

**FORSCHUNGSBERICHTE DER
ABTEILUNG MEDIZINISCHE BIOMETRIE,
UNIVERSITÄT HEIDELBERG**



Nr. 41

**Die Bewertung von Tumor-Nachsorgemaßnahmen aus
biometrischer Sicht**

Juli 2002

**INSTITUT FÜR MEDIZINISCHE BIOMETRIE
UND INFORMATIK**

RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Forschungsberichte der
Abteilung Medizinische Biometrie, Universität Heidelberg

Nr. 41

**Die Bewertung von Tumornachsorgemaßnahmen aus
biometrischer Sicht**

PROF. DR. DR. ULRICH ABEL

DR. CHRISTINE WOLLERMANN

Institut für Medizinische Biometrie und Informatik (IMBI) der Medizinischen Fakultät der
Universität Heidelberg

Heidelberg, Juli 2002

Impressum:

Reihentitel: Forschungsberichte der Abteilung Medizinische Biometrie,
Universität Heidelberg

Herausgeber: Prof. Dr. Norbert Victor

Anschrift: Im Neuenheimer Feld 305, 69120 Heidelberg

Druck: Hausdruckerei der Ruprecht-Karls-Universität Heidelberg

elektronischer Bezug: <http://www.biometrie.uni-heidelberg.de>

ISSN: 1619-5833

Inhaltsverzeichnis

1. Einleitung	2
2. Das Problem der Validierung	6
2.1 Vorbemerkung	6
2.2 Wahrer Zustand und Außenkriterium	7
2.3 Test-Positivität	12
2.4 Sensitivität, Spezifität und Vorhersagewerte	13
3. „Effizienz“ und Nutzen von Nachsorgeprogrammen – methodische Aspekte	19
3.1 Notwendige Bedingungen für die Effizienz	19
3.2 Kritik der Standardargumentation für das diagnostische Monitoring	21
3.3 Zum Design methodisch fundierter Studien	28
4. „Effizienz“ und Nutzen von Nachsorgeprogrammen – randomisierte Studien	32
4.1 Mammakarzinom	32
4.2 Kolorektales Karzinom	38
5. Zusammenfassung	51
6. Literatur	53

1. Einleitung

Die Tumornachsorge ist ein Instrument der Patientenbetreuung, mit dem gleichzeitig mehrere Ziele verfolgt werden (Gallmeier u. Keding, 1994; Reitzig, 1995; Schumacher et al 1995; Northover, 1998, Delbrück u. Bamberg, 2000, Sakorafas et al., 2000, Taylor, 2000), zum Beispiel

- die Fahndung nach Krankheits"rückfällen" (metachrone Metastasen oder Lokalrezidive)
- die frühzeitige Entdeckung neuer Primärtumoren
- die psychosoziale Unterstützung der Patienten
- die Kontrolle der Behandlungsergebnisse
- spezielle Hilfe bei Defektheilungen
- die Erfassung und Behandlung von Langzeitfolgen der Krankheit oder Therapie
- die klinische Forschung.

Ergänzt wird diese Liste zuweilen durch den Gesichtspunkt „Ermöglichung einer externen Kontrolle („audit“) der Ergebnisse des behandelnden Chirurgen“ (Mc Ardle, 2000; Northover, 2000), doch handelt es sich hierbei wohl hauptsächlich um einen nützlichen Begleitaspekt, der für sich genommen schwerlich zur Begründung oder Rechtfertigung der Nachsorge dienen kann.

Das wohl herausragende Einzelziel ist das zuerst genannte: die möglichst frühzeitige Entdeckung von Lokalrezidiven oder Metastasen. Sie erfolgt mittels einer diagnostischen Überwachung im Rahmen regelmäßiger Kontrolluntersuchungen. Die Diagnostik, die dabei betrieben wird, kann recht aufwendig sein und z.B. im Falle des Brustkrebses eine Röntgenuntersuchung des Brustkorbes, eine szintigraphische Untersuchung der Knochen, einen Ultraschall des Bauchraums sowie die Bestimmung von Tumormarkern umfassen. Offenbar kommt es bei diesen Kontrollen darauf an, Krankheitsrückfälle zu entdecken, bevor sie ohne die regelmäßige Diagnostik erkannt worden wären, also bevor sie symptomatisch werden (Schumacher, 1995; Cocconi, 2001).

Die diagnostische Überwachung der Patienten ist heute bei vielen Krebsformen ein fester Bestandteil der Krebsnachsorge. Ihre spezielle Ausgestaltung - und damit insbesondere ihre Intensität - ist äußerst heterogen, sie hängt nicht nur von der Tumorentität, sondern auch vom betreuenden Arzt bzw. Zentrum ab. Cooper et al. (1999) fanden bei einer Untersuchung anhand der Daten von Medicare, daß es bei den Versicherten beträchtliche Variationen in den Follow-up-Testungen des kolorektalen Karzinoms gibt, und zwar mit sehr auffälligen systematischen Unterschieden zwischen geographischen Regionen. Dies verwundert angesichts der zahlreichen Optionen und Möglichkeiten nicht. So haben Kievit und Bruinvels (1995) angemerkt, daß sich aus acht unterschiedlichen diagnostischen Maßnahmen, die zu fünf Nachsorgeterminen einzeln oder in Kombination eingesetzt werden können, über 30000 Nachsorgestrategien zusammenstellen lassen, von denen eine große Zahl durchaus realistisch ist.

Unterschiede in der diagnostischen Überwachung ergeben sich natürlich auch bei Vergleichen verschiedener Zeiträume. Sie hängen zusammen mit der Entwicklung neuer diagnostischer Techniken, aber auch mit dem Grad der Umsetzung von wissenschaftlichen Erkenntnissen über die Validität und den Nutzen solcher Maßnahmen (s. z.B. die vergleichende Studie von Ciatto et al., 1998) oder von publizierten Empfehlungen (z.B. Mille et al., 2000).

Fundierte zusammenfassende Empfehlungen für Follow-up-Schemata nach potentiell kurativer Primärtherapie finden sich bei Edelman et al. (1997), und zwar für folgende Krebsformen: Keimzelltumor (ohne Seminom), Non-Hodgkin-Lymphom, M. Hodgkin, Mammakarzinom, Kolonkarzinom, Prostatakarzinom, Lungenkrebs. Für die diagnostischen Maßnahmen ist jeweils der Grad der aus klinischen Studien gewonnenen Evidenz angegeben, die ihre jeweilige Verwendung stützt. Darüber hinaus werden Schätzungen der hierdurch verursachten Kosten vorgenommen.

Selbstverständlich haben auch onkologische Gesellschaften Leitlinien für die Nachsorge herausgegeben, die allerdings ebenfalls nicht einheitlich sind. Für das kolorektale Karzinom sind bei Anthony et al. (2000) verschiedene Empfehlungen nordamerikanischer Gesellschaften in knapper Form tabellarisch zusammengestellt, eine weitere wurde unlängst von der American Society for Gastrointestinal Endoscopy (2000) veröffentlicht. Bei Berman et al. (2000) findet sich für eine Auswahl diagnostischer Maßnahmen eine kritische Wertung. Die Arbeit ist ähnlich konzipiert wie die der ASCO (Desch et al., 1999) – d.h. die Empfehlungen sind versehen jeweils sowohl mit dem Grad der Evidenz aufgrund

existierender Studien als auch mit einem Gewicht der Empfehlung -, jedoch sind die Gesamtheiten der von der ASCO und Berman et al. abgehandelten Maßnahmen nicht identisch, und für die Verfahren, die in beiden Abhandlungen betrachtet werden, stimmen die Urteile der beiden Autorengruppen nicht immer überein.

Für das Mammakarzinom hat eine Expertenkommission der American Society of Clinical Oncology im Jahr 1997 detaillierte Empfehlungen für Maßnahmen der Nachsorgediagnostik, ebenfalls jeweils zusammen mit dem Grad der Evidenz und dem Gewicht der Empfehlung, publiziert (American Society of Clinical Oncology, 1997). Für den Spezialfall der Tumormarker beim Mammakarzinom sowie beim kolorektalen Karzinom waren analog angelegte Empfehlungen bereits 1996 von der ASCO herausgegeben worden (American Society of Clinical Oncology, 1996; vgl. auch American Society of Clinical Oncology, 1998).

In Deutschland hat die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) Leitlinien für die Nachsorge unterschiedlicher Krebserkrankungen erarbeitet und im Internet verfügbar gemacht (<http://www.uni-duesseldorf.de/AWMF/II/index.html>). Gleiches gilt für manche Tumorzentren, so z.B. das Tumorzentrum Heidelberg/Mannheim (<http://www.tumorzentrum-hdma.de/tzreihe.htm>) oder das Tumorzentrum München (<http://www.krebsinfo.de/ki/manuale.html>).

Die Kosten, die die Rezidiv/Metastasendiagnostik in der Tumornachsorge verursacht, sind ebenso heterogen wie die Programme selbst. Für die USA haben Virgo et al. (1995) eine ökonomische Analyse von 11 verschiedenen Überwachungsstrategien beim kolorektalen Karzinom durchgeführt. Die von Medicare zu erstattenden Kosten für die Diagnostik in einer 5-jährigen Nachsorge reichten (auf Basis von Zahlen des Jahres 1992) von \$561 bis \$16 492 pro Patient; letzterer Betrag würde bei flächendeckendem Einsatz des entsprechenden Programms Gesamtkosten von jährlich 1,65 Milliarden Dollar verursachen (Anthony et al., 2000). Beim Mammakarzinom würde nach Schapira (1993) ein im wesentlichen auf klinische Untersuchungen beschränktes Minimalprogramm, verglichen mit einer intensivierten Überwachung, dem amerikanischen Gesundheitswesen (in Preisen von 1990) jährlich über 600 Millionen Dollar einsparen.

Eckardt und Bernhard (1997; zitiert nach Anonym, 1997) bezifferten die Untersuchungskosten des in Deutschland üblichen Nachsorgeprogramms bei

sozialversicherten Patienten auf rund 2700,-- DM pro Patient für eine 5-jährige Nachsorge, wodurch geschätzte jährliche Gesamtkosten von 47,7 Millionen DM entstehen. Hingegen würde ein rationaleres Programm, das 10% weniger Untersuchungstermine vorsieht, auf 30% der Untersuchungen verzichtet und 20% der Patienten (solche mit minimalem Rezidivrisiko) ausschließt, nur jährliche Kosten von rund 19 Millionen DM verursachen.

Die Idee der Nachsorgeprogramme ist scheinbar bestechend plausibel, und in Befragungen hat sich herausgestellt, daß die Mehrzahl der Krebspatienten nach der Primärtherapie eine regelmäßige diagnostische Nachkontrolle wünschen (The Givio Investigators, 1994; Kjeldsen et al., 1999; weitere Referenzen finden sich bei Grunfeld et al., 1999), wohl auch oder vor allem, weil sie davon überzeugt sind, daß dies ihre Prognose verbessert (Edelman et al., 1998). Dies steht nicht im Widerspruch zu den Ergebnissen einer Studie von Gulliford et al. (1997) an 196 Brustkrebspatientinnen, in der sich herausstellte, daß zwei Drittel der Patienten – vor die Alternative gestellt – für seltenere Kontrolluntersuchungen optieren würden.

Der Verweis auf die vermeintliche Plausibilität ist freilich ein ungeeignetes und auch gefährliches Argument, wenn es um die Bewertung medizinischer Interventionen geht (Abel u. Windeler, 1995). Hinzu kommt, daß wie wir sehen werden, im vorliegenden Fall die Plausibilität in Wahrheit nur schlecht begründet ist.

Im folgenden wollen wir uns mit dem Wert der diagnostischen Nachsorgeprogramme aus Sicht des Methodikers näher befassen. Ganz allgemein umfaßt die Bewertung diagnostischer Test oder von Testprogrammen die Untersuchung mehrerer Aspekte (Abel, 1993):

- Praktikabilität
- Validität
- Effizienz (Wirksamkeit)
- Kosten/Nutzen-Relation (aus Sicht des Patienten oder der Gesellschaft).

Hier wollen wir uns schwerpunktmäßig mit dem zweiten und dritten Aspekt auseinandersetzen.

Die Validität einer diagnostischen Maßnahme gibt, grob gesprochen, an, ob und in welchem Maße die Diagnostik das entdeckt, was sie entdecken soll. Ausgedrückt und gemessen wird sie gewöhnlich durch Größen wie die Sensitivität, die Spezifität, den positiven und negativen

prädiktiven Wert oder die Treffsicherheit. Mit dem Begriff der Effizienz oder Wirksamkeit versucht man zu erfassen, ob und inwieweit die Maßnahme zum eigentlichen Ziel beiträgt, das immer einen Nutzen für den Patienten beinhalten muß (Abel u. Windeler, 1996); in der Onkologie ist der Nutzen einer therapeutischen oder diagnostischen Maßnahme, also auch der Nachsorge, letztlich stets daran zu messen, ob sie die Lebenserwartung oder die Lebensqualität der Patienten verbessert (Gallmeier u. Keding 1994; Millat u. Borie, 1998).

2. Das Problem der Validierung

2.1 Vorbemerkung

Aus methodischer Sicht handelt es sich bei der Rezidiv/Metastasendiagnostik in der Tumornachsorge im wesentlichen um die Situation einer diagnostischen Früherkennung, wengleich mit einer weit höheren Prävalenzrate der wahren Positiven, als sie in Screeningprogrammen in der Früherkennung der primären Erkrankung vorkommen. Das bedeutet, daß alle aus der Krebsfrüherkennung bekannten methodischen Probleme (Abel, 1993, S. 89ff) auch hier eine Rolle spielen. Hinzu kommen einige Aspekte, die mit der speziellen Art der Diagnostik in der Nachsorge zu tun haben.

Bei oberflächlicher Betrachtung ist die Validierung der Rezidivdiagnostik trügerisch einfach. Vermutlich rührt dies daher, daß man - vielleicht ungewollt - stets die positiven Befunde und ihren prädiktiven Wert im Blickfeld hat. Die größeren Schwierigkeiten werden jedoch von negativen Befunden aufgeworfen. So ist bereits die Frage, wie ein falsch-negativer Befund überhaupt präzise definiert ist, keineswegs trivial. Problematisch ist somit der negative prädiktive Wert und die Sensitivität der Testung. Letzteres gilt um so mehr, als die konventionelle Definition der Sensitivität, wie wir sehen werden, nicht verwendet werden kann.

Insgesamt ist es nicht ganz einfach, die Validität der Rezidivdiagnostik in der Tumornachsorge exakt in einer Weise zu definieren, die erstens der Intuition entspricht und zweitens einer verzerrungsfreien Schätzung zugänglich ist. Den Autoren ist keine Veröffentlichung über das Thema Nachsorgediagnostik und ihre Validität bekannt, in der eine präzise Klärung der recht verwickelten Begriffe und Probleme unternommen wurde.

In der Literatur wird die Validität meist für einzelne, in der Nachsorgediagnostik eingesetzte diagnostische Maßnahmen oder Tests behandelt, und mit dieser Problemstellung wollen wir uns auch im folgenden befassen. Natürlich ist auch die Frage nach der Validität ganzer diagnostischer Nachsorgeprogramme sinnvoll, sie ist aber methodisch nicht prinzipiell anders gelagert. (Unterschiede gibt es allerdings in der Definition der Außenkriterien, s.u.).

2.2 Wahrer Zustand und Außenkriterium

Im folgenden sei D eine diagnostische Maßnahme, die zu evaluieren ist, z.B. die Messung eines Tumormarkers; (t_0, \dots, t_n) seien die Nachsorgezeitpunkte im postoperativen Follow-up, z.B. $t_0=3$ Monate, $t_1=6$ Monate etc. Der Einfachheit halber wollen wir annehmen, daß D zu jedem dieser Termine durchgeführt werden soll. Ferner soll zu jedem Follow-up-Zeitpunkt eine Reihe weiterer diagnostischer Maßnahmen zur Anwendung kommen, zu denen mindestens die Anamnese und körperliche Untersuchung zählt.

Der in Nachsorgeprogrammen eigentlich interessierende Zustand ist das Rezidiv bzw. die metachrone Metastase. (Um die methodische Erörterung nicht unnötig zu komplizieren, wollen wir den vergleichsweise seltenen Fall der metachronen Primärtumoren, für den andere Erwägungen bei der Validitätsbestimmung vonnöten sind, im folgenden ausklammern.) Es ist keineswegs einfach, eine für Validierungsstudien brauchbare Definition des „tatsächlichen Vorliegens“ von Rezidiven und Metastasen zu geben.

Das Problem liegt darin, daß das neoplastische Gewebe, aus denen manifeste Rezidive und Metastasen hervorgehen, wenn es überhaupt vorhanden ist, in Wahrheit im allgemeinen schon zu Beginn des Follow-up angelegt war (z.B. als Mikrometastase); es war diesem Zeitpunkt lediglich nicht diagnostisch feststellbar. Wir wollen diesen Zustand (tatsächliches Vorhandensein/Nichtvorhandensein von Resttumorgewebe/Mikrometastasen) mit den Symbolen $Z+$ bzw. $Z-$ bezeichnen. $Z+$ und $Z-$ sind zwar präzise definiert, stellen aber keine direkt beobachtbaren Größen dar. Und sie sind auch nicht Zustände, deren Entdeckung man in Wahrheit vom diagnostischen Test verlangt.

Dies kann man sich an einem Beispiel klarmachen: Angenommen, in der postoperativen Kontrolle eines kolorektalen Karzinoms trete ein einmaliger positiver CEA-Befund auf, die anschließende Abklärung sei jedoch negativ. Drei Jahre später werde eine Lebermetastase aufgrund klinischer Symptome entdeckt. In diesem Fall ist Z positiv, doch kann der positive

CEA-Befund (speziell dann, wenn die nachfolgenden Werte wieder im normalen Bereich lagen) schwerlich als richtig positiv gewertet werden.

Umgekehrt wird man die in einem 3-jährigen Follow-up bei einem Patienten gemessenen negativen CEA-Befunde nicht als falsch-negativ einstufen, nur weil der Patient im 7. postoperativen Jahr ein manifestes Rezidiv erleidet.

Mit anderen Worten: In der Rezidivdiagnostik ist der „wahre Zustand“, um den es geht, schwer definierbar. Andererseits ist dies für Zwecke der Validierung auch nicht unbedingt erforderlich, denn hier wird ohnehin generell der Begriff des „wahren Zustands“ durch die Spezifikation eines „Außenkriteriums“ (manchmal auch „gold standard“ oder „Referenzdiagnostik“ genannt) ersetzt. Das Außenkriterium stellt eine *Vereinbarung* dar, die bei der Validierung die Stelle des „wahren Zustands“ einnimmt; alle Validitätsindizes (Sensitivität, Spezifität usw.) sind in einer Validierungsstudie, auch wenn dies nicht immer explizit erwähnt wird, nur relativ zu A definiert und schätzbar (Jensen u. Abel, 2000).

