens genauso klein ist wie der von Verfahren

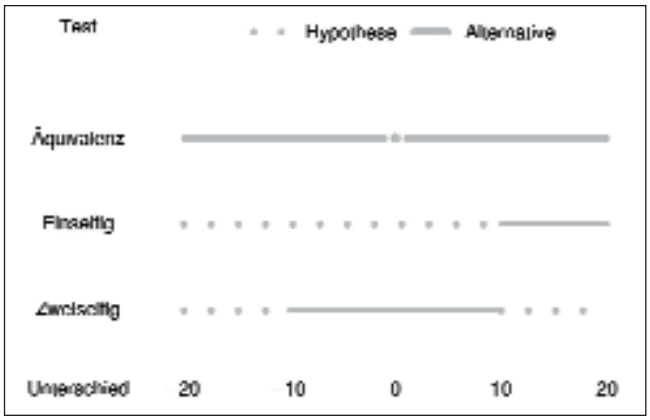

Abbildung 1 Verhältnis von Hypothese und Alternative in Abhängigkeit von Äquivalenztest, einseitiger Test und zweiseitiger Test

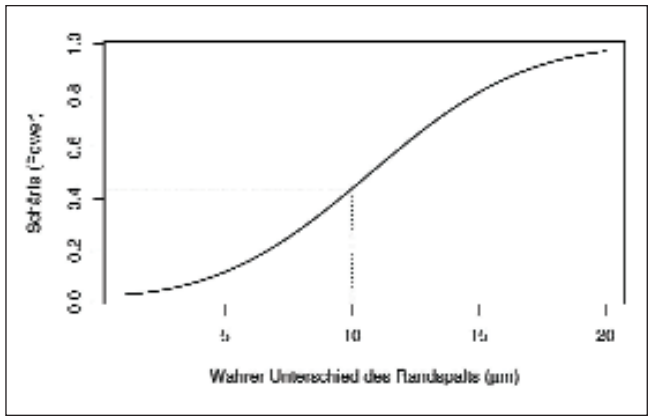

Abbildung 2 Trenn-Schärfe (Power) als Funktion des wahren Unterschieds des Randspalts

2 bei D=10 µm Toleranz. Der P-Wert des einseitigen t-Tests für D=10 µm ist 0,12. Zu dem üblichen Signifikanzniveau alpha=0,05 bedeutet dieses Ergebnis also Entscheidung für die Hypothese: Es kann nicht geschlossen werden, dass Verfahren 1 mindestens genauso gut ist wie Verfahren 2 (bei einer Toleranz von 10 µm).

Möglichkeit 2: Äquivalenztest (Abb. 1)

Hypothese: $|M1 - M2| > D$

Alternative: $|M1 - M2| < D$

Im Unterschied zum einseitigen Test bedeutet die Alternative beim Äquivalenztest Gleichheit der beiden Gruppen. Übersetzt in unser Beispiel: Der Unterschied der Randspalten nach Verfahren 1 und Verfahren 2 beträgt höchstens 10 µm. Die Testentscheidung des Äquivalenztests für D=10 µm kann unter Berücksichtigung eines 90%-Konfidenzintervalls wie folgt hergeleitet werden [für Details siehe z. B. 4]:

Der Äquivalenztest zum Niveau alpha=0,05 entscheidet genau dann für die Alternative, wenn das 90 %-Konfidenzintervall in dem Äquivalenzbereich von (–D bis D) enthalten ist. In unseren Daten ist die untere Grenze des 90%-Konfidenzintervalls –6,7 µm und die obere Grenze 12,8 µm. Also entscheidet der Äquivalenztest für die Hypothese: die beiden Verfahren sind nicht statistisch signifikant gleich.

Möglichkeit 3: Trennschärfe des zweiseitigen Tests auf Unterschied (Abb. 1)

Es gibt noch eine dritte reizvolle und gebräuchliche Variante für unsere Problemstellung, nämlich eine Analyse der Trennschärfe, auch Güte oder *Power* genannt, eines zweiseitigen Tests auf Unterschied:

Hypothese: $M1 - M2 = 0$

Alternative: $|M1 - M2| > 0$

Der P-Wert des paarigen t-Tests ist 0,6; es besteht also kein signifikanter Unterschied zwischen den beiden Verfahren. Daraus kann jedoch nicht direkt auf keinen Unterschied (also Gleichheit) der Randspalten geschlossen werden, wie wir unter Möglichkeiten 1 und 2 gesehen haben.

Die Trennschärfe eines Tests auf Unterschied hängt von dem klinisch relevanten Unterschied D ab. In Abbildung 2 haben wir die Trenn-Schärfe als Funktion von D aufgetragen.

Dazu sei bemerkt, dass die Trennschärfe entscheidend vom Stichprobenumfang abhängt. Es gilt: Je kleiner der Stichprobenumfang, desto weniger scharf ist der Test und desto ungenauer ist die Entscheidungsfähigkeit zugunsten der Hypothese.

„Man glaubt an Sachen, die man nicht beweisen kann. Deswegen glaubt man an Gott. Aber Zahnmedizin ist keine Religion."

Sandro Palla (Zürich), am 31. Mai 2005 in Basel auf der interdisziplinären Fortbildungswoche der Schweizerischen Zahnärzte-Gesellschaft (SSO) „Evidence Based Dentistry. Möglichkeiten und Grenzen".

Schlussbemerkung

Es ist ein weit verbreiteter Irrtum zu glauben, ein nicht-signifikantes Ergebnis sei auch ein Ergebnis. Wie wir in diesem EbM-Splitter erläutert haben, müssen spezielle statistische Methoden, Äquivalenztest oder Analyse der Trennschärfe (*Power*) des Tests, betrachtet werden, um auf Gleichheit von zwei Gruppen schließen zu können. In unserem Beispiel waren die Gruppen weder signifikant verschieden noch signifikant gleich. In gewisser Weise ist ein Test auf Unterschied vergleichbar mit einem Fernglas, bei dem die Sehschärfe eingestellt werden kann. Beim Testen ist das Einstellrädchen der Stichprobenumfang; wenn dieser sehr klein ist, ist das Bild, das uns der Test anhand der Daten über den interessierenden Sachverhalt liefert, verschwommen: wir können dann nicht einmal sehr große Unterschiede statistisch nachweisen. Andererseits, wenn der Stichprobenumfang sehr groß ist, sehen wir so scharf, dass auch klinisch bedeutungslose Unterschiede erkennbar werden.

Literatur

1. Gerds T, Türp JC, Antes G: Über die Signifikanz eines statistischen Tests und die zugehörigen Fehlentscheidungen. Dtsch Zahnärztl Z 60, 549-550 (2005)
2. Shodadai SP, Türp JC, Gerds T, Strub JR: Is there a benefit of using an arbitrary face-bow for the fabrication of a stabilization appliance? Int J Prosthod 14, 517-522 (2001)
3. Türp JC, Antes G: Was versteht man unter „Evidenzbasierter Medizin"? Dtsch Zahnärztl Z 56, 74 (2001)
4. Wellek S: Statistische Methoden zum Nachweis von Äquivalenz. Fischer, Stuttgart 1994
5. Witkowski S, Komine F, Gerds T, Strub JR: Marginal accuracy of titanium copings fabricated by casting and CAD/CAM techniques. J Prosthet Dent (2005) [eingereicht]

Thomas Gerds, Freiburg
Jens C. Türp, Basel
Gerd Antes, Freiburg