In den existierenden Studien zur Validierung diagnostischer Verfahren in der Tumornachsorge wird generell (vermutlich zumeist ohne daß dies den Autoren bewußt ist) eine *operationale Definition* des Krankheitsrückfalls benutzt: Rezidive/Metastasen werden als tatsächlich vorhanden/nicht vorhanden betrachtet, wenn sie im Laufe des begrenzten Follow-up der Studie, also bis t_n , entdeckt werden, sei es, indem sie symptomatisch werden oder durch die Routinediagnostik in der Nachsorge (sowie im Falle positiver Befunde u.U. einer Folgediagnostik) entdeckt werden.

Ein in dieser Weise operational definiertes Außenkriterium A mit dem Zustand Z gleichzusetzen ist ein verbreiteter Irrtum. Die Nichtübereinstimmung von A und Z ist jedoch offensichtlich: So wird A im Vergleich zu Z (d.h. A als diagnostischer Test mit Außenkriterium Z) falsch-negative Befunde produzieren.

Die Wahrscheinlichkeit für einen positiven Befund des Außenkriteriums A (dies gilt übrigens gleichermaßen für die zu validierende Maßnahme D der Nachsorgediagnostik) hängt von der Größe, Lage und biologischen Beschaffenheit des Relapses ab. In der Gesamtpopulation wächst die Wahrscheinlichkeit für einen positiven Befund monoton mit der seit der Operation verstrichenen Follow-up-Dauer an.

Wegen der Progredienz der Krankheit gilt, daß A um so besser mit Z übereinstimmt, je länger der Follow-up dauert.

So betrug in der sehr umfangreichen Studie (n=1017) von Moertel et al. (1993; vgl. auch Goldberg et al., 1998), die wohl nach wie vor die am sorgfältigsten geplante und durchgeführte Untersuchung zur Validität von CEA im Monitoring des potentiell kurativ operierten kolorektalen Karzinoms ist, die Schicksalskontrolle bei den nicht zuvor verstorbenen Patienten 5 bis 8 Jahre. Nach dieser Zeitspanne ist nur noch mit sehr wenigen Rückfällen zu rechnen.

Aus dem Gesagten folgt auch: Zwei Validierungen derselben diagnostischen Maßnahme (z.B. CEA) werden nicht nur dann unterschiedliche Ergebnisse liefern, wenn die Außenkriterien, also die zur Feststellung des „wahren Zustands“ herangezogenen Tests, verschieden sind, sondern auch dann, wenn sich die Dauer des Follow-up in den Studien unterscheidet.

Denstman et al. (1986) stellten in ihrer relativ umfangreichen (n=214) Studie zur Validität von CEA in der Nachsorge des kolorektalen Karzinoms fest, daß mit zunehmender Länge des Follow-up die Sensitivität und Spezifität monoton abnahmen. Für die Sensitivität ergab sich bei Zugrundelegung eines Schwellenwertes für die Testpositivität von 5ng/ml nach einer Follow-up-Dauer von einem Jahr ein Schätzwert von 89%, nach einem fünfjährigen Follow-up hingegen von 60%, für die Spezifität betragen die entsprechenden Werte 90% bzw. 71%.

Die Feststellung des Außenkriteriums darf bei der Validierung nicht ausschließlich oder häufiger durch einen positiven Befund von D ausgelöst werden, denn andernfalls tritt eine Verzerrung in Form des sogenannten *Verifikationsbias* (auch „*Work-up bias*“ genannt) auf. Wenn z.B. positive Befunde von D stets zu einer Abklärung führen, negative jedoch nur selten, und wenn für die Validierung nur die durch das Außenkriterium abgeklärten Fälle herangezogen werden, dann wird der negative Vorhersagewert von D und die Sensitivität überschätzt. Dieses Phänomen ist in der methodischen Literatur recht eingehend betrachtet worden (Abel 1993, S. 95ff).

Bei der Nachsorgediagnostik tritt der Verifikations-Bias in einer etwas anderen Form auf: Hier werden positive und negative Fälle nicht mit unterschiedlicher Wahrscheinlichkeit,

sondern in unterschiedlicher *Weise* abgeklärt. Naheliegend - und leider wohl üblich - ist es, daß Patienten mit positiven Befunden des zu validierenden Tests gründlicher untersucht werden als solche mit negativen. De facto besteht also bei diesem Vorgehen das Außenkriterium aus *zwei* Komponenten, A1 und A2, von denen die eine bei D+, die andere bei D- zum Einsatz kommt.

Eine Auswahl von Literaturzitate möge dies für den Fall des CEA bei kolorektalen Karzinoms belegen:

„When a significant CEA elevation occurred, the test was repeated. If the repeated CEA value remained elevated, it was considered to be an indication fo a second-look operation“ (Martin et al, 1977)

„Patients with increasing CEA values were studied for evidence of tumor recurrence by physical examination, chest roentgenography, barium enema, sigmoidoscopic examination, liver chemistry analyses and liver scan“ (Steele et al., 1980)

„Investigations for possible tumor recurrence were undertaken for either suggestive clinical findings, or else a confirmed rise of more than 1ng/ml between any 2 consecutive CEA levels regardless of their absolute levels.“ (Boey et al., 1984)

„Abnormal CEA values resulted in subsequent early retesting of CEA or other diagnostic procedures“ (Denstman et al., 1986).

Das methodische Problem bei einer solchen Aufspaltung des Außenkriteriums in A1 (für D+) und A2 (für D-) ist zunächst, daß *unabhängig vom Test D*, der ja evaluiert werden soll, gar nicht mehr gesagt werden kann, was A+ überhaupt bedeutet. Und die in einer Studie geschätzte Validität, wird sich von jener unterscheiden, die man bei Verwendung eines einheitlichen, bei allen Patienten angewandten Außenkriteriums erhalte.

Relativ gering dürfte die Verzerrung dann sein, wenn der zu evaluierende Test D in ein umfassendes diagnostisches Programm eingebettet ist, das routinemäßig und unabhängig von dem Befund von D durchgeführt wird.

Ist dies nicht der Fall, werden also außer D etwa nur klinische Befunde erhoben, so wird die Verzerrung um so größer sein, je kürzer der Follow-up ist. Zu der Verzerrung tragen nämlich alle Rückfälle bei, die bis zum Ende des Follow-up von D nicht angezeigt und in der klinischen Untersuchung nicht entdeckt werden, jedoch von der gründlicheren Diagnostik A1 entdeckt würden. Je kürzer der Follow-up ist, um so häufiger sind solche Fälle, und zwar deshalb, weil die Inzidenzrate der Rückfälle mit der Follow-up-Dauer abnimmt.

Anzumerken wäre noch, daß in den publizierten Validierungsstudien dem Verifikationsbias keine Beachtung geschenkt wird.

Ein anderes Problem, das bei der Spezifikation des Außenkriteriums auftritt, ist der sogenannten „*Informationsbias*“.

Informationsbias tritt auf, wenn der zu evaluierende Diagnostest und das Außenkriterium nicht unabhängig voneinander festgestellt werden, wenn der Befund von D also bei der Feststellung von A bekannt ist und das Ergebnis von A beeinflussen kann (oder umgekehrt). Dies ist bei diagnostischen Maßnahmen, die vom subjektiven Urteil abhängen, nie auszuschließen. So muß man damit rechnen, daß die Interpretation eines Röntgenbildes oder die Befundung in einer Sonographie auch davon beeinflusst wird, ob dem Diagnostiker ein unmittelbar zuvor gemessener extrem erhöhter Tumormarkerwert beim betreffenden Patienten bekannt war oder nicht.

Informationsbias führt dazu, daß sämtliche Validitätsmaße zu günstig eingeschätzt werden. Sicher eliminieren läßt er sich durch gegenseitige *Verblindung* bei der Befundung durch D und A, was aber natürlich entsprechende Vorkehrungen in der Planung der Evaluierung voraussetzt.

Den Verfassern ist keine Validierungsstudie von diagnostischen Maßnahmen in der Tumornachsorge bekannt, die dem Problem des Informationsbias methodisch Rechnung trägt.

Bei der simultanen Evaluierung mehrerer diagnostischer Maßnahmen D_1, \dots, D_k , die zu einem Nachsorgeprogramm gehören, tritt schließlich ein weiteres Problem auf. Die zu evaluierende Maßnahme darf niemals *Teil* des Außenkriteriums sein, da sonst eine Verzerrung auftritt, die in der Methodik als *Einbeziehungsbias* bekannt ist. Einbeziehungsbias führt naheliegenderweise zu einer Überschätzung der Validität. Für die Abklärung von D_i dürfen also nur die übrigen Maßnahmen herangezogen werden. Mit anderen Worten: Das Außenkriterium ist für jedes der D_i verschieden; es besteht nämlich aus einer Kombination der

D_j ($j \neq i$) nebst u.U. weiteren diagnostischen Maßnahmen. Die gewonnenen Validitätsschätzungen für die D_i sind nicht mehr unmittelbar miteinander vergleichbar. Richard u. McLeod (1997) bemerkten in ihrer Übersichtsarbeit treffend: „Most studies looking at the value of follow-up programs have not assessed the different tests independently from one another.“

2.3 Test-Positivität

Zumindest dann, wenn die Diagnostik D aus der Messung eines Tumormarkers besteht oder eine solche Messung umfaßt, wirft die Frage, unter welchen Umständen der Befund von D als positiv zu werten ist, definitorische Probleme auf, die allen zeitabhängigen, wiederholten Messungen zueigen sind.

Grundsätzlich gibt es bei Verlaufsmessungen unbegrenzt viele Möglichkeiten und Vorgehensweisen, Test-Positivität zu definieren. Grundlage kann hierfür z.B. der einzelne Befund oder eine mathematische Zusammenfassung mehrerer aufeinanderfolgender Werte sein; dieser Einzel- oder Gesamtbefund kann für sich genommen beurteilt werden oder aber verglichen werden mit dem Ausgangswert des Patienten oder – alternativ - mit einer sogenannten Baseline aus mehreren zu Beginn des Follow-up gemessenen Werten; die Test-Positivität kann durch einen Schwellenwert (für den absoluten Meßwert, die relative Veränderung oder auch die Steigung des Anstiegs aufeinanderfolgender Meßwerte) definiert werden, wobei für die Spezifikation dieses Schwellenwertes wiederum ein Kontinuum von Möglichkeiten existiert.

Dabei ist zu beachten, daß verschiedene Schwellenwerte jeweils unterschiedliche binäre Tests definieren. Durch Verschiebung des Schwellenwerts kann man die Validität der zugehörigen Tests verändern: Je höher der Schwellenwert gewählt wird, um so geringer wird die Sensitivität und um so höher die Spezifität des resultierenden Tests ausfallen.

Hierbei tritt im übrigen das Problem auf, daß die Schätzung der Validität infolge der sogenannten *Resubstitution* verzerrt sein kann. Wenn man nämlich einen Datensatz für die Spezifikation eines geeigneten Schwellenwertes für den Diagnostest benutzt und anschließend dieselben Daten heranzieht, um die Validität des so definierten Tests zu evaluieren, dann wird sowohl die Sensitivität als auch die Spezifität tendenziell überschätzt. Diese Verzerrung ist im Mittel um so größer, je kleiner die Stichprobe ist (Abel u. Berger, 1986).

Im Licht dieser Bemerkungen ist die Heterogenität der in publizierten Studien zu findenden Kriterien für Testpositivität bei Tumormarkerbestimmungen nicht verwunderlich.

So wurden beispielsweise in den Validierungen des Tumormarkers CEA beim kolorektalen Karzinom von Steele et al. (1980), Wood et al. (1980), Minton et al. (1985), Staab et al. (1985), Abel et al. (1986), Denstman et al. (1986) und McCall et al. (1994) jeweils unterschiedliche Definitionen für auffällige Befunde benutzt, bei Lavin et al. (1981) sowie Moertel et al. (1993) wurden sogar an einem Datensatz verschiedene Kriterien ausprobiert und miteinander verglichen. Erwähnenswert ist auch die Untersuchung von Langhammer et al. (1994), in der an Daten aus der Nachsorge von 2224 Mammakarzinompatientinnen retrospektiv verschiedene Schwellenwerte der Tumormarker CEA und CA15-3 ausprobiert wurden.

2.4 Sensitivität, Spezifität und Vorhersagewerte

Die Sensitivität einer diagnostischen Maßnahme D ist allgemein definiert als (bedingte) Wahrscheinlichkeit dafür, daß D positiv befundet, wenn das zugrundegelegte Außenkriterium positiv ist; die Spezifität ist die Wahrscheinlichkeit für einen negativen Befund von D bei Personen, die einen negativen Befund von A aufweisen.

In der Situation der Nachsorgediagnostik sind bereits diese scheinbar harmlosen Definitionen aus mehreren Gründen problematisch.

1. Die Validität einer diagnostischen Maßnahme ist generell kontextabhängig. Mindestens drei Aspekte sind zu berücksichtigen:

a) Die Eigenschaften der Population.

Sie beeinflusst sowohl die Sensitivität als auch die Spezifität einer Maßnahme. Insbesondere wird z.B. die Sensitivität zur Entdeckung von Rezidiven beeinflusst von der Verteilung der Art und Lage der Rezidive in der Population; die Spezifität wird allgemein durch „symptomatisch konkurrierende“ Erkrankungen vermindert (z.B. gutartige Gewebsveränderungen), und deren Häufigkeit hängt ihrerseits von der Altersverteilung der Population ab.

b) Die Unsicherheit der Befundung durch die Maßnahme selbst.

Im Falle von Labortests drückt sich diese Unsicherheit durch zufällige und systematische Abweichungen der Ermittlungsergebnisse vom wahren oder richtigen Wert aus; bei

bildgebenden Verfahren äußert sie sich in der Beobachtersvariabilität. Stets spielen schwer faßbare Aspekte wie die Qualität der verwendeten Geräte oder die Erfahrung des Diagnostikers eine Rolle.

c) Die Einbettung des Tests in eine diagnostische Strategie.

Für eine Maßnahme ergibt sich generell eine unterschiedliche Validität, je nachdem ob sie für sich genommen evaluiert wird oder als Baustein einer komplexen Teststrategie. Im Falle der Tumornachsorge scheint dies keine Rolle zu spielen, da selbst bei komplexen Nachsorgeschemata die einzelnen Tests nicht im Rahmen einer *Sequenz* evaluiert werden, sondern als Früherkennungstests gleichberechtigt nebeneinander stehen (parallele Testdurchführung nebst Interpretation nach der „believe-the-positive-Regel“). Allerdings hat die Einbettung Auswirkungen auf die Art des Außenkriteriums (s. den nächsten Punkt).

2. Wie wir gesehen haben, ist das Außenkriterium A in Validierungsstudien der Nachsorgediagnostik nicht gleichzusetzen mit dem Zustand „Z“ („Resttumorgewebe/ Mikrometastasen vorhanden bzw. nicht vorhanden“).

Die sich in einer Validierungsstudie ergebende apparente Sensitivität und Spezifität ist daher nicht identisch mit der, die sich bei Kenntnis von Z+ und Z- ergäbe. Der Unterschied zwischen Z und A hängt wegen der Progredienz der Erkrankung von der Länge des Follow-up ab. Man kann mathematisch zeigen, daß durch Benutzung von A die Sensitivität von D höher ausfällt als bei Benutzung von Z.

Wie wir gesehen haben, werden in publizierten Validierungsstudien de facto zwei Außenkriterien eingesetzt, nämlich eine intensive Abklärung A1 für D+ und eine weniger intensive A2 für D-. Dies hat zur Folge, daß die so berechneten Validitätsindizes sich von jenen unterscheiden, die sich bei Verwendung eines einheitlichen Kriteriums, etwa A1, ergäben. Dieser work-up bias führt dazu, daß sämtliche publizierten Sensitivitätsschätzungen verzerrt, nämlich tendenziell zu hoch sind.

3. Offenbar interessiert man sich bei der Validierung überhaupt nicht für die konventionell definierte Sensitivität. Denn bei buchstabengetreuer Auslegung würden zu ihr auch Fälle (im Zähler) gerechnet, die nicht für den Wert der Nachsorge sprechen. Gemeint sind vor allem Fälle, in denen die Nachsorge den Relapse erst entdeckt, wenn oder sogar nachdem er klinisch symptomatisch geworden ist.

Bei der intensivierten Nachsorgediagnostik geht es jedoch nicht um die Entdeckung per se, sondern um eine *Vorverlegung* der Entdeckung oder - anders ausgedrückt – um eine Entdeckung des Rückfalls bei asymptomatischen Patienten. Diese Vorverlegung, also das Intervall zwischen einem positivem Befund von D und dem Auftreten klinischer Krankheitssymptome, wird „lead time“ genannt. (Wir wollen sie mit L abkürzen.)

Trägt man dieser Überlegung Rechnung, so könnte man versuchen, die Sensitivität, leicht modifiziert, so zu definieren

(*) Se = Wahrscheinlichkeit dafür, daß bei einem Patienten mit A+ die lead time L positiv ist.

(Dies entspricht der von McCall et al., 1994 gegebenen Definition). Für die Spezifität hat übrigens eine eventuell geforderte (Mindest)-lead time keine Bedeutung; hier kann die klassische Begriffsbestimmung verwendet werden.

Die Definition (*) erscheint harmlos, doch zeigt eine nähere Analyse, daß auch sie in dieser Form nicht sinnvoll ist. Erstens löst sie ohne weitere Präzisierung nicht das angesprochene Problem: Natürlich ist man nicht an *jeder* (beliebig kleinen) lead time interessiert, sondern man wird verlangen, daß sie ein Minimum Δ_0 nicht unterschreitet ($L > \Delta_0$).

Zweitens umfaßt (*) nach wie vor solche Fälle, bei denen der mittels der Sensitivität quantifizierte Begriff des „Entdeckens“ intuitiv fehl am Platz ist. Dies trifft zu für positive Befunde von D, die transient sind und denen z.B. viele Monate oder Jahre negativer Befunde folgen. Wenn bei einem derartigen Verlauf während der Phase, in der D negativ ist, ein Tumorrückfall etwa durch die Symptome entdeckt wird, kann man gewiß nicht von einem richtig positiven Befund von D sprechen. Mit anderen Worten: Man wird als zusätzliche Bedingung verlangen, daß der letzte positive Befund von D, der vor dem durch Symptome angezeigten Rückfall beobachtet wurde, nicht mehr als eine maximale Zeitspanne Δ_1 zurückliegt. Diese Bedingung werde als $B \leq \Delta_1$ abgekürzt. Hieraus ergibt sich die Definition

(**) Se = Wahrscheinlichkeit dafür, daß bei einem Patienten mit A+ gilt:
 $L \leq \Delta_0$ und $B \leq \Delta_1$.

Die so definiert Sensitivität hängt dann auch von Δ_0 und Δ_1 ab; sie wird beispielsweise um so kleiner ausfallen, je größer Δ_0 gewählt wird.

Daß auch sie nicht sinnvoll ist, hängt mit den Vorhersagewerten zusammen. Der positive (negative) Vorhersagewert einer diagnostischen Maßnahme ist definiert als bedingte Wahrscheinlichkeit dafür, daß ein positiver (negativer) Testbefund durch das Außenkriterium bestätigt wird.

Betrachten wir nun die Sensitivität und den positiven Vorhersagewert. Bei beiden handelt es sich um bedingte Wahrscheinlichkeiten; der Nenner ihrer Schätzwerte ist durch die Bedingung der Wahrscheinlichkeit gegeben: Der Nenner des Schätzwerts der Sensitivität besteht aus der Anzahl der Fälle, bei denen in der Validierungsstudie A+ festgestellt wird, der Nenner des positiven Vorhersagewerts aus allen Fällen, bei denen in der Validierungsstudie D+ beobachtet wird. Eventuelle Forderungen an die lead-time wirken sich auf den Zähler, nicht aber den Nenner aus.

Wie ist nun mit einem Fall zu verfahren, der als (D+,A+) eingestuft ist, bei dem jedoch die lead time gleich Null ist? Rechnet man ihn als richtig-positiv, so steht dies im Widerspruch zu der in (**) enthaltenen Forderung; rechnet man ihn als falsch-positiv, so wird er im Nenner des positiven Vorhersagewerts berücksichtigt, jedoch fälschlich aus dem Nenner der Sensitivität weggelassen; rechnet man ihn als falsch-negativ, so ist es umgekehrt. (Man beachte, daß bei einer willkürlichen Einstufung einiger positiver Test-Befunde als falsch-negativ der negative Vorhersagewert der Maßnahme auch dann gering ausfallen kann, wenn im Gefolge negativer Test-Befunde niemals ein Rezidiv beobachtet wird.) Eine Klassifikation von (D+,A+) als richtig-negativ kommt schließlich offenbar überhaupt nicht in Frage. Mit anderen Worten: Die Definition (**) ist mit einer inhaltlich sinnvollen Definition der Vorhersagewerte nicht vereinbar.

Für Methodiker sei angemerkt, daß sich die Inadäquatheit von (*) und (**) auch darin zeigt, daß bei Zugrundelegung einer positiven lead-time in den Definitionen der Sensitivität sowie eventuell des positiven Vorhersagewertes die Bayes-Formel für den Übergang von Sensitivität und Spezifität zu den Vorhersagewerten nicht mehr gültig ist.

Daß die aufgezeigten Probleme und die durch sie verursachten Begriffsverwirrungen nicht nur akademischer Natur sind, lehrt z.B. die Studie von Lavin et al. (1981), in der man die folgenden Begriffsbestimmungen findet: „Correctly predicted recurrence rate: percentage with recurrence among those patients with follow-up CEA event occurring beforehand“ und

„Incorrectly predicted recurrence rate: percentage with follow-up CEA event occurring among those patients with no recurrence“ (sic!).

Erwähnt sei noch ein weiteres methodisches Problem mit den Definitionen (*) und (**). Die hier präzisierte Wahrscheinlichkeit (Sensitivität) läßt sich in Validierungsstudien kaum schätzen. Der Grund liegt darin, daß die Bedingung der relevanten lead time ($\text{lead time} > \Delta_0$) beim einzelnen Patienten *unbeobachtbar* ist, wenn im Falle von D+ entweder sofort oder nach Bestätigung durch aufwendigere diagnostische Maßnahmen mit einer Therapie begonnen wird, die die Krankheitssymptome, welche ohne Therapie aufgetreten wären, verhindern kann.

4. Bei der Validierung einer Rezidivdiagnostik sind einige weitere Dinge von Bedeutung.

Erstens ist bei der Schätzung der Validitätsmaße der Informationsbias (s.o.) zu vermeiden, was jedoch in retrospektiven Analysen kaum möglich ist. Validierungen sind daher *prospektiv* durchzuführen und sorgfältig zu planen (wechselseitige Verblindung von A und D).

Zweitens sind Schätzungen der Validitätsmaße mit einer statistischen Ungenauigkeit behaftet, die durch Angabe eines Vertrauensbereiches für den wahren Parameter deutlich zu machen ist. Die Ungenauigkeit ist um so größer, je kleiner die Fallzahl ist.

Drittens spricht eine hohe Sensitivität des Programms nicht für die diagnostische Testung, denn, wie wir gesehen haben, läßt sich z.B. im Falle quantitativer Tests eine hohe Sensitivität auf Kosten der Spezifität willkürlich allein durch Verschiebung des Schwellenwertes erreichen. Übrigens gilt gleiches im allgemeinen, wenn auch nicht mathematisch zwingend, für die Vorhersagewerte. Moertel et al. (1993) fanden, daß durch Senkung der Schwellenwerte für Positivität des CEA (15, 10, 5, 3ng/ml) der auf Basis desselben Datensatzes berechnete positive prädiktive Wert des CEA von 96% auf 57% abnahm, während der negative prädiktive Wert sich von 69% auf 77% erhöhte. Sogar eine Sensitivität von 100% ist ohne weiteres erreichbar, nämlich mit Nonsenstests, die jeden Patienten als Test-positiv klassifizieren.

Umgekehrt spricht aber eine *niedrige* Sensitivität *gegen* den Test, weil dann eine hohe Validität auch bei hundertprozentiger Spezifität nicht erreichbar ist.

Mit Blick auf die angestellten Überlegungen verwundert es nicht, daß die publizierten Angaben zur Validität der Nachsorgediagnostik äußerst widersprüchlich sind bzw. eine enorme Spannweite aufweisen, vgl. die von Fletcher (1986), Edelman et al. (1997), Virgo et al. (1997), Richard u. McLeod (1997), Millat u. Borie (1998), Bedenne u. Jouve (1999) und Berman et al. (2000) zusammengetragenen Angaben. So reicht die von Richard u. McLeod (1997) angegebene Spannweite für die in unterschiedlichen Studien gefundene Sensitivität des CEA in der Nachsorge des kolorektalen Karzinoms von 58% bis 89%, für die Spezifität von 75% bis 98%. Noch größer ist die Ergebnisspreizung, die Fletcher für das CEA angibt: Sensitivität: 17-89%, Spezifität: 34-91%.

Zahlreiche Studien widmen sich der Frage, welcher Prozentsatz von Krankheitsrückfällen zuerst durch die einzelnen diagnostischen Maßnahmen eines Nachsorgeprogramms entdeckt werden.

In der retrospektiven Untersuchung von Pandya et al. (1985) wurden 175 Rezidive, die in der Nachsorge von 856 Mammakarzinompatientinnen entdeckt wurden, nach der Lokalisation des Relapses und dem ersten Indikator aufgeschlüsselt. Insgesamt am besten schnitten die Symptome und die körperliche Untersuchung ab, die in 63 bzw. 32 Fällen den Relapse als erste anzeigten. Erwartungsgemäß erwies sich für jedes diagnostische Verfahren, daß der Anteil der Rezidive, die hierdurch erstmalig entdeckt wurden, stark von der Lokalisation des Relapses abhing. So waren die Symptome bei 10 von 69 lokoregionären Rückfällen, bei 43 von 66 Knochenmetastasen, bei 19 von 44 viszeralen Metastasen und bei 1 von 10 der übrigen Rückfälle der erste Indikator.

Ähnliche Untersuchungen oder Zusammenstellungen, zumeist mit weniger gründlichen Aufschlüsselungen, finden sich z.B. bei Winchester et al. (1979), Camunas et al. (1991), Schapira (1993), Loprinzi (1994), Peethambaram et al. (1997) und Bleeker et al. (2001).

Die in diesen Studien gefundenen Entdeckungsraten können allerdings nicht als Argument zugunsten des Einsatzes einer Maßnahme dienen: Erstens fehlen die erforderlichen Angaben der Spezifität, und zweitens bleibt unklar, ob mit der Vorverlegung der Entdeckung auch eine relevante lead-time verbunden ist.

Diese Frage nach der lead-time – oder auch nur die einfachere Frage nach der Zeit bis zur Entdeckung von Rückfällen durch ein bestimmtes diagnostisches Verfahren - ist prinzipiell in einarmigen Studien dieses Typs nicht verzerrungsfrei zu klären, und Untersuchungen, die dies unternehmen (z.B. Minton et al., 1985; Staab et al., 1985; Bleeker et al., 2001), sind fehlerhaft. Der Grund liegt darin, daß - wie bereits erwähnt - in solchen Studien die Ereignisse, die gemessen werden, nicht statistisch unabhängig voneinander sind. So fallen bei der Berechnung des Zeitintervalls bis zur Entdeckung von Rezidiven durch Symptome alle Rezidive aus der Berechnung heraus, die zuvor durch andere Verfahren diagnostiziert wurden.

Gleichwohl sind Untersuchungen der Frage, welche diagnostischen Verfahren Rezidive zuerst entdecken, nicht ohne jeden Wert; eine Aussagekraft können sie durchaus im *negativen* Sinne haben. So wurde in manchen Untersuchungen (z.B. Scanlon et al., 1980; Pandya et al., 1986; Shapira et al., 1993; Donnelly et al., 2001) beobachtet, daß die meisten Rückfälle erstmalig durch die Symptome und klinische Untersuchungen auffällig werden. Je höher dieser Prozentsatz ist, um so geringer ist zwangsläufig der mögliche Wert einer zusätzlichen Diagnostik.

Dieser Argumentationslinie folgt die Feststellung in einem Editorial mit Bezug auf das kolorektale Karzinom (Anonym, 1997): „Der Anteil von Rezidiven, die primär über Sonographie, CT, Röntgenthorax oder Endoskopie entdeckt werden, ist dagegen äußerst gering. Vor allem das „Routinelabor“ ist bei der Suche nach einem Rezidiv so ineffizient, daß man es getrost aus dem Nachsorgeprogramm streichen kann.“

3. „Effizienz“ und Nutzen von Nachsorgeprogrammen – methodische Aspekte

3.1 Notwendige Bedingungen für die Effizienz

Zwar haben sich die Krebsnachsorgeprogramme bereits in den siebziger und achtziger Jahren zu einer formidablen Maschinerie entwickelt, doch wurde die Frage nach Effizienz (Wirksamkeit) und Nutzen anscheinend erst in den neunziger Jahren näher untersucht. Dies ist um so erstaunlicher, als bereits einige einfache Überlegungen ausreichen, um die Rationale der Programme sehr zweifelhaft erscheinen zu lassen.

Auch ohne profunde methodische Kenntnisse leuchtet sofort ein, daß sich die Nachsorge nur dann günstig auf die Prognose der Patienten auswirken kann, wenn nach der Entdeckung eines Tumorrückfalls auch eine lebensverlängernde Therapie zur Verfügung steht. Dabei kommt allerdings erschwerend hinzu, daß sogar *fünf* Bedingungen erfüllt sein müssen, damit die Nachsorge einen erkennbaren lebensverlängernden Effekt haben kann (vgl. Kievit u. Bruinvels, 1995; Schumacher et al., 1995; Anonym, 1997):

Erstens muß es in der Anwendungspopulation eine gewisse Mindestinzidenzrate der Rezidive geben.

Zweitens muß die Diagnostik der Tumorrezidive oder Metastasen eine hohe Validität aufweisen. Sie muß also die Rückfälle erkennen können, bevor sie symptomatisch werden. Dabei darf sie aber andererseits nur wenige falsch-positive Befunde anzeigen, denn dies würde den Patienten nicht nur unnötig beunruhigen, sondern würde ihn, wenn die Diagnostik zu therapeutischen Interventionen führt, auch unnötig oder unnötig früh mit den Neben- oder Folgewirkungen der Therapie belasten. Infolge einer zu geringen Spezifität kann die Testung also Schaden anrichten und sogar die Lebenserwartung der Gesamtheit der Patienten verkürzen.

Drittens muß die lead time nicht nur positiv sein, sondern mindestens einige Monate betragen, denn andernfalls ist ein Profit für den Patienten von vornherein nicht zu erwarten.

Viertens muß, wie gesagt, die aufgrund von positiven Befunden eingeleitete Therapie lebensverlängernd wirken.

Fünftens, und diese Bedingung ist besonders hart, muß es für die Lebenserwartung der Patienten einen Unterschied ausmachen, ob man die Behandlung früher oder etwas später verabreicht; später, das heißt hier: wenn der Rückfall ohnehin symptomatisch wird.

Manche Autoren ersetzen (im Falle des kolorektalen Karzinoms) die vierte und fünfte Bedingung durch die Forderung (F), die Diagnostik müsse einen hohen Anteil resezierbarer Rezidive entdecken (z.B. Kievit, 2000). Diese Substitution ist allerdings nicht stichhaltig. Nicht nur impliziert (F), wie wir weiter unten sehen werden, nicht die Bedingungen 4 und 5;

sondern auch die Umkehrung gilt nicht: Die Bedingungen 4 und 5 können auch erfüllt sein, wenn die Diagnostik überhaupt keine resezierbaren Rückfälle zu entdecken vermag. Es würde ausreichen, daß eine *systemische* Therapie zur Verfügung steht, deren früherer Einsatz im Falle des Rezidivs von Vorteil ist. (Der Nachweis eines solchen Vorteils muß in randomisierten Vergleichen einer sofortigen mit einer verzögerten Anwendung geführt werden.)

Häufig wird als weitere Bedingung genannt, daß die Diagnostik auch unter Kosten-Nutzen oder Kosten-Effektivitätsaspekten günstig abschneiden müsse. Diese Aspekte sind von Bedeutung für die *Entscheidung*, ob eine intensivierete Nachsorgediagnostik im Einzelfall durchgeführt werden sollte oder gar ihre routinemäßige Einführung sinnvoll ist, sie berühren jedoch strenggenommen nicht die eingeschränktere Frage nach der Wirksamkeit (Lebensverlängerung).

3.2 Kritik der Standardargumentation für das diagnostische Monitoring

Macht man sich klar, wie schwer erfüllbar die fünf genannten Anforderungen zusammengenommen sind, so verwundert es eigentlich, warum man erst so spät angefangen hat, am Nutzen der Nachsorge zu zweifeln. Bis zur Mitte der neunziger Jahre wurde eine engmaschige, intensive Nachsorgediagnostik mit folgenden Argumenten gerechtfertigt bzw. befürwortet (vgl. Liberati et al., 1997); man findet sie, wenn auch seltener, bis in die jüngste Zeit:

Argument 1: Die Entdeckung von Rückfällen wird durch das Monitoring vorverlegt.

Dieses Argument ist sicherlich die ursprüngliche Grundlage für Nachsorgediagnostik überhaupt. Daß es auch in jüngerer Zeit noch eine praktische Rolle spielt, zeigen die 1996 herausgegebenen Empfehlungen der ASCO für die Durchführung von CEA-Messungen in der Nachsorge des Kolokarzinoms, die sich maßgeblich auf diese Begründung stützen (American Society of Clinical Oncology 1996).

Mit diesem Argument haben wir uns bereits mehrfach auseinandergesetzt. Zusammenfassend sei nochmals festgestellt: Eine Vorverlegung per se ist unerheblich. Aber selbst der Nachweis einer *relevanten* Vorverlegung spricht nur dann für die Diagnostik, wenn sie mit einer Verlängerung der Lebenserwartung der Patienten einhergeht. Die Lebenserwartung muß sich

in der Gesamtkohorte, auf die die Maßnahme angewandt wird verbessern. Dies kann man selbstverständlich nicht einfach voraussetzen; man kann sich, wie wir gesehen haben, nicht einmal darauf berufen, daß es sonderlich plausibel sei.

Argument 1, in dem es um die Sensitivität der Maßnahmen geht, überzeugt nicht, denn es ist nur eine *notwendige*, aber keine *hinreichende* Bedingung dafür, daß die Diagnostik sich positiv auf die Lebenserwartung auswirken kann. Zu beachten ist, daß dies eine Asymmetrie der Aussagekraft von Studien zur Folge hat: Follow-up-Studien, die die Validität der Nachsorge untersuchen, besagen (auch dann wenn man von den besprochenen methodischen Schwächen einmal absieht) nichts über den Nutzen der Programme, wenn ihr Resultat *günstig* ist; ein *ungünstiges* Resultat (d.h. die Erkenntnis, daß das Programm nur eine geringe Validität aufweist) macht hingegen jedweden Nutzen unwahrscheinlich und unplausibel.

Argument 2: In Untersuchungen zeigt sich, daß die durch diagnostische Tests oder Verfahren entdeckten Rückfälle eine bessere Prognose aufweisen als jene, auf die man durch Symptome aufmerksam wird.

Ein Beispiel für eine Arbeit, die diesem Aspekt Bedeutung beimißt, ist die Meta-Analyse von Rosen et al. (1998), in der mehrfach ausdrücklich die deutlich längere Restüberlebenszeit nach Rückfall hervorgehoben wird, die Patienten mit intensivierter im Vergleich zu weniger intensiver Nachsorgediagnostik aufweisen. Einer analogen Argumentation folgt die Studie von Eckardt et al. (1994), in der beobachtet wurde, daß Patienten, deren Rezidiv durch Kontrolluntersuchungen im asymptomatischen Zustand entdeckt wurde, eine bessere Prognose aufwiesen als Patienten, deren Rückfall durch Symptome angezeigt wurde.

Argument 2 ist aus zwei Gründen untauglich. Erstens tritt dieser Effekt notwendig und rein rechnerisch ein, wenn die lead time positiv ist („lead-time bias“). Er folgt zwangsläufig aus der Vorverlegung, auch wenn die Früherkennung ursächlich überhaupt nichts zur tatsächlichen Lebenserwartung des Patienten beiträgt. Man würde den Effekt auch mit einer Nonsensediagnostik erhalten, die bei jedem Patienten 3 Monate nach der Operation einen Rückfall „diagnostiziert“.

Zweitens ist anzunehmen, daß in der Nachsorge der aus der Krebsfrüherkennung bekannte „length bias“ (Abel 1993; S. 115) ebenfalls zum Tragen kommt: Die Wahrscheinlichkeit, daß Rückfälle zwischen den Nachsorgeterminen durch Symptome auffällig werden, ist um so

höher, je schneller und aggressiver der Krebs wächst. Im Umkehrschluß folgt daraus, daß die in der Nachsorgediagnostik erkannten Rückfälle eine Selektion darstellen, in der langsam wachsende Krebserkrankungen überrepräsentiert sind. Da das Überleben nach dem Rückfall stark von der Wachstumsgeschwindigkeit abhängt, hat dies zur Folge, daß die in der Nachsorgediagnostik erkannten Rückfälle auch dann im Mittel eine bessere Prognose aufweisen, wenn die Nachsorge ursächlich keinen Einfluß auf die Prognose ausübt.

Kurz gesagt ist anzunehmen, daß Patienten, bei denen durch die intensive Nachsorgediagnostik ein Rückfall entdeckt wird, *von vornherein* eine günstigere Prognose aufweisen als Patienten, bei denen der Rückfall durch Symptome angezeigt wird.

Argument 3a: Unter den potentiell kurativ resezierbaren Rezidiven (z.B. operable Lokalrezidive, isolierte Lebermetastasen), die in Studien der intensivierten diagnostischen Nachsorge aufgefunden werden, wird ein hoher Prozentsatz durch die Nachsorgediagnostik (und nicht erst durch die Symptome) entdeckt.

Argument 3b: In vergleichenden Studien „intensive versus weniger intensive diagnostische Nachsorge“ sind die Rückfälle, die in der intensiver überwachten Gruppe entdeckt werden, zu einem größeren Anteil potentiell kurativ resezierbar als die in der Vergleichsgruppe entdeckten.

Man beachte, daß die beiden Argumente ähnlich, aber nicht äquivalent sind. Auffällig ist, daß die Argumente 3a,b beim kolorektalen Karzinom, kaum jedoch z.B. beim Mammakarzinom, eine Rolle spielen. Der Grund liegt möglicherweise darin, daß beim Mammakarzinom – anders als beim kolorektalen Karzinom – Rückfälle, die als potentiell kurativ behandelbar betrachtet werden, relativ selten sind.

Wir wollen zunächst Beispiele dafür anführen, daß Argumente dieses Typs tatsächlich dazu herangezogen werden, den Erfolg oder Nutzen der Nachsorge zu begründen.

O'Dwyer et al. (2001) befassen sich mit Studienergebnissen, die Argumente des angeführten Typs für das kolorektale Karzinom untermauern (wobei allerdings keine scharfe Unterscheidung zwischen 3a und 3b vorgenommen wird); im Abstract ihrer Arbeit schreiben sie: “Analyses of Relapses following adjuvant therapy support an even more

aggressive schedule, with the goal of maximizing the proportion of patients who may be operated on with curative intent...”

Bei Ohlsson et al. (1995) heißt es: „Success of the follow-up is determined by the proportion of asymptomatic patients undergoing curative treatment of recurrence“.

Ergebnisse der Art, wie sie in Argument 3a angesprochen werden, finden sich in der Literatur recht häufig (vgl. die Zitate bei Anthony et al., 2000). Hervorzuheben ist die umfangreiche Intergroup-Studie INT-0035 von Goldberg et al. (1998), in der es um den Wert von CEA ging. Unter 248 Rezidiven in dieser Studie wurde in 109 Fällen eine Reoperation mit kurativer Zielsetzung durchgeführt. Als häufigster „motivierender Faktor“ wurde in 41 Fällen ein erhöhter CEA-Wert genannt, gefolgt von den Symptomen (27 Fälle).

In einer neueren Untersuchung von Bleeker et al. (2001), die in eine randomisierte Therapiestudie an 496 Kolonkarzinompatienten des Stadiums Dukes C eingebettet war, wurde über 213 Zweitoperationen wegen eines Rezidivs berichtet, von denen 42 mit kurativer Zielsetzung durchgeführt wurden. Unter diesen 42 Fällen war Ultraschall/CT der Leber das Instrument, welches am häufigsten den Rückfall identifiziert hatte. Im Vergleich hierzu wurden nur 12 der 42 Fälle erstmalig durch Symptome auffällig.

Argument 3b gehörte zu den Untersuchungsgegenständen der in der Meta-Analyse von Rosen et al. (1998) berücksichtigten vergleichenden Studien. Rosen et al. errechneten aus diesen Ergebnissen, daß bei intensiver überwachten Kolonkarzinompatienten mit schätzungsweise 2.4-fach höherer Wahrscheinlichkeit eine kurative Resektion durchführbar war als bei weniger intensiv überwachten. (Die nichtadjustierten zusammengefaßten Anteile betragen 25.8% versus 8.8%.)

Einen Unterschied zugunsten des intensivierten Follow-up beim kolorektalen Karzinom hatten zuvor bereits Bruinvels et al. (1994) in einer Meta-Analyse konstatiert. In fünf vergleichenden Studien mit insgesamt 2824 Patienten wurden 971 Rückfälle registriert, von denen 388 operativ behandelt wurden, 113 davon potentiell kurativ. Von 215 Rückfällen in den Gruppen mit intensiverem Follow-up waren 76 als potentiell kurativ resezierbar dokumentiert, unter 173 Rückfällen in den Gruppen mit minimalem Follow-up hingegen nur

37. Als adjustierter Schätzwert für die Differenz der Wahrscheinlichkeit für eine kurative Operation (intensiver versus minimaler Follow-up) ergab sich 23.1%; in Studien, bei denen CEA-Messungen eingesetzt wurde, betrug die Differenz 25.5%.

Die Argumente 3a und 3b behaupten per se keinen Vorteil für den Patienten; offenkundig müssen Autoren, die sie benutzen, stillschweigend unterstellen, daß bei Patienten, deren Rezidiv potentiell kurativ reseziert wird, die Lebenserwartung durch den Eingriff verbessert wird oder daß ein Teil von ihnen gar geheilt ist.

Diese Überzeugung ist verbreitet, sie beruht maßgeblich auf den relativ günstigen Überlebensraten, die für Patienten mit potentiell kurativ operierten Lebermetastasen kolorektaler Karzinome beobachtet wurden.

So führten Audisio et al. (1996) ein Follow-up bei 505 Patienten mit kolorektalem Karzinom durch. 141 Rezidive wurden registriert, von denen 32 potentiell kurativ reoperiert werden konnten. Audisio et al. betrachteten 18 Patienten aus dieser Gruppe als „geheilt“, da diese Patienten im Median 93 Monate (Spannweite: 15 bis 147 Monate) nach dem Eingriff noch krankheitsfrei waren.

D'Angelica et al. (1997) analysierten die Überlebenszeiten von 456 Patienten, bei denen im Zeitraum 1985 bis 1991 eine Resektion von Lebermetastasen durchgeführt worden war. 96 dieser Patienten lebten 5 Jahre nach der Operation noch, 71 von ihnen waren zum letzten Follow-up-Termin krankheitsfrei.

Man vergleiche auch die Reviews von Müller et al. (1991) oder Scheele u. Altendorf-Hofmann (1999). Freilich existiert bis heute keine randomisierte Studie, die eine Lebensverlängerung durch potentiell kurative Operation isolierter Rezidive nachweist, und vermutlich dürften ihrer Durchführung auch in Zukunft ethische Bedenken entgegenstehen (Poston, 2001).

Wir wollen annehmen, daß die Therapie tatsächlich die Überlebenszeit verbessert, und werden zeigen, daß die Argumente 3a,b dennoch nicht stichhaltig sind.

Für Argument 3a ist dies leicht ersichtlich. Dieses Argument bezieht sich auf ein und dieselbe Studie, die Entdeckung durch die Nachsorgediagnostik und durch Symptome sind hier keine unabhängigen Ereignisse. Letztlich kann man aus dem in 3a beschriebenen Sachverhalt lediglich auf eine *Vorverlegung* der Entdeckung resezierbarer Rezidive schließen, über das Ausmaß der Vorverlegung und die Frage, ob es sich überhaupt auf den Erfolg der Therapie auswirkt, sind keine eindeutigen Schlüsse möglich.

Wie schwach das Argument 3a, selbst mit der Zusatzannahme der Heilbarkeit resezierbarer Rückfälle ist, wird an der Kontroverse um die bereits erwähnte Studie von Moertel et al. (1993) deutlich. So schrieben Cohen u. Paty. (1994) in einem Kommentar: "There are approximately 150 000 new colorectal cancer patients per year, patients with early cancer and synchronous metastases are excluded, and so there are about 100 000 patients to be followed up. One half of these cancers will recur (50 000); one third of these will have only liver disease (17 000); one third of these will have resectable hepatic metastases (5500); and one third of these will be cured (1850)." Moertel (1994) antwortete hierauf, daß vielleicht einige dieser Patienten geheilt werden können, daß sie ja aber nicht notwendig *wegen* des CEA-Monitorings geheilt würden. Das Monitoring könnte sogar schädlich sein, weil sich Patienten mit falsch-negativen Befunden in falscher Sicherheit wiegen.

Argument 3b ist stärker und, wenn man unterstellt, daß ein Teil der entdeckten resezierbaren Rezidive geheilt werden kann, sogar scheinbar zwingend. Dieser Anschein täuscht jedoch. Einen Vorteil durch die Nachsorge impliziert 3b nicht einmal dann, wenn es sich bei den vergleichenden Studien, auf die sich das Argument 3b bezieht, um randomisierte Studien handelt.

Zwei Einwände sind zu erheben. Zum einen würde der behauptete Nutzen in einer Untergruppe der Patienten nur dann einen Nutzen für die *Gesamtheit* der Patienten in der Nachsorge implizieren, wenn die sensitivere Diagnostik nicht - infolge eines höheren Anteils falsch-positiver Befunde - einen kompensatorischen *Schaden* (z.B. durch unnötige Resektionen) verursacht.

Zum anderen ist aber selbst ein Nutzen in Patientenuntergruppen unsicher; aus dem Argument und den Zusatzannahmen folgt er jedenfalls nicht logisch. Die Gegenargumentation ist recht subtil. Um sie leichter verständlich zu machen, wollen wir annehmen, daß die Vergleiche, um

die es in 3b geht, mit einem Minimalprogramm, bestehend aus klinischer Untersuchung und Anamnese, vorgenommen werden.

Generell ist eine potentiell kurative Resektion eines Rezidivs nur möglich, wenn es sich tatsächlich um einen *isolierten* Rückfall handelt (im Sinne des in Kapitel 2 definierten Außenkriteriums Z), wenn also keine okkulten Mikrometastasen vorhanden sind. In solchen Fällen ist die Heilungschance auch dann noch recht groß, wenn das Rezidiv weiterwächst, bis es symptomatisch wird. Dies bedeutet aber: Zwar werden in der Gruppe mit intensiverer Nachsorge mehr resezierbare Rezidive auffällig als in der Vergleichsgruppe, jedoch muß es sich bei der Differenz der Anteile zur Vergleichsgruppe zum größten Teil um solche Fälle handeln, die in Wahrheit zum Zeitpunkt der Entdeckung bereits okkulte Metastasen gebildet haben; denn es sind dies ja Fälle, die in der Vergleichsgruppe nicht auftraten, oder – mit anderen Worten: die, wenn sie Symptome ausbilden, nicht als resezierbar eingestuft werden. Solche Fälle sind aber nicht heilbar. Die Differenz der Anteile potentiell kurativ resezierbarer Rezidive impliziert also nicht zwingend eine höhere Heilungsrate.

Angemerkt sei, daß die Argumente des Typs 3a,b nicht stringenter werden, wenn man sie durch Angaben zu Überlebensraten der potentiell kurativ resezierten Rezidive ergänzt. Diese Argumentationsfigur ist, namentlich beim kolorektalen Karzinom, keineswegs selten; mitunter verbirgt sie sich hinter Definitionen oder Berechnungen:

Beispiele

Anthony et al. (2000) definieren den „estimated benefit“ eines Nachsorgeprogramms als Produkt aus den Faktoren Rezidivrate, Resektabilitätsrate und 5-Jahres-Überlebensrate nach Resektion.

Bei der Empfehlung der ASCO (1996) für CEA-Messungen in der Nachsorge des kolorektalen Karzinoms spielte eine Rolle, daß die durch CEA frühzeitiger entdeckten resezierbaren Rezidive eine 50-prozentige 2-Jahres-Überlebensrate aufweisen.

Staub et al. (2000) definieren „efficacy of follow-up“ als „number of patients with asymptomatic tumor recurrence, solely detected by the follow-up program and who

survived after surgical treatment for recurrence, divided by all patients in the follow-up program.“ Diese Patienten wurden (nach einer medianen Follow-up-Dauer von 24 Monaten nach der Reoperation) als geheilt betrachtet (!).

Rosen et al. (1998) untersuchten als einen ihrer Wirksamkeitsparameter die Fünf-Jahres-Überlebensrate nach potentiell kurativer Resektion.

In der Kosten-Nutzen-Analyse von Norum u. Olsen (1997) basierte die Nutzenberechnung auf der Unterstellung, daß 2% der Patienten durch die Nachsorge nebst Therapie geheilt werden. Dieser Wert wurde geschätzt auf Basis der Studie von Camunas et al. (1991), in der von 151 Patienten in der Nachsorge drei nach einem Relapse mit kurativer Zielsetzung nachoperiert wurden und 40,43 bzw. 69 später noch lebten (!).

Wir haben aufgezeigt, daß die übliche Argumentation zugunsten der Nachsorgediagnostik nicht stichhaltig ist. Fletcher (1993) kritisiert sie in seinem Editorial zur erwähnten Studie von Moertel et al. (1993) treffend als „wishful thinking and reliance on simple arguments (it is better to find recurrence sooner; resection of apparently localized recurrence will cure the patient“). Die Diskussion lehrt aufs Neue, daß Plausibilitätsbetrachtungen grundsätzlich ungeeignet sind, um die Wirksamkeit einer medizinischen Maßnahme zu belegen.

3.3 Zum Design methodisch fundierter Studien

Die Wirksamkeit einer diagnostischen Maßnahme muß in *vergleichenden* Studien gezeigt werden. Solche Studien zählen zur sogenannten Phase IV der Evaluierung einer diagnostischen Maßnahme (Köbberling et al., 1991) und haben methodisch denselben Charakter wie Studien zur Wirksamkeit von Therapien. Zu vergleichen ist in der Studie eine Gruppe von Patienten, bei denen die Maßnahme angewandt wird, mit einer Gruppe, bei der sie nicht angewandt wird.

Hier wie dort kommt der Vergleichbarkeit der Gruppen eine eminente Bedeutung zu. Insbesondere gilt das für die Strukturgleichheit, also die Balancierung in Bezug auf die prognostische Ausgangslage. Eine Unbalance in prognostischen Faktoren kann in einer Vielzahl von Ursachen begründet sein (für eine ausführliche Diskussion s. Abel u. Windeler, 1998).

Das einzige existierende Verfahren, um eine Balance in bezug auf *alle*, bekannte wie unbekannt oder nicht erhobene, prognostische Faktoren zu erreichen, ist die *Randomisation*, die zufällige Zuteilung der Patienten zu den Vergleichsgruppen. Sie gewährleistet, daß die Gruppen in sämtlichen initialen Charakteristika nur noch zufällig bedingte Unterschiede aufweisen. Die Randomisation gilt nicht zu unrecht als goldener Standard der Therapieevaluierung. Sie bietet nicht nur erkenntnistheoretische, sondern auch pragmatische, „metawissenschaftliche“ Vorteile, indem sie z.B. die Verteidigung von Studienergebnissen erleichtert (Beweislastumkehr!) und den Impact der Studien erhöht (Abel u. Koch, 1999). Folgerichtig haben die meisten in der Evaluationsforschung tätigen Mediziner heute den Leitspruch verinnerlicht, daß eine Studie immer randomisiert werden muß, wenn dies möglich ist. Die Cochrane-Collaboration hat in ihren Meta-Analysen bisher stets nur randomisierte Studien berücksichtigt.

Nichtrandomisierte Studien zur Tumornachsorge können aus naheliegenden Gründen schwerwiegende Verzerrungen aufweisen. Dies liegt daran, daß die Entscheidung über die Art der Nachsorge möglicherweise nicht unabhängig vom Stadium ist, sicherlich auch nicht von der Institution und vom Arzt. Die Qualität und der Erfolg der Primärbehandlung ist aber von Institution zu Institution und von Arzt zu Arzt verschieden. Auch Vergleiche von Patienten, die die Kontrolluntersuchungen sorgsam wahrnehmen, mit solchen, die dies nicht tun oder die Nachsorge gar ablehnen (ein solcher Vergleich findet sich z.B. bei Eckardt et al., 1994), können nicht als unverzerrt angesehen werden, denn man kann nicht einfach unterstellen, daß die „Compliance“ von den prognostischen Faktoren der Patienten, der Begleitbehandlung oder anderen wichtigen Determinanten statistisch unabhängig ist. Daß solche Vergleiche unzulässig sind, lehren die Richtlinien der Arzneimittelzulassung, in denen eine Gegenüberstellung der Ergebnisse bei „compliant“ und „noncompliant“ Patienten nicht als Wirksamkeitsnachweis akzeptiert wird.

In historisch kontrollierten Studien kommen zusätzliche Verzerrungen zum Tragen, unter denen dieses Studiendesign im Bereich der Onkologie generell leidet. Zu denken ist etwa, um nur einige Punkte zu nennen, die Verbesserung der Früherkennung oder der Primärdiagnose, an Stadienwanderung oder Fortschritte in der Qualität der Primärbehandlung (Operation).

Im Zusammenhang mit diagnostischen Maßnahmen in der Tumornachsorge besitzen daher publikationsbasierte Meta-Analysen wie die von Bruinvels et al. (1994), in die lediglich nichtrandomisierte Studien einfließen, oder die von Rosen et al. (1998), in der eine Zusammenfassung von randomisierten mit nichtrandomisierten Studien vorgenommen wurde, wenig Aussagekraft. Die Ergebnisse leiden nicht nur unter der Verzerrung, die die einzelnen nichtrandomisierten Studien aufweisen, sondern auch darunter, daß in der Gesamtheit der nichtrandomisierten Studien mit einem höheren Publikationsbias zu rechnen ist. Beim Publikationsbias handelt es sich um eine Auswahlverzerrung des veröffentlichten Materials, die dadurch zustandekommt, daß Studien mit "auffälligen" Ergebnissen (z.B. mit signifikanten Unterschieden zwischen den Vergleichsgruppen) aus vielerlei Gründen – und beinahe unvermeidlich (Abel u. Windeler 1995) – mit höherer Wahrscheinlichkeit publiziert werden als solche mit unauffälligen Ergebnissen. Die Existenz dieses Phänomens wurde in mehreren Untersuchungen belegt. Berlin et al. (1989) stellten fest, daß sich der Selektionseffekt bei nichtrandomisierten Studien stärker bemerkbar macht als bei randomisierten und bei kleineren Studien stärker als bei größeren.

Bestätigt wird das soeben Besprochene durch eine unlängst erschienene Analyse der publizierten Studien zur Nachsorge des kolorektalen Karzinoms (Kievit, 2000). In dieser Analyse ergab sich, daß die Studienqualität - und speziell die Stringenz des Studiendesigns - negativ mit dem Urteil der Autoren über die Effektivität der Nachsorge korreliert war.

Man beachte allerdings auch, daß wegen der Tendenz der Verzerrungen, in *positiver* Richtung zu wirken (Abel u. Windeler, 1995), *Null*resultate aus umfangreichen nichtrandomisierten Studie durchaus beträchtliches Gewicht besitzen. Diese Asymmetrie der Aussagekraft wird oft nicht richtig gewürdigt.

Einschränkend sei hinzugefügt, daß auch die Randomisation nicht die Verzerrungsfreiheit von Studien garantiert. Die Randomisation schafft Vergleichbarkeit lediglich hinsichtlich der Basischarakteristika der Patienten, nicht jedoch in bezug auf alle Faktoren, die nach der Aufnahme in eine Studie das Ergebnis beeinflussen können, wie beispielsweise die Qualität der Durchführung der Prüftherapie, die ärztliche Zuwendung, die Begleitbehandlung, das häusliche Umfeld oder die Art der Ergebnisfeststellung und Beobachtung.

In jüngster Zeit wird zunehmend Kritik an den publizierten randomisierten Studien zur Tumornachsorge geübt. Der Einwand - er wurde gegen die Studien beim kolorektalen Karzinom geäußert - lautet, daß ja nur eine kleine Untergruppe von der Nachsorge profitieren könne, nämlich Patienten, bei denen ein resezierbares Rezidiv entdeckt wird. Studien, die an der Gesamtheit der in die Nachsorge einbezogenen Patienten durchgeführt werden, seien bei weitem zu klein, d.h. sie hätten zu geringe Macht, um günstige Effekte in dieser Untergruppe nachzuweisen (z.B. Renehan u. O'Dwyer, 1998; Anthony et al., 2000; MacKenzie u. O'Dwyer, 2000; Cocconi, 2001). Diese Auffassung beruht auf einem Mißverständnis über die Rolle klinischer Studien.

Studien sollen die klinische Realität möglichst getreu widerspiegeln. Wenn eine Maßnahme in einer Population angewandt wird, dann ist der Nutzen in der *Gesamtheit* der zu dieser Population gehörenden Patienten zu zeigen, und die Bringschuld, einen relevanten Gesamtnutzen nachzuweisen, liegt überdies beim Verfechter der Maßnahme. Wenn die Maßnahme zwar in einer (nicht a priori bestimmbar) Untergruppe einen Vorteil zeigt, nicht jedoch in der gesamten Anwendungspopulation, dann ist sie auch nicht von Nutzen. Der eventuelle Vorteil in der Untergruppe wird dann entweder durch einen Schaden bei den übrigen Patienten kompensiert, oder er ist so klein, daß er, bezogen auf die Gesamtheit der Patienten, lediglich zu einem auch in großen Studien nicht erkennbaren, klinisch irrelevanten Vorteil führt.

Oder anders ausgedrückt: Da die Untergruppe nicht vorab bestimmbar ist, ist der *erwartete* Nutzen bei jedem Patienten, der der Maßnahme unterzogen wird, ausschlaggebend. Dieser ergibt sich aber aus dem Effekt in der Gesamtpopulation.

Das hier widerlegte Argument ist ganz analog zur Forderung, Studien zu Lebensverlängerung einer systemischen Therapie bei fortgeschrittenen Krebserkrankungen mögen so angelegt werden, daß sie einen Überlebenszeitgewinn bei den Respondern nachweisen können. Auch dies ist fehl am Platz. Der Vorteil ist in der Gesamtheit der behandelten Patienten zu zeigen.

Audisio u. Robertson (2000) haben auf ein weiteres Problem hingewiesen, das ihrer Meinung nach alle randomisierten Vergleiche unterschiedlich intensiver Nachsorgediagnostik-Programme aufweisen. (Das Problem betrifft freilich strenggenommen alle vergleichenden Studien, es hat mit der Zufallszuweisung nichts zu tun.) Bei einigen Patienten in der Gruppe

mit minimalem Follow-up werden diagnostische Maßnahmen eingesetzt werden, die zum intensivem Follow-up zählen; umgekehrt werden in der Gruppe mit intensivem Follow-up nicht alle Patienten die Nachsorge regelmäßig wahrnehmen. Audision u. Robertson bezeichnen dies als Kontamination, und merken an, daß als Folge hiervon eine Verzerrung in Richtung Nullergebnis zu erwarten ist.

Das Problem ist allerdings aus Therapiestudien wohlbekannt. Zu entgegnen ist auch hier, daß Studien einem pragmatischen Ansatz folgend die *Realität* wiedergeben sollen. Studienobjekt sind nicht hypothetische Patienten, die den beiden Regimes exakt folgen; vielmehr werden Patientengruppen verglichen, bei denen die *intendierten* Schemata der Nachsorge verschieden sind, wobei die tatsächliche *praktische Durchführung* der klinischen Realität entspricht, inklusive aller Imperfektionen und „Kontaminationen“. Vorteile für ein Regime sind in diesem realen Kontext zu zeigen.

Im folgenden sollen für die besonders gut untersuchten und wichtigen Beispiele des Mammakarzinoms und des kolorektalen Karzinoms Ergebnisse aus randomisierten Studien dargestellt werden.

4. „Effizienz“ und Nutzen von Nachsorgeprogrammen – randomisierte Studien

4.1 Mammakarzinom

Beim Mammakarzinom ist die Situation, was die in 3.1 aufgeführten notwendigen Bedingungen für einen Nutzen der Nachsorge betrifft, mit Blick auf Ergebnisse von Therapiestudien nicht als sonderlich günstig einzustufen. Ganz besonders trifft dies für die Bedingungen 4 und 5 zu.

Die intensive diagnostische Nachsorge beim Mammakarzinom zielt in der Hauptsache darauf ab, frühzeitig Fernmetastasen zu entdecken. Das metastasierte Mammakarzinom gilt aber nach wie vor als unheilbar, jedwede Behandlung hat primär palliativen Charakter (Muss, 2002). Ob eine systemische Therapie die Lebenserwartung im metastasierten Stadium verlängern kann, ist unklar, da diese Behandlungen niemals in direkten Vergleichen mit unbehandelten Kontrollen geprüft wurden.

Manche renommierte Onkologen bestreiten eine deutliche Lebensverlängerung. So heißt es in den Empfehlungen des Tumorzentrums München: „Eine Verlängerung der Gesamtüberlebenszeit durch die Therapie ist im metastasierten Krankheitsstadium bisher nicht eindeutig zu belegen“ (Sauer, 2001). In der Tat wird diese skeptische Sicht durch mehrere Indizien gestützt: In randomisierten Studien zeigte sich kein Überlebenszeitvorteil einer Chemotherapie im Vergleich zu einer Hormontherapie; aggressivere Chemotherapien scheinen insgesamt, d.h. bei zusammenfassender Würdigung der zahlreichen zu dieser Frage durchgeführten Studien, hinsichtlich der Lebensdauer nicht besser abzuschneiden als weniger aggressive; und Responder auf eine anfängliche Chemotherapie profitieren anscheinend nicht davon, daß man nach Eintritt der Response die Therapie fortsetzt¹.

Erst recht gibt es keine randomisierten Therapiestudien, die nahelegen, daß die Bedingung 5 erfüllt ist, d.h. Studien, in denen eine sofortige Therapie einen Überlebenszeitvorteil im Vergleich zu einer später begonnenen Therapie („wait-and-see“) gezeitigt hat.

Die ersten und bisher einzigen randomisierten Studien zum Wert einer intensiven Nachsorge beim Mammakarzinom wurden im Jahr 1994 veröffentlicht. Es handelt sich um zwei hervorragend geplante, umfangreiche Untersuchungen, die beide fast zeitgleich, aber unabhängig voneinander, von italienischen Studiengruppen unter Beteiligung einer großen Zahl von Zentren durchgeführt wurden (Rosselli del Turco et al, 1994; The Givio Investigators, 1994).

Der Aufbau der beiden Studien war sehr ähnlich. Generell wurden nur potentiell kurativ operierte Patientinnen aufgenommen. Nach der Operation wurden die Patientinnen per Zufallszug entweder der Gruppe mit intensiver Nachsorge-Diagnostik zugeteilt oder der Vergleichsgruppe, in der man sich mit klinischen Untersuchungen nebst jährlichen Mammographien begnügte. Die intensive Diagnostik umfaßte über die klinischen Untersuchungen und die Mammographien hinaus regelmäßige Röntgenaufnahmen des Brustkorbs und szintigraphische Untersuchungen der Knochen, in der Studie der Givio Investigators zusätzlich noch Ultraschalluntersuchungen der Leber sowie Blutuntersuchungen. *Tabelle 1a,b* gibt einen Überblick über das Design und die Hauptergebnisse der beiden Studien. Die Schicksalskontrolle war in der Studie von Rosselli del Turco bis 5 Jahre nach Randomisation nahezu vollständig (die Information über den Vitalstatus fehlte nur für 5

¹ Eine Übersicht über Studien zu diesen Fragestellungen ist beim Erstautor erhältlich.

Patienten); in der GIVIO-Studie ist lediglich vermerkt, daß 123 Patientinnen (9.8%) vorzeitig die Nachsorge abbrachen oder vor dem Relapse „lost-to-follow-up“ waren. Diese Patienten wurden mit den verfügbaren Daten in der Auswertung berücksichtigt.

Tabelle 1a

Die randomisierte Studie der GIVIO-Studiengruppe (1994)

Design: Multizentrische, randomisierte Studie, 26 teilnehmende Allgemeinkrankenhäuser in Italien. Randomisation der Patientinnen innerhalb von 6 Wochen nach Operation

Patienten: Histologisch gesichertes, operables, unilaterales, nichtinflammatorisches Mammakarzinom, Stadium T1-T3, N0-N1, M0
Alter bis 70 Jahre

Interventionen (Vergleichsgruppen):

1. Intensives Follow-up-Programm (n=655), Dauer: 5 Jahre
 - alle 3 Monate (nach 2 Jahren alle 6 Monate): körperliche Untersuchung
 - alle 6 Monate: Röntgenthorax
 - alle 12 Monate: Knochenszintigramm, Ultraschall der Leber, Mammographie der kontralateralen Mamma
 - bei jedem Nachsorgetermin: Laborparameter (z.B. alk. Phosphatase)
2. Kontrollgruppe (minimales Follow-up-Programm (n=665), Dauer: 5 Jahre
 - alle 3 Monate (nach 2 Jahren alle 6 Monate): körperliche Untersuchung
 - alle 12 Monate: Mammographie der kontralateralen Mamma
 - zusätzliche Diagnostik nur, wenn klinisch indiziert

Zielgrößen:

- Überlebenszeit
- DFI
- lead time
- Lebensqualität (7 gesundheitsbezogene Aspekte, die anhand von Selbstausskünften anhand 4 verschiedener Fragebögen ermittelt wurden); Erhebungen 6,12,24 und 60 Monate nach der Primäroperation

Ergebnisse:

1. Der Prozentsatz der Metastasen, die im asymptomatischen Stadium entdeckt wurden, war in Gruppe 1 höher (31% versus 21%).
2. Die Zeit bis zur Entdeckung der Metastasen/Rezidive unterschied sich nur wenig (im Mittel weniger als 1 Monat).
3. Die Lebensdauerverteilungen in Gruppe 1 und 2 waren nahezu identisch.
4. Die Lebensqualitäts-Scores in Gruppe 1 und 2 waren zu jedem Erhebungszeitpunkt beinahe identisch.

Tabelle 1b

Die randomisierte Studie von Rosselli Del Turco et al. (1994)

Design: Multizentrische, randomisierte Studie, 12 teilnehmende onkologische Zentren („breast clinics“) in Italien. Randomisation der Patientinnen innerhalb von 6 Monaten nach Operation

Patienten: Histologisch gesichertes, operables, unilaterales, invasives Mammakarzinom ohne Evidenz für Metastasen
Alter bis 70 Jahre

Interventionen (Vergleichsgruppen):

1. Intensives Follow-up-Programm (n=622), Dauer: 5 Jahre
 - alle 3 Monate (nach 2 Jahren alle 6 Monate): körperliche Untersuchung
 - alle 6 Monate: Röntgenthorax, Knochenszintigramm
 - alle 12 Monate: Mammographie
2. Kontrollgruppe: minimales Follow-up-Programm (n=621), Dauer: 5 Jahre
 - alle 3 Monate (nach 2 Jahren alle 6 Monate): körperliche Untersuchung
 - alle 12 Monate: Mammographie
 - zusätzliche Diagnostik nur, wenn klinisch indiziert

Zielgrößen: - Überlebenszeit
- DFI, DFS
- lead time

Ergebnisse:

1. Die Zeit bis zur Entdeckung der Metastasen/Rezidive (DFI) war in Gruppe 2 signifikant länger ($p=0.01$), das DFS (bei dem auch nicht nur Metastasen/Rezidive, sondern auch Todesfälle als Ereignis gezählt werden) war grenzwertig signifikant länger ($p=0.07$).
2. Die Lebensdauerverteilungen in Gruppe 1 und 2 waren nahezu ununterscheidbar.

Als Fazit stellte man in beiden Untersuchungen fest, daß es zwar mit Hilfe der intensiveren Diagnostik tatsächlich gelang, Metastasen oder lokale Rückfälle im Krankheitsverlauf etwas früher zu diagnostizieren (wobei Edelman et al. (1998) freilich anmerken, daß nur 10% der Rückfälle ausschließlich durch Labortests oder radiologische Methoden entdeckt wurden); jedoch waren die Überlebenskurven der beiden Vergleichsgruppen nahezu ununterscheidbar.

Bekräftigt wurde dieses Ergebnis noch einmal durch später mitgeteilte Langzeitergebnisse der Studie von Rosselli del Turco (Palli et al., 1999): Erneut ergaben die nunmehr auf einer

zehnjährigen Schicksalskontrolle beruhenden Analysen praktisch identische Überlebenskurven.

Unlängst wurde von der Cochrane Collaboration ein Review zur Frage der Follow-up-Strategien beim Brustkrebs herausgegeben. Dieser Review, der mittlerweile mehrfach aktualisiert wurde (neueste Version: Rojas et al., 2002), enthält unter anderem eine originaldatenbasierte Meta-Analyse der beiden in Tabelle 1 aufgeführten Studien. Aus der Zusammenfassung beider Studien ergab sich für die weniger intensive Überwachung verglichen mit der intensiveren ein nahezu gleiches Mortalitätsrisiko (Hazard Ratio = 0.96, 95%-Konfidenzintervall: 0.8 bis 1.15) und ein grenzwertig signifikant längeres krankheitsfreies Überleben (Hazard Ratio = 0.84, 95%-Konfidenzintervall: 0.71 bis 1.00).

Die Untersuchung zur Lebensqualität, die zu den Hauptfragestellungen der GIVIO-Studie zählte, war kaum weniger ernüchternd als die zur Überlebensdauer. Die Patientinnen waren gebeten worden, jeweils 6, 12, 24 und 60 Monate nach Aufnahme in die Studie vier Fragebögen auszufüllen, in denen unterschiedliche Aspekte der Lebensqualität erhoben wurden, soweit diese vom Gesundheitszustand abhingen. Im einzelnen handelte es sich um folgende Instrumente:

- Functional Living Index-Cancer Scale (68 Items)
- Sickness Impact Profile (74 Items)
- Profile of Mood States (104 Items)
- Cancer Inventory of Problem Situation (75 Items).

In der Studie interessierte man sich für 7 Lebensqualitäts-Konstrukte: „quality-of life perception“, „overall health perception“, „body image“, „emotional well-being“, „social functioning“, „symptoms“, „satisfaction with care“. Um Scores für diese 7 Aspekte zu ermitteln, wurden auf der Basis psychometrischer Analysen aus den vier Fragebögen 18 Items ausgewählt, die anschließend den 7 Konstrukten zugeordnet wurden.

Die Datenanalyse ergab, daß zu allen vier Zeitpunkten die mittleren Bewertungen der 7 gemessenen Aspekte der Lebensqualität in den beiden Vergleichsgruppen nahezu gleich hoch waren, wobei minimale Nachteile sogar eher auf Seiten der Gruppe mit der intensiven

Überwachung lagen. Die Studiengruppe kommentiert dies in ihrer Veröffentlichung so: "...one of the arguments frequently used to support an intensive surveillance protocol is its ability to reassure patients. Our results do not support this view."

Wenngleich die Ergebnisse der randomisierten Studien für manche Onkologen vermutlich überraschend waren, wurden sie nicht nur zur Kenntnis genommen, sondern breit und in aller Offenheit diskutiert und in mehreren Zentren beinahe umgehend umgesetzt. Die Geschwindigkeit dieses Vorgangs ist – auch aus wissenschaftshistorischer Sicht – um so bemerkenswerter, als die Resultate langjährige Lehrmeinungen in Frage stellten und zweifellos auch gewisse Interessen berührten.

Die praktische Konsequenz, die z.B. manche Tumorzentren noch im Jahr der Publikation aus den Studie zogen (vgl. z.B. Schmid u. Sauer, 1994), war die, daß sie aus ihrem Nachsorgeprogramm für Brustkrebspatientinnen alle technischen und laborchemischen Routineuntersuchungen herausnahmen.

Dabei spielte gewiß auch die Einsicht eine Rolle, daß die Nachsorge, wenn sie nicht wirksam ist, bei manchen Patienten durchaus Schaden anrichten kann. So kommentierte der Münchner Mediziner H.S. Fießl (1994) die Ergebnisse der Studien mit den Worten: "Offensichtlich gibt es nicht einmal eine rein medizinische Begründung für ein Intensiv-Überwachungsprogramm. Nicht zu reden von der psychischen Belastung, der die Frauen ausgesetzt sind: Die häufigen Arztbesuche, das Warten auf die Ergebnisse, die laufende Erinnerung daran, eine potentiell tödliche Krankheit zu haben und die psychische Achterbahn falsch-positiver Befunde mit der Konsequenz weiterer Untersuchungen und möglicher Übertherapie."

Nicht weniger kritisch äußern sich manche andere Onkologen, so z.B. Gallmeier u. Keding (1994): "Häufig leiden Kranke nur an den für sie schwer zu interpretierenden Ergebnissen überflüssiger Untersuchungen oder an den Erwartungsängsten schon Tage vor den Nachsorgeterminen." Kleeberg (1994) bemerkt²: "Die bittere Erfahrung ist aber, daß eine noch so engmaschige Nachsorge am Schicksal der Patienten nichts ändert. Eine frühzeitige Aufdeckung einer Metastase bedeutet nicht eine Verlängerung der Lebenszeit, sondern eine Verlängerung der Leidenszeit."

² Behandlung von Krebspatienten in der Praxis. Interview mit Prof. Dr. med U.R. Kleeberg, Hamburg. Cancer Care Nr. 1, S. 10-11 (1994)

Eine Reihe neuerer Empfehlungen für Onkologen - so z.B. seitens der amerikanischen Gesellschaft für klinische Onkologie (American Society of Clinical Oncology, 1997), der Deutschen Krebshilfe oder manchen Landesärztekammern (vgl. die entsprechenden Meldungen in der Ärzte-Zeitung am 9. 2. 1998 und am 14. 4. 1999) - tragen diesen Einsichten Rechnung und empfehlen jetzt für Brustkrebspatientinnen nach der operativen Entfernung des Tumors ein Nachsorgekonzept, das sich auf eine Anamnese, eine körperliche Untersuchung sowie Mammographien beschränkt und eine intensive apparative Diagnostik nur dann vorsieht, wenn Befunde oder Symptome des Patienten dies nahelegen.

Natürlich gibt es - wie immer in der Wissenschaft, wenn neue Ergebnisse etablierte Praktiken in Frage stellen - auch Widerstand, oder aber die Umsetzung erfolgt trotz grundsätzlicher Akzeptanz nur mit erheblicher Verzögerung. Die Arbeitsgruppe um Rosselli del Turco (Ciatto et al, 1998) stellte im Nachgang zu ihrer Aufsehen erregenden Studie eine systematische Erhebung der Nachsorgemodalitäten beim Brustkrebs in 138 klinischen Zentren Europas an, von denen 55 außerhalb Italiens lagen. 87% der Zentren führten eine Routine-Nachsorge mit mindestens einem diagnostischen Test auf stumme Fernmetastasen durch, jedoch gaben 36% der Einrichtungen zu Protokoll, daß diese Kontrollen ihrer Einschätzung nach nichts an der Prognose der Patientinnen änderte, und 40% gingen davon aus, daß sie keinen Einfluß auf die Lebensqualität hatten.

Ungeachtet dessen läßt sich anhand des zugänglichen Schrifttums nicht erkennen, daß die durchaus verbreiteten Verteidigungs- und Abwehrstrategien der Medizin gegen neue, unbequeme Erkenntnisse (Windeler, 1992) im Falle der Nachsorge des Mammakarzinoms in breitem Maßstab angewandt wurden.

4.2 Kolorektales Karzinom

Die zweite Krebsform, bei der man den Wert der Nachsorge inzwischen gründlich untersucht hat, ist das kolorektales Karzinom.

Bei dieser Krebslokalisierung sind die Aussichten einer Tumornachsorge, gemessen an den Ergebnissen aus Therapiestudien, a priori günstiger einzuschätzen als beim Mammakarzinom. In der Literatur finden sich mehrere randomisierte Studien, die im fortgeschrittenen Stadium insgesamt einen Überlebenszeitvorteil durch zytostatische Chemotherapie belegen (Colorectal Cancer Collaborative Group, 2000). Auch gibt es aus mindestens einer randomisierten Studie

Hinweise darauf, daß eine sofortige Chemotherapie einem abwartenden Vorgehen überlegen ist (The Nordic Gastrointestinal Tumor Adjuvant Therapy Group, 1992).

Hinzu kommt, daß zumindest gewisse Hinweise darauf existieren (s.o.), daß in der Untergruppe der Patienten mit resezierbaren Rezidiven die Zweitoperation zur Verbesserung der Prognose beiträgt. Nochmals sei allerdings darauf hingewiesen, daß diese Erfolge selbst wenn sie real sind, einen Nutzen der Nachsorge bei den betroffenen Patienten zwar nahelegen, aber keineswegs zwingend implizieren. Stattdessen gilt umgekehrt: Wenn die Effizienz der Tumornachsorge ausschließlich an der Behandelbarkeit isolierter Rezidive festgemacht wird, dann geben die Resultate der Follow-up-Studien eher zur Skepsis Anlaß. Dies wollen wir an Beispielen verdeutlichen.

Safi u. Beger (1991) analysierten Daten, die in Zeitraum 1978 bis 1989 im aufwendigen Nachsorgeprogramm der Abteilung für Allgemeine Chirurgie der Universitätsklinik Ulm angefallen waren. Im Kollektiv von 1054 Patienten wurde während einer medianen Beobachtungszeit von 38 Monaten bei 350 Patienten ein Tumorrezidiv festgestellt (75 lokoregionäre Rezidive, 275 Fernmetastasen). In 56 Fällen erfolgte eine Zweitoperation mit kurativer Zielsetzung. 21 Patienten waren am Ende der Beobachtungszeit tumorfrei. Safi und Beger bemerken hierzu: „Aufwand und Ergebnis des zur Zeit praktizierten Tumornachsorgeprogramms beim kolorektalen Karzinom stehen in einem Mißverhältnis“.

Böhm et al. (1993) werteten die Nachsorgedaten von 487 Patienten aus, bei denen im Zeitraum 1980 bis Ende 1989 eine R₀-Resektion eines kolorektalen Karzinoms vorgenommen worden war. Bei 139 Patienten war ein Tumorrezidiv dokumentiert, davon in 46 Fällen ausschließlich ein Lokalrezidiv, in 76 Fällen ausschließlich eine Fernmetastasierung. Bei 36 Patienten konnte eine Resektion des Rezidivs mit kurativer Zielsetzung vorgenommen werden, nur 9 Patienten aus dieser Gruppe waren zwei Jahre später noch am Leben.

Wolf u. Cohen (1997) standen für ihre Analysen, die sich primär mit dem Wert des CEA befaßten, Nachsorgedaten von 1899 Patienten zur Verfügung, die in der prospektiven Colorectal Cancer Database des Memorial Sloan Kettering Cancer Centers gespeichert waren. 272 Rezidive waren unter diesen Patienten dokumentiert, in 50 Fällen wurde eine

potentiell kurative Zweitoperation unternommen. 26 dieser Patienten waren beim letzten Follow-up noch krankheitsfrei. Dies entspricht 1.3% der Gesamtkohorte. Nur die Hälfte dieser Fälle war überdies durch CEA-Erhöhungen angezeigt worden.

Eine Reihe weiterer Untersuchungen sind zu ähnlichen Ergebnissen gelangt (s. z.B. die von Gallmeier u. Keding (1994) oder Schumacher et al. (1995) mitgeteilten Daten).

Hervorzuheben ist die bereits mehrfach erwähnte umfangreiche Studie von Moertel et al. (1993), in der es um die Bedeutung des CEA in der Nachsorge ging.

Tabelle 2 gibt einen knappen Überblick über das Design und die wichtigsten Ergebnisse dieser Untersuchung. Die Studie zeichnet sich dadurch aus, daß sie einen Vergleich ermöglichte, der allerdings nicht auf einer Zufallszuteilung beruhte. Moertel et al. stellten fest, daß das CEA in der Diagnostik von Krankheitsrückfällen einen nicht unbeträchtlichen Anteil falsch-positiver und falsch-negativer Befunde produzierte (1-Sensitivität=41%, 1-Spezifität=16%; man beachte allerdings die in Kapitel 2 diskutierten methodischen Probleme, denen auch diese Schätzwerte unterliegen). Der zusätzliche Anteil der Patienten, bei denen das CEA-Monitoring zu einer Entdeckung isolierter Metastasen führte, die man potentiell kurativ operieren konnte und die ein Jahr später noch krankheitsfrei waren, betrug lediglich 0.3% aller Patienten ($23/1017 = 2.3\%$ in der Gruppe mit CEA-Monitoring versus $4/200 = 2\%$ in der Gruppe ohne Monitoring).

Tabelle 2

Die Studie von Moertel et al. (1993)

Design: Retrolektive multizentrische Kohortenstudie, eingebettet in die randomisierte Studie zum Vergleich von 5-FU + Levamisol versus Beobachtung in der adjuvanten Therapie des Kolonkarzinoms (Moertel et al, 1990); elektive postoperative CEA-Messungen; mediane Follow-up-Dauer: 6 Jahre (Spannweite: 5-8.5 Jahre)

Patienten: Histologisch gesichertes, operables, Kolonkarzinom, Dukes B2,C; n=1217

Messungen, Auswertungen:

Postoperative CEA-Messungen, soweit durchgeführt; CEA-Messungen nicht standardisiert. Primär wurde ein Wert von CEA > 5ng/mL als „positiver“ Befund angesehen.

CEA-Messungen zum oder nach dem Zeitpunkt der Feststellung eines Rezidivs blieben unberücksichtigt.

Vergleich von der Patienten mit CEA-Monitoring (Gruppe 1, n=1017) mit solchen ohne Monitoring (Gruppe 2, n=200).

- Zielgrößen:**
- Sensitivität und Spezifität des CEA-Monitorings in Gruppe 1
 - Anteil von Patienten mit potentiell kurativer Resektion eines Rezidivs, in Abhängigkeit vom CEA-Wert und aufgeschlüsselt nach Lokalisation des Rezidivs
 - Überlebenszeit bei diesen Patienten

Ergebnisse:

1. Gruppe 1: Bei 247 (59%) von 417 Patienten mit Krankheitsrückfall wurde zuvor ein positiver CEA-Befund festgestellt; bei 98 (16%) von 600 Patienten ohne Rückfall wurden ebenfalls erhöhte CEA-Werte gemessen.
2. Der Vergleich von Gruppe 1 und 2 ergab folgendes Resultat:

	CEA-Monitoring		kein Monitoring
	CEA-Erhöhung	keine Erhöhung	
Anzahl Patienten	345	672	200
potentiell kurative Resektion des Rez.	47	38	23
davon krankheitsfrei nach > 1 Jahr	10	13	4

In einem Editorial zur Studie von Moertel et al. wies Fletcher (1993) darauf hin, daß im Licht dieser Resultate die CEA-Bestimmung in der Nachsorge zwangsläufig eine sehr ungünstige Kosten-Nutzen-Relation aufweisen müsse. Er schreibt: „The financial costs are considerable, almost \$500 000 per possible cure according to the simple calculations by Moertel and colleagues, and perhaps much higher. This is far most costly than most clinical care considered in our society to be worth paying for it. But it is not just a matter of money. There is also the inconvenience and discomfort of repeated blood tests, worry over what the result might be, extra surgery, and knowing about recurrence of cancer earlier than one would ordinarily, even though many of these patients cannot be cured. I am skeptical that the benefits outweigh the costs, all things considered.“

Bis 1994 lagen zur Frage des Nutzens der Nachsorge beim kolorektalen Karzinom lediglich nichtrandomisierte Studien vor. Erwähnenswert ist die Matched-Pair-Studie von Pichlmaier (1994), in der 183 Patienten, die seit 1977 an einem Nachsorgeprogramm teilgenommen hatten, 183 Merkmalszwillingen aus derselben Klinik gegenübergestellt wurden, bei denen keine standardisierte Nachsorge durchgeführt worden war. Die Überlebenskurven waren über einen Zeitraum von 5 Jahren nahezu identisch. Auch die stratifizierte Analyse für die Stadien Dukes B und C ergab keinen Hinweis auf Vorteile der Nachsorgegruppe. Zwar fehlen in der Publikation die Details der Studie, doch kommt dem Resultat, da es sich um ein Nullergebnis handelt, aus den oben besprochenen Gründen beträchtliches Gewicht zu.

Im Zeitraum 1995-99 wurden fünf randomisierte Studien veröffentlicht, in denen ein intensives diagnostisches Monitoring mit dem Verzicht hierauf oder einem weniger intensiven verglichen wurde (Mäkelä et al., 1995; Ohlsson et al., 1995; Kjeldsen et al., 1997; Pietra et al., 1998; Schoemaker et al., 1998). Erwähnenswert ist eine weitere, in Italien durchgeführte randomisierte Studie (Barillari et al., 1996) an 607 Patienten, in der es um den Vergleich zweier Nachsorgestrategien ging, bei denen Koloskopien jährlich bzw. im 24-monatigen Abstand durchgeführt wurden. Diese Studie enthält jedoch keinen Vergleich der Gruppen in Bezug auf Wirksamkeitsparameter, insbesondere keinen Vergleich von Überlebenskurven.

In Tabelle 3a-e sind die wesentlichen Merkmale des Designs und die Hauptresultate der publizierten Studien aufgeführt. Die intensivere Diagnostik war nicht in allen Studien gleich geartet, doch umfaßte sie neben den üblichen klinischen Untersuchungen typischerweise regelmäßige Koloskopien, CT oder sonographische Untersuchungen der Leber, Röntgenthorax sowie diverse Laboruntersuchungen, zu denen insbesondere die Bestimmung des Tumormarkers CEA zählte.

Tabelle 3

Veröffentlichte randomisierte Studien zur Tumornachsorge beim kolorektalen Karzinom

a) Mäkelä et al. (1995)

Patienten:

106 Patienten, Dukes A-C, Primäroperation am Oulu University Hospital, Aufnahmezeitraum: 1988-1990

Vergleichsgruppen

Gruppe 1 (konventionelles Follow-up, n=54): Kontrolluntersuchungen 3,6,9,12,15,18, 21,24,30,36,42,48,54,60 Monate nach Op; zu jedem Termin: Anamnese, klinische Untersuchung, Blutbild und Leukozyten-Gesamtzahl, CEA, Hämokult-Test, Röntgenthorax, Sigmoidoskopie (nur bei Patienten mit Rektumkarzinom); jährlich: Barium-Kontrastdarstellung

Gruppe 2 (intensiviertes Follow-up, n=52): die gleichen Untersuchungstermine wie Gruppe 1; keine Barium-Kontrastdarstellungen; zusätzliche Untersuchungen: 3 Monate postoperativ sowie anschließend jährlich Koloskopie, alle 6 Monate Ultraschall der Leber, jährlich CT der Leber und des Operationsorts.

Ergebnisse:

In Gruppe 1 wurden 21 Rückfälle beobachtet, in Gruppe 2 22. In Gruppe 1 waren 13 Rückfälle asymptomatisch, in Gruppe 2 16. Mit kurativer Zielsetzung konnten 3 versus 5 Rückfälle reoperiert werden. Das krankheitsfreie Überleben in der intensiv überwachten Gruppe war signifikant kürzer (im Mittel 5 Monate, $p=0.002$). Die Gesamtüberlebenskurven unterschieden sich nicht erkennbar und überschneiden sich mehrfach ($p=0.5$).

b) Ohlsson et al. (1995)

Patienten:

107 Patienten, Dukes A-C, Primäroperation in chirurgischen Abteilungen in Lund und Helsingborg, Aufnahmezeitraum: 1983-1986

Vergleichsgruppen

Gruppe 1 (intensives Follow-up, n=53): Kontrolluntersuchungen 3,6,9,12,15,18,21,24, 30,36,42,48,60 Monate nach Op; zu jedem Termin: Anamnese, klinische Untersuchung, Laborwerte, inklusive CEA, Hämokult-Test, Röntgenthorax, Sigmoidoskopie (nur bei

Patienten mit Rektumkarzinom); 9,21,42 Monate postoperativ endoskopische Untersuchung der Anastomose; 3,15,30,60 Monate postoperativ Koloskopie; 3,6,12,18,24 Monate CT der Hüfte (bei Patienten mit abdominoperinealer Resektion)

Gruppe 2 (keine routinemäßige Nachsorge, n=54)

Ergebnisse:

In Gruppe 1 wurden 17 Rückfälle beobachtet, in Gruppe 2 18. In Gruppe 1 waren 12 Rückfälle asymptomatisch, in Gruppe 2 3. Eine potentiell kurative Reoperation war in 4 versus 3 Rückfällen möglich. Das DFS in der intensiv überwachten Gruppe war nichtsignifikant kürzer (mediane Intervalle bis Relapse: 1.7 vs. 2 Monate). Die Gesamtüberlebenskurven waren in Gruppe 1 etwas günstiger (5-Jahres-Überlebensraten 78 vs. 71%), allerdings war der Unterschied nicht signifikant. Bei Patienten mit Relapse zeigten die Überlebenskurven von Gruppe 1 und 2 keinen Unterschied.

c) Kjeldsen et al. (1997,1999)

Patienten:

597 Patienten unter 76 Jahren, Dukes A-C, Einwohner des Bezirks Funen, Aufnahmezeitraum: 1983-1994

Vergleichsgruppen

Gruppe 1 (intensives Monitoring, n=290): Kontrolluntersuchungen 6,12,18,30,36, 48,60,120,150 180 Monate nach der Operation des Primärtumors

Gruppe 2 (Kontrolle, n=307): Untersuchungen 60, 120, 180 Monate nach der Operation

Die Untersuchungen waren ansonsten in beiden Gruppen identisch: Anamnese, klinische Untersuchung, digitale rektale Untersuchung, ggf. gynäkologische Untersuchung, Hämokkult-Test, Koloskopie, Röntgenthorax, Blutbild und Leberenzyme.

Ergebnisse:

In Gruppe 1 wurden 76 Rückfälle beobachtet, in Gruppe 2 80. In Gruppe 1 waren 38 Rückfälle asymptomatisch, in Gruppe 2 11 ($p=0.02$). Mit kurativer Zielsetzung konnten 9 versus 0 Rückfälle reoperiert werden ($p<0.01$). Das krankheitsfreie Überleben in der intensiv überwachten Gruppe war nichtsignifikant kürzer (im Mittel 9 Monate). Die Überlebensdauer nach dem Rückfall war in Gruppe 1 signifikant kürzer ($p<0.01$). Die Gesamtüberlebenskurven unterschieden sich nicht erkennbar und überschneiden sich mehrfach ($p=0.9$). Die gesundheitsbezogene Lebensqualität der Patienten, die 1996 noch lebten (n=350), war in Gruppe 1 nur marginal besser (s. die Zusatzuntersuchung von Kjeldsen et al., 1999).

d) Pietra et al. (1998)**Patienten:**

207 Patienten, Dukes B/C, Primäroperation in der Abteilung für Allgemeinchirurgie der Universitätsklinik Parma, Aufnahmezeitraum: 1987-1990

Vergleichsgruppen

Gruppe 1 (konventionelles Follow-up, n=103): Kontrolluntersuchungen 6,12 Monate nach Op, dann jährlich; zu jedem Termin: Anamnese, klinische Untersuchung, CEA, Ultraschall der Leber; jährlich: Röntgenthorax, Koloskopie

Gruppe 2 (intensiviertes Follow-up, n=104): Kontrolluntersuchungen zunächst in halben Abständen der Gruppe 1, nach 5 Jahren jährlich; zu jedem Termin: Anamnese, klinische Untersuchung, CEA, Röntgenthorax; jährlich: CT der Leber, Koloskopie; der Schwerpunkt der Studie lag auf der Entdeckung von Lokalrezidiven.

Ergebnisse:

In beiden Gruppen wurden je 41 Rezidive beobachtet. In Gruppe 1 wurden 20 Lokalrezidive festgestellt, davon 9 bei Rektumkarzinompatienten, in Gruppe 2 26, davon 14 bei Rektumkarzinompatienten. Acht Lokalrezidive in Gruppe 1 und 2 in Gruppe 2 wurden aufgrund von Symptomen diagnostiziert. Eine potentiell kurative Reoperation war bei 2 versus 17 Lokalrezidiven möglich. Das Intervall bis zur Diagnose eines Lokalrezidivs betrug im Mittel 10.3 versus 20.2 Monate. Gruppe 1 hatte ein signifikant kürzeres Überleben, mit 5-Jahres Überlebensraten von 58.3% versus 73.1%, $p < 0.02$.

Bei Patienten mit Lokalrezidiv war der Unterschied zwischen Gruppe 1 und 2 besonders ausgeprägt (5-Jahres-Überlebensraten: 0 versus 38.4%, $p < 0.01$).

Nach Ansicht der Autoren scheinen vor allem Patienten mit Krebs des Rektums von der häufigeren Kontrolle zu profitieren.

e) Schoemaker et al. (1998)**Patienten:**

325 Patienten <85 Jahre, Dukes A-C, Primäroperation am einem von zwei teilnehmenden südaustralischen Zentren

Vergleichsgruppen

Gruppe 1 (konventionelles Follow-up, n=158): Kontrolluntersuchungen 3,6,9,12,15,18,21,24,30,36,42,48,54,60 Monate nach Op; zu jedem Termin: klinische Untersuchungen, Blutbild mit Leukozytenzahl, Leberfunktionstests, CEA, Hämokultttest.

Gruppe 2 (intensiviertes Follow-up, n=167): gleiches zeitliches Nachsorgeschema wie Gruppe 1; zusätzliche Untersuchungen: jährlicher Röntgenthorax, Leber-CT, Koloskopie

Ergebnisse:

Die Arbeit schlüsselt die entdeckten Rezidive/metachronen Primärtumoren auf nach der diagnostischen Methode, die zur Entdeckung führte. Beispiel: Die Zahl der entdeckten Lebermetastasen betrug 23 (Gruppe 1) versus 20 (Gruppe 2). In Gruppe 2 waren 12 dieser Metastasen asymptomatisch, in Gruppe 1 Null ($p=0.00001$). Die Gesamtüberlebenskurven unterschieden sich nicht signifikant, mit 5-Jahres-Überlebensraten von rund 70% vs. 76% ($p=0.1986$).

Nicht aufgeführt ist eine randomisierte Studie der Arbeitsgruppe um John Northover. Von dieser Untersuchung, in der es ausschließlich um die Bedeutung von CEA-Messungen in der Nachsorge ging, existiert keine Originalpublikation, Details wurden lediglich auf Kongressen und in persönlichen Mitteilungen berichtet und können teilweise Kurzberichten oder der Sekundärliteratur entnommen werden (z.B. Lennon et al., 1994; Millat u. Borie, 1998; Bedenne u. Jouve, 1999). Demnach war die Studie eingebettet in eine Kohortenstudie an 1447 Patienten, die im Zeitraum 1982 bis 1993 wegen eines primären kolorektalen Karzinoms operiert wurden. CEA-Messungen wurden 3 Jahre lang monatlich, weitere zwei Jahre in 3-monatigen Abständen durchgeführt. Die 213 Patienten, bei denen eine auffällige CEA-Erhöhung beobachtet wurde, teilte man zufällig in zwei Gruppen ein: In der einen Gruppe wurden die Ärzte über die CEA-Erhöhung informiert, in der zweiten Gruppe unterblieb dies. Geprüft wurde somit die Arbeitshypothese, daß die Kenntnis eines erhöhten CEA-Werts eine erfolgreiche Intervention - beispielsweise eine erneute Operation - ermöglicht, die zu einer Lebensverlängerung führt. Wie man sich leicht klarmacht, läuft dies de facto ebenfalls auf einen Vergleich unterschiedlich intensiven diagnostischen Monitorings hinaus.

Die Rekrutierung für die Studie wurde auf Empfehlung des Data Monitoring Sub-Committee abgebrochen. Zwar wurden in der „informierten“ Gruppe häufiger Reinterventionen durchgeführt (65% vs. 28%), auch solche mit kurativer Zielsetzung (16% vs. 8%), jedoch ergab sich nach den 1994 auf einem Kongreß mitgeteilten Zwischenergebnissen für die über die Testergebnisse informierte Gruppe sogar eine etwas *niedrigere* 5-Jahres-Überlebensrate (20%) als für die Kontrollgruppe (22%), mit einem Hazard Ratio von 0.84 (95%-Konfidenzintervall: 0.62-1.13).

Die Resultate der in Tabelle 3a-e dargestellten Studien sind nicht ganz einheitlich, doch fällt auf, daß sich in vier der fünf Studien kein statistisch signifikanter Vorteil in der

Überlebensdauer für die Gruppe mit intensiver Überwachung ergab. Eine Ausnahme bildet die Studie von Pietra et al. (1998), die sich speziell auf die Entdeckung von Lokalrezidiven konzentrierte.

Mit Blick auf die Erörterungen in Kapitel 2 und 3 ist der Befund der Studien von Kjeldsen et al. (1997) von besonderem Interesse, daß zwar die Entdeckung der Rezidive durch intensiviertes Monitoring vorverlegt wird, daß dies jedoch durch eine kürzere Restüberlebenszeit nach dem Rezidiv erkaufte und kompensiert wird.

Die Ergebnisse der randomisierten Studien wurden weithin als enttäuschend empfunden und führten zu einem Überdenken der üblichen aufwendigen Nachsorgeprogramme. So kommentierten Smith u. Bear (1998) in einem Editorial der Zeitschrift "Gastroenterology" die negativen Resultate der australischen Studiengruppe mit den Worten: "It is time to stop doing these extra tests and make sure that our patients actually receive good, simple follow-up care." Und McArdle (2000) zog in seiner Übersichtsarbeit das Fazit „Clearly, follow up as currently practised is ineffective“. Auch in Deutschland gingen zahlreiche kritische Kommentare durch die Presse (vgl. z.B. Ostendorf, 1999).

Die dänische Gruppe nahm eine Schätzung der Kosten vor, die mit dem häufigeren Follow-up in ihrer Studie verbunden war (Kjeldsen et al., 1999). Danach beliefen sich diese Kosten auf über 8 Millionen dänische Kronen (in Kursen von 1999 umgerechnet mehr als 1.3 Millionen US-Dollar), wobei die indirekten Kosten, wie z.B. Transporte, sogar unberücksichtigt blieben. Kronborg (2000) merkte zu diesen Berechnungen an: „...a small benefit from intensive follow-up (1-2% increase in long-term survival) will be at least four times as expensive as that obtained by screening persons of average risk for colorectal cancer, when measured as cost per life year without colorectal cancer.“

In zwei publikationsbasierten Meta-Analysen wurden die Ergebnisse der randomisierten Studien mathematisch-statistisch zusammengefaßt.

Die Meta-Analyse von Renehan u. O'Dwyer (2000), die nur in Form eines Leserbriefes publiziert wurde, berücksichtigte neben den in Tab. 3a-e aufgeführten Untersuchungen auch die oben beschriebene Studie von Northover. Für die Zielgröße „Überleben“ im Vergleich „intensives versus weniger intensives diagnostisches Monitoring“ errechnete sich ein

geschätztes summarisches relatives Risiko (Odds ratio) von 0.79 ($p=0.04$), mit einem 95%-Konfidenzintervall von 0.63 bis 0.99. Bei Ausschluß der Studie von Northover fiel das relative Risiko auf 0.76, 95%-Konfidenzintervall: (0.59-0.97). Es wurde keine statistisch signifikante Heterogenität der Ergebnisse der sechs Studien festgestellt ($p>0.1$).

Unlängst wurde von der Arbeitsgruppe um A.G. Renehan eine ausführliche Fassung ihrer Meta-Analyse, beschränkt auf die fünf publizierten randomisierten Studien, vorgelegt (Renehan et al., 2002). Das Ergebnis ist ähnlich, aber nicht gleichlautend zu dem im Jahr 2000 veröffentlichten (Zielgröße Überlebensdauer: OR=0.81, 95%-Konfidenzintervall von 0.70 bis 0.94). Darüber hinaus finden sich in dieser Arbeit Odds Ratios des Vergleichs „intensiviertes Follow-up versus Kontrollen“ für verschiedene Subtypen von Tumorrezidiven. Die Werte – errechnet nach zwei biometrischen Methoden, dem „fixed effects model“ und dem „random effects model“ - lauten wie folgt (Angaben in Klammern = 95% Konfidenzintervalle):

	OR (fixed effects)	OR (random effects)
alle Rezidive	0.95 (0.82-1.11)	0.95 (0.81-1.11)
Lokalrezidive	1.17 (0.91-1.52)	1.19 (0.91-1.54)
isolierte Lokalrezidive	1.62 (1.12-2.32)	1.59 (1.10-2.30)
Lebermetastasen	0.78 (0.59-1.05)	0.77 (0.77-1.03)
isolierte Lebermetastasen	1.13 (0.61-2.08)	1.02 (0.51-2.02)
Lungenmetastasen	0.61 (0.35-1.05)	0.62 (0.35-1.08)
intraluminale Rezidive	1.25 (0.64-2.44)	1.26 (0.63-2.51)
metachrone Primärtumoren	1.50 (0.63-3.54)	1.55 (0.28-3.63)

Die Rezidive wurden in den Gruppen mit intensiverem Follow-up im Mittel 8.5 Monate (95%-Konfidenzintervall: 7.6-9.4 Monate) früher entdeckt als in den Kontrollgruppen. Die Autoren merken kritisch an, daß in den Publikationen der Studien zahlreiche wichtige Angaben fehlen und daß insbesondere in keiner Studie erkennbare Vorkehrungen getroffen waren, um die Ärzte in bezug auf Patientenzuteilung oder Ergebnisfeststellung zu verblinden.

Zu ganz ähnlichen Ergebnisse gelangte die im Rahmen eines Cochrane Reviews vorgenommene Meta-Analyse von Jeffery et al. (2002). Berücksichtigt wurden nur die fünf

publizierten Studien. Für die Zielgröße „Überleben“ ergab sich ein Odds Ratio von 0.67, 95%-Konfidenzintervall: 0.53-0.84. Es wurde keine signifikante Heterogenität gefunden.

Eine Risikoreduktion ergab sich insbesondere für die Vergleiche „mehr versus weniger diagnostische Tests“ und „bildgebende Untersuchungen der Leber versus keine solchen Untersuchungen“. Diese beiden Vergleiche stützen sich auf dieselben drei Studien (Mäkela et al, Ohlsson et al, Schoemaker et al.) und resultierten deshalb in identischen Schätzungen des relativen Risikos (0.66, 95%-Konfidenzintervall: 0.46-0.95).

Die Autoren des Reviews ziehen folgendes Fazit: „The results of our review suggest that there is an overall survival benefit for intensifying the follow-up of patients after curative surgery for colorectal cancer. Because of the wide variation in the follow-up programmes used in the included studies it is not possible to infer from the data the best combination and frequency of clinic (or family practice) visits, blood tests, endoscopic procedures and radiological investigations to maximise the outcomes for these patients. Nor is it possible to estimate the potential harms or costs of intensifying follow-up for these patients in order to adopt a cost-effective approach in this clinical area.“

Möglicherweise ist bei der Interpretation der Resultate der Meta-Analysen noch mehr Zurückhaltung angebracht. Wie Anthony et al. (2000) anmerkten, leiden die Studien unter der Heterogenität sowohl der Interventionen als auch der Ergebnisse (man beachte, daß die Heterogenität der Ergebnisse in der Meta-Analyse wegen der geringen Zahl der berücksichtigten Studien in der entsprechenden statistischen Testung nur schwer auffällig werden kann). Dies läßt eine Zusammenfassung der Studien fragwürdig erscheinen.

Aus dem Rahmen fällt insbesondere die Studie von Pietra et al. (1998), die von Kievit (2000) als „outlying study“ bezeichnet wird. In dieser Studie erklärt sich – dies ergibt ein überschlägiges Nachrechnen - der Überlebenszeitunterschied im wesentlichen durch den Vorteil in der Untergruppe von Patienten mit Lokalrezidiven. Auch wenn man unterstellt, daß dieser Vorteil real ist, rechtfertigt er doch nicht die besonders aufwendigen diagnostischen Maßnahmen, die zur Entdeckung von Fernmetastasen dienen und eingesetzt werden.

Bei Jeffrey et al. (2002) finden sich Hinweise auf zwei neue umfangreiche, multizentrische randomisierte Studien, eine italienische und eine britische. Die italienische Studie (GILDA Group unter Leitung von R. Fossati) wurde 1998 aktiviert und soll mindestens 2920 Patienten

des Stadiums Dukes B/C rekrutieren. In die britische Studie, in der es unter anderem um den Wert von CEA sowie von CT und Ultraschall im Follow-up geht, sollen 4890 Patienten aufgenommen werden.

Mit der Größe der Studien will man – dies wird an der Stellungnahme von Fossati (2001) deutlich - offenbar sicher stellen, daß die Studien eine ausreichende Macht zur Entdeckung von Vorteilen in der Untergruppe von Patienten mit potentiell kurativ resezierbaren Rezidiven aufweisen. Man beachte aber, daß der in Kapitel 3 gemachte Einwand gegen die Interpretation auch bei einem positivem Resultat der Studien gültig bleibt.

5. Zusammenfassung

Der vorliegende Report befaßt sich in zwei Teilen mit Fragen der Validierung und Wirksamkeitsbestimmung diagnostischer Maßnahmen in der Tumornachsorge.

1. Zwar gibt es in der Literatur eine Fülle von Follow-up-Studien, die die Bestimmung der Sensitivität, Spezifität oder der Vorhersagewerte von Test- oder Untersuchungsverfahren in der Tumornachsorge zum Gegenstand haben, doch wird in diesen Arbeiten den methodischen Aspekten der Validierung nur geringe Beachtung geschenkt. Die publizierten Schätzwerte sind durchweg schlecht fundiert.

Die folgenden Probleme werden im Report erörtert:

a) Die Validität ist kontextabhängig; sie hängt ab von Charakteristika der Population, der Unsicherheit der Befundung selbst, der Einbettung in den diagnostischen Prozeß und vom Außenkriterium; im Falle quantitativer Tests ist sie zudem eine Funktion des Positivitätskriteriums (z.B. des Schwellenwerts).

b) Die veröffentlichten Sensitivitätsschätzungen leiden generell nicht nur unter statistischer Unsicherheit (sie wird in den Studien im allgemeinen nicht quantifiziert), sondern auch unter mehreren Verzerrungen. An erster Stelle ist ein ausgeprägter Verifikationsbias zu nennen, der dadurch zustande kommt, daß positive und negative Testbefunde in unterschiedlicher Weise abgeklärt werden (es werden de facto zwei Außenkriterien benutzt); zweitens spielt Informationsbias eine Rolle, da die Bestimmungen des Testbefunds und des Außenkriteriums nicht gegeneinander verblindet sind. Bei quantitativen Tests (Tumormarkern) ist in Studien, in denen das Positivitätskriterium erst anhand der Daten ermittelt wird, mit „Resubstitutionsbias“ zu rechnen.

c) Die klassische Definition der Sensitivität ist im Zusammenhang mit der Nachsorgediagnostik unbefriedigend; Versuche, die wünschenswerte „Vorverlegung“ definitorisch zu berücksichtigen sind andererseits nicht mit einer inhaltlich sinnvollen Definition der Vorhersagewerte vereinbar.

2. Der zweite Teil des Reports diskutiert die Frage der Wirksamkeit des diagnostischen Monitorings in bezug auf die Lebenserwartung der Patienten. Fünf notwendige Bedingungen müssen erfüllt sein, damit eine derartige Wirkung möglich ist. Unter Berücksichtigung dieser Gesichtspunkte werden die üblichen Standardargumente für das Monitoring einer methodischen Kritik unterzogen. Es zeigt sich, daß keines dieser Argumente haltbar ist.

In einer Erörterung von Aspekten des Designs klinischer Studien zur Wirksamkeit wird herausgestellt, daß nichtrandomisierte Studien notwendigerweise unter einer Reihe spezifischer Verzerrungen leiden. Ferner wird die in der neueren Literatur mit Bezug auf das kolorektale Karzinom aufgestellte Forderung widerlegt, vergleichende Studien intensiverer mit weniger intensiver Nachsoragediagnostik mögen so angelegt sein, daß sie einen Überlebenszeitvorteil in der Untergruppe der Patienten entdecken können, bei denen ein resezierbares Rezidiv auftritt.

In den beiden letzten Abschnitten wird ein Überblick über die publizierten randomisierten Studien und Meta-Analysen zur Wirksamkeit des postoperativen Monitorings beim Mammakarzinom und kolorektalen Karzinom gegeben.

6. Literatur

Abel, U., Berger, J.(1986): Comparison of resubstitution, data splitting, the Bootstrap, and the Jackknife as methods for estimating the validity indices of new marker tests. *Biometrical J.* 28, 899-908

Abel, U., Quentmeier, A., Herfarth, Ch. (1986): Monitoring plasma CEA in the follow-up of colorectal cancer: the validation of a class of diagnostic criteria for early detection of local recurrence and metastases. *Theoret. Surgery* 1, 136-142

Abel, U.(1993): *Die Bewertung diagnostischer Test.* Hippokrates-Verlag, Stuttgart

Abel, U., Windeler, J. (1995): Erkenntnistheoretische Aspekte klinischer Studien. 1. Irrtümer in der Bewertung medizinischer Therapien - Ursachen und Konsequenzen. *Internistische Praxis* 35, 613-629

Abel, U., Windeler, J. (1996): Erkenntnistheoretische Aspekte klinischer Studien. 2. Die Wirksamkeit medizinischer Therapien. *Internistische Praxis* 36, 375-90

Abel, U., Windeler, J. (1998): Erkenntnistheoretische Aspekte klinischer Studien. 4. Vergleichbarkeit in klinischen Studien. A. "Strukturgleichheit" und die Bedeutung der Randomisation. *Internistische Praxis* 38, 613-24

Abel, U., Koch, A. (1999): The role of randomization in clinical studies: myths and beliefs. *J. Clin. Epidemiol.* 52, 487-97

American Society for Gastrointestinal Endoscopy (2000): Guidelines for colorectal cancer screening and surveillance. *Gastrointestinal Endoscopy* 51, 777-782

American Society of Clinical Oncology (1996): Clinical practice guidelines for the use of tumor markers in breast and colorectal cancer. *J. Clin. Oncology* 14, 2843-2877

American Society of Clinical Oncology (1997): Recommended breast cancer surveillance guidelines. *J. Clin. Oncol.* 15, 2149-2156

American Society of Clinical Oncology (1998): Update of recommendations for the use of tumor markers in breast and colorectal cancer. *J. Clin. Oncology* 16, 793-795

Anonym (1997): Krebsnachsorge: Abschied von Illusionen. *Münch. Med. Wschr.* 139, 22-23

Anthony, Th., Fleming, J.B., Cieligk, S.C., Sarosi, G.A., Kim, L.T., Gregorcyk, S.G., Simmang, C.L., Turnage, R.H. (2000): Postoperative colorectal cancer surveillance. *J. Am. Coll. Surg.* 190, 737-749

Audisio, R., Setti-Carraro, P., Segala, M., Capko, D., Andreoni, B., Tiberio, G. (1996): Follow-up in colorectal cancer patients: a cost-benefit analysis. *Annals of Surgical Oncology* 3, 349-357

Audisio, R.A., Robertson, C. (2000): Colorectal cancer follow-up: perspectives for future studies. *Europ. J. Surgical Oncol.* 26, 329-337

- Barillari, P., Ramacciato, G., Manetti, G., Bovino, A., Sammartino, P. (1996): Surveillance of colorectal cancer. *Dis. Colon Rectum* 38, 388-393
- Bedenne, L., Jouve, J.L. (1999): Surveillance des cancer colorectaux après exérèse. *La Presse Médicale* 28, 651-656
- Berlin, J.A., Begg, C.B., Louis, T.A. (1989): An assessment of publication bias using a sample of published clinical trials. *JAMA* 84, 381-392
- Berman, J.M., Cheung, R.J., Weinberg, D.S. (2000): Surveillance after colorectal cancer resection. *Lancet* 355, 395-399
- Bleeker, W.A., Mulder, N.H., Hermans, J., Otter, R., Plukker, J.T.M. (2001): Value and cost of follow-up after adjuvant treatment of patients with Dukes' C colonic cancer. *Brit. J. Surgery* 88, 101-106
- Boey, J., Cheung, H.C., Lai, C.K., Wong, J. (1984): A prospective evaluation of serum carcinoembryonic antigen (CEA) levels in the management of colorectal carcinoma. *World J. Surgery* 8, 279-286
- Böhm, B., Schwenk, W., Hucke, H.P., Stock, W. (1993): Does methodic long-term follow-up affect survival after curative resection of colorectal carcinoma? *Dis Colon Rectum* 36, 280-286
- Bruinvels, D.J., Stiggelbout, A.M., Kievit, J., van Houwelingen, H.C., Habbema, D.F., van de Velde, C.J.H. (1994): Follow-up of patients with colorectal cancer. A meta-analysis. *Annals of Surgery* 219, 174-182
- Camuñas, J., Enriques, J.M., Devesa, J.M., Morales, V., Millan, I. (1991): Value of follow-up in the management of recurrent colorectal cancer. *Europ. J. Surgical Oncology* 17, 530-535
- Ciatto, S., Muraca, M.C., Rosselli del Turco, M. (1998): Survey of the practice of follow-up for the early detection of distant metastases in breast cancer patients in Europe. *The Breast* 7, 72-74
- Cocconi, G. (2001): Follow-up nel carcinoma coloretale. *Tumori* 87 (Suppl 1), S37-S38
- Cohen, A.M., Paty, P. (1994): CEA for monitoring colon cancer. *Letter. JAMA* 271, 346
- Colorectal Cancer Collaborative Group (2000): Palliative chemotherapy for advanced colorectal cancer: systematic review and meta-analysis. *Brit. Med. J.* 321, 531-535
- Cooper, G.S., Yuan, Z., Chak, A., Rimm, A.A. (1999): Geographic and patient variation among Medicare beneficiaries in the use of follow-up testing after surgery for non-metastatic colorectal carcinoma. *Cancer* 85, 2124-2131
- D'Angelica, M., Brennan, M.F., Fortner, J.G., Cohen, A.M., Blumgart, L.H., Fong, Y. (1997): Ninety-six five-year survivors after liver resection from metastatic colorectal cancer. *J. Am. College of Surgeons* 185, 554-559

- Delbrück, H., Bamberg, M. (2000): La cure est morte! Vive la rehabilitation! Editorial. *Onkologie* 6, 1-2
- Denstman, F., Rosen, L., Khubchandani, I.T., Sheets, J.A., Stasik, J.J., Riether, R.D. (1986): Comparing predictive decision rules in postoperative CEA monitoring. *Cancer* 58, 2089-2095
- Desch, C.E., Benson III, A.B., Smith, T.J., Flynn, P.J., Krause, C., Loprinzi, C.L., Minsky, B.D., Petrelli, N.J., Pfister, D.G., Somerfield, M.R. (1999): Recommended colorectal cancer surveillance guidelines by the American Society of Clinical Oncology. *J. Clin. Oncol.* 17, 1312-1321
- Donnelly, J., Mack, P., Donaldson, L.A. (2001): Follow-up of breast cancer: time for a new approach. *Int. J. Clin. Pract* 55, 431-433
- Eckardt, V.F., Stamm, H., Kanzler, G., Bernhard, G. (1994): Improved survival after colorectal cancer in patients complying with a postoperative endoscopic surveillance program. *Endoscopy* 26, 523-527
- Eckardt, V.F., Bernhard, G. (1997): Nachsorge beim kolorektalen Karzinom. *Dt. Ärzteblatt* 93A, 456-462
- Edelman, M. J., Meyers, F.J., Siegel, D. (1997): The utility of follow-up testing after curative cancer therapy. *J. Gen. Intern. Med.* 12, 318-331
- Edelman, M. J., Siegel, D., Meyers, F.J. (1998): Follow-up testing after cancer treatment: is it indicated. *Western J. Med.* 168, 528-529
- Fletcher, R.H. (1986): Carcinoembryonic antigen. *Ann. Internal Medicine* 104, 66-73
- Fletcher, R.H. (1993): CEA monitoring after surgery for colorectal cancer. When is the evidence sufficient? Editorial. *JAMA* 270, 987-988
- Fossati, R. (2001): Il follow-up del carcinoma coloretale operato radicalmente. *Tumori* 87 (Suppl. 1), S35-S36
- Füeßl, H.S. (1994): Welche Nachsorge bei Mammakarzinom? *Münch. Med. Wschr.* 136/Nr 26, S. 16
- Gallmeier, W.M., Keding, G. (1994): Nachsorge bei Krebspatienten. *Münch. Med. Wschr.* 136, 610-616
- Goldberg, R., Fleming, T.R., Tangen, C.M., Moertel, C.G., Macdonald, J.S., Haller, D.G., Laurie, J.A. (1998): Surgery for recurrent colon cancer: strategies for identifying resectable recurrence and success rates after resection. *Ann. Internal Medicine* 129, 27-35
- Grunfeld, E., Fitzpatrick, R., Mant, D., Yudkin, P., Adewuyi-Dalton, R., Stewart, J., Cole, D., Vessey, M. (1999): Comparison of breast cancer patient satisfaction with follow-up in primary care versus specialist care: results from a randomized controlled trial. *Brit. J. of General Practice* 49, 705-710

Gulliford, T., Opumu, M., Wilson, E., Hanham, I., Epstein, R. (1997): Popularity of less frequent follow up for breast cancer in randomised study: initial findings from the hotline study. *BMJ* 314, 174-177

Jeffery, G.M., Hickey, B.E., Hider, P. (2002): Follow-up strategies for patients treated for non-metastatic colorectal cancer (Cochrane Review). In: *The Cochrane Library*, Issue 1. Oxford: Update Software

Jensen, K., Abel, U. (2000): Methodik diagnostischer Validierungsstudien - Fehler in der Studienplanung und Auswertung. *Med. Klinik* 95, Sondernummer, 54-60

Kievit, J., Bruinvels, D.J. (1995): Detection of recurrence after surgery for colorectal cancer. *Europ. J. Cancer* 31A, 1222-1225

Kievit, J. (2000): Colorectal cancer follow-up: a reassessment of empirical evidence on effectiveness. *Europ. J. Surgical Oncology* 26, 322-328

Kjeldsen, B.J., Kronborg, O., Fenger, C., Jørgensen, O.D. (1997): A prospective randomized study of follow-up after radical surgery for colorectal cancer. *Br. J. Surgery* 84, 666-669

Kjeldsen, B.J., Thorsen, H., Whalley, D., Kronborg, O. (1999): Influence of follow-up on health-related quality of life after radical surgery for colorectal cancer. *Scand. J. Gastroenterol.* 34, 509-515

Köbberling, J., Richter, K., Trampisch, H.-J., Windeler, J. (1991): *Methodologie der medizinischen Diagnostik*. Springer, Berlin – Heidelberg – New York

Kronborg, O. (2000): Colorectal cancer follow-up. Editorial. *Europ. J. Surgical Oncol.* 26, 321

Langhammer, H.R., Ellgas, W., Laubenbacher, Chr., Erbas, B., Busch, R., Riesinger, U., Senekowitsch, R. (1994): Ca 15-3 und CEA in der Rezidivdiagnostik des Mammakarzinoms unter Berücksichtigung von Lokalisation und Ausmaß der Metastasierung. *TumorDiagnostik u. Therapie* 15, 96-103

Lavin, P.T., Day, J., Holyoke, D., Mittelman, A., Chu, T.M. (1981): A statistical evaluation of baseline and follow-up carcinoembryonic antigen in patients with resectable colorectal carcinoma. *Cancer* 47, 823-826

Lennon, T., Houghton, J., Northover, J. (1994): Post-operative CEA monitoring and second-look surgery in colorectal cancer: trial results. Abstract no 30. *Br. J. Cancer* 22 (Suppl.), 16

Liberati, A., Torri, V., Apolone, G. (1997): Assessing the effectiveness of follow-up care for colorectal cancer: a great conceptual and methodological challenge for clinical oncology. Editorial. *Annals of Oncology* 8, 1059-1062

Loprinzi, C.L. (1994): It is now the age to define the appropriate follow-up of primary breast cancer patients. Editorial. *J. Clin. Oncol.* 12, 881-883

- MacKenzie, S., O'Dwyer, P.J. (2000): Surveillance after colorectal cancer resection. Letter. *Lancet* 355, 1095
- Mäkelä, J.T., Laitinen, S.O., Kairaluoma, M.I. (1995): Five-year follow-up after radical surgery for colorectal cancer. *Arch. Surg.* 130, 1062-1067
- Martin, E.W., James, K.K., Hurtubis, P.E., Catalano, P., Minton, J.P. (1977): The use of CEA as an early indicator for gastrointestinal tumor recurrence and second-look procedures. *Cancer* 39, 440-446
- McArdle, C. (2000): ABC of colorectal cancer. Effectiveness of follow-up. *BMJ* 321, 1332-1335
- McCall, J.L., Black, R.B., Rich, C.A., Harvey, J.R., Baker, R.A., Watts, J.M., Touli, J. (1994): The value of serum carcinoembryonic antigen in predicting recurrent disease following curative resection of colorectal cancer. *Dis. Colon Rectum* 37, 875-881
- Millat, B., Borie, F. (1998): Faut-il surveiller les malades opérés d'un cancer colorectal? *J. Chir.* 135, 177-181
- Mille, D., Roy, T., Carrère, M.-O., Ray, I., Ferjaoui, N., Späth, H.-M., Chauvin, F., Philip, T. (2000): Economic impact of harmonizing medical practices: compliance with clinical practice guidelines in the follow-up of breast cancer in a French Comprehensive Cancer Center. *J. Clin. Oncol.* 18, 1718-1724
- Minton, J.P., Hoehn, J.L., Gerber, D.M., Horsley, J.S., Conolly, D.P., Salwan, F., Fletcher, W.S., Cruz Jr., A.B., Gatchell, F.G., Oviedo, M., Meyyer, K.K., Leffall Jr., L.D., Berk, R.S., Stewart, P.A., Kurucz, S.E. (1985): Results of a 400-patient carcinoembryonic antigen second-look colorectal cancer study. *Cancer* 55, 1284-1290
- Moertel, Ch., Fleming, Th.R., Macdonald, J.S., Haller, D.G., Laurie, J.A., Goodman, P.J., Ungerleider, J.S., Emerson, W.A., Tormey, D.G., Glick, J.H. (1990): Levamisole and Fluorouracil for adjuvant therapy of resected colon carcinoma. *New Engl. J. Med.* 322, 352-358
- Moertel, Ch., Fleming, Th., Macdonald, J.S., Haller, D.G., Laurie, J.A., Tangen, C. (1993): An evaluation of the carcinoembryonic antigen (CEA) test for monitoring patients with resected colon cancer. *JAMA* 270, 943-947
- Moertel, Ch. (1994): CEA for monitoring colon cancer. Letter to the Editor. *JAMA* 271, 346-347
- Müller, J.M., Schmidt, A., Strauß, J.M., Zieren, H.U. (1991): Resektion von Lebermetastasen kolorektaler Karzinome. Anspruch und Wirklichkeit. *Dt. Med. Wschr.* 116, 681-688
- Muss, H.B. (2002): Management of metastatic breast cancer: current standards and emerging concept. (Online program certified by accredited CME providers). Transcript of a presentation as delivered by the faculty at the „Optimal Therapy for Primary and Metastatic Breast Cancer: Emergencies Standards and New Approaches“ Symposium, San Antonio, Texas, Dec 13, 2001.

Seton Hall University School of Graduate Medical Education. www.medscape.com/viewprogram/1021_pnt

Northover, J. (1998): Realism or nihilism in bowel cancer follow-up? *The Lancet* 351, 1074-1076

Northover, J. (2000): Which type of follow-up? *Hepato-Gastroenterology* 47, 335-336

Norum, J., Olsen, J.A. (1997): A cost-effectiveness approach to the Norwegian follow-up programme in colorectal cancer. *Annals of Oncology* 8, 1081-1087

O'Dwyer, P.J., Stevenson, J.P., Haller, D.G., Rotman, N., Giantonio, B.J. (2001): Follow-up of stage B and C colorectal cancer in the United States and France. *Seminars in Oncology* 28 (Suppl. 1), 45-49

Ohlsson, B., Breland, U., Ekberg, H., Graffner, H., Tranberg, K.-G. (1995): Follow-up after curative surgery for colorectal carcinoma. *Dis. Colon Rectum* 38, 619-626

Ostendorf, G-M. (1999): Krebs-Nachsorge: Trend zur individuellen Betreuung. <http://arzt-online.de/docs/1999/04/09/065a1501.asp>

Palli, D., Russo, A., Saieva, C., Ciatto, S., Rosselli del Turco, M., Distante, V., Pacini, P. (1999): Intensive vs clinical follow-up after treatment of primary breast cancer: 10 year update of a randomized trial. *JAMA* 281, 1586

Pandya, K.J., McFadden E.T., Kalish, L.A., Tormey, D.C., Taylor, S.G., Falkson, G. (1985): A retrospective study of earliest indicators of recurrence in patients on Eastern Cooperative Oncology Group adjuvant chemotherapy trials for breast cancer. *Cancer* 55, 202-205

Peethambaram, P., Weiss, M., Loprinzi, C.L., Novotny, P., O'Fallon, J.R., Erlichman C., O'Connell, M.J., Laurie, J.A (1997): An evaluation of postoperative follow-up tests in colon cancer patients treated for cure. *Oncology* 54, 287-292

Pichlmaier, H. (1994): Kosten-Nutzen-Überlegungen in der onkologischen Chirurgie. *Münch. Med. Wschr.* 136, 226-230

Pietra, N., Sarli, L., Costi, R., Ouchemi, C., Gratarola, M., Peracchia, A. (1998): Role of follow-up in management of local recurrences of colorectal cancer. *Dis. Colon Rectum* 41, 1127-1133

Poston, G. (2001): The argument for liver resection in colorectal liver metastases. *Hepato-Gastroenterology* 48, 345-346

Reitzig, P. (1995): Tumornachsorge – Nutzen oder sinnloser Aufwand? In: H. Günther, G. Ehninger (Hrsg.): *Individuelle Therapieentscheidungen bei „unbegrenzten Möglichkeiten“*. Pechstein Verlag, Dobersdorf; S. 111-119

Renahan, A., O'Dwyer, S.T. (1998): A selective policy in follow-up for bowel cancer. Letter to the Editor. *Lancet* 351, 1891

- Renahan, A., O'Dwyer, S.T. (2000): Surveillance after colorectal cancer resection. Correspondence. *Lancet* 355, 1095-1096
- Renahan, A.G., Egger, M., Saunders, M.P., O'Dwyer, S.T. (2002): Impact on survival of intensive follow-up after curative resection for colorectal cancer: systematic review and meta-analysis of randomised trials. *BMJ* 324, 1-8
- Richard, C.S., McLeod, R.S. (1997): Follow-up of patients after resection for colorectal cancer: a position paper of the Canadian Society of Surgical Oncology and the Canadian Society of Colon and Rectal Surgeons. *Canadian J. Surg.* 40, 90-100
- Rojas, M.P., Telaro, E., Russo, A., Fossati, R., Palli, D., Rosselli Del Turco, M., Confalonieri, C., Liberati, A. (2002): Follow-up strategies for women treated for early breast cancer. (Cochrane Review): In: *The Cochrane Library*, Issue 1, Oxford: Update Software.
- Rosen, M., Chan, L., Beart, R.W., Vukasin, P., Anthone, G. (1998): Follow-up of colorectal cancer. A meta-analysis. *Dis Colon Rectum* 41, 1116-1126
- Rosselli del Turco, M., Palli, D., Cariddi, A., Ciatto, S., Pacini, P., Distante, V. (1994): Intensive diagnostic follow-up after treatment of primary breast cancer. *JAMA* 271, 1593-1597
- Safi, F., Beger, H.G. (1991): Ist die Tumornachsorge beim kolorektalen Karzinom sinnvoll? *Dt. Med. Wschr.* 116, 1001-1007
- Sakorafas, G.H., Tsiotou, A.G., Pavlakis, G. (2000): Follow-up after primary treatment for breast cancer. *Acta Oncologica* 39, 935-940
- Sauer, J. (Hrsg.) (2001): Empfehlungen zur Diagnostik, Therapie und Nachsorge. Mammakarzinome. 8. überarbeitete Auflage 2001. Tumorzentrum München. <http://www.krebsinfo.de/ki/empfehlung/mamma>
- Scanlon, E.F., Oviedo, M.A., Cunningham, M.P., Caprini, J.A., Khandekar, J.D., Cohen, E., Robinson, B., Stein, E. (1980): Preoperative and follow-up procedures on patients with breast cancer. *Cancer* 46, 977-979
- Schapira, D.V. (1993): Breast cancer surveillance: a cost-effectiveness strategy. *Breast Cancer Res. Treatm.* 25, 107-111
- Scheele, J., Altendorf-Hofmann, A. (1999): Resection of colorectal liver metastases. *Langenbecks Archives of Surgery* 384, 313-327
- Schmid, L., Sauer, H. (1994): Neue Empfehlungen für die Mammakarzinom-Nachsorge. *Münch. Med. Wschr.* 136, 413-414
- Schoemaker, D., Black, R., Giles, L., Toouli, J. (1998): Yearly colonoscopy, liver CT, and chest radiography do not influence 5-year survival of colorectal cancer patients. *Gastroenterology* 114, 7-14

Schumacher, T., Schmitz, S., Diehl, V. (1995): Prognoserelevanz der onkologischen Nachsorge. *Versicherungsmedizin* 47, 166-170

Smith, Th. J., Bear, H.D. (1998): Standard follow-up of colorectal cancer patients: finally we can make practice guidelines based on evidence. *Gastroenterology* 114, 211-213

Staab, H.J., Anderer, F.A., Stumpf, E., Hornung, A., Fischer, R., Kieninger, G. (1985): Eighty-four potential second-look operations based on sequential carcinoembryonic antigen determinations and clinical investigations in patients with recurrent gastrointestinal cancer. *Am. J. Surgery* 149, 198-204

Staib, L., Link, K.H., Beger, H.G. (2000): Follow-up in colorectal cancer: cost-effectiveness analysis of established and novel concepts. *Langenbeck's Arch. Surg.* 385, 412-420

Steele Jr., G., Zamcheck, N., Wilson, R., Mayer, R., Lokich, J., Rau, P., Maltz, J. (1980): Results of CEA-initiated second-look surgery for recurrent colorectal cancer. *Am. J. Surgery* 139, 544-548

Taylor, I. (2000): Quality of follow-up of the cancer patient affecting outcome. *Surgical Oncology Clinics of North America* 9, 21-25

The Givio Investigators (1994): Impact of follow-up testing on survival and health-related quality of life in breast cancer patients. *JAMA* 271, 1587-1592

The Nordic Gastrointestinal Tumor Adjuvant Therapy Group (1992): Expectancy or primary chemotherapy in patients with advanced asymptomatic colorectal carcinoma: a randomized trial. *J. Clin. Oncol.* 7, 1437-1446

Virgo, K., Vernava, A.M., Longo, W.E., McKirgan, L.W., Johnson, F.E. (1995): Cost of patient follow-up after potentially curative colorectal cancer treatment. *JAMA* 273, 1837-1841

Winchester, D.P., Sener, S.F., Khandekar, J.D., Oviedo, M.A., Cunningham, M.P., Caprini, J.A., Burkett, F.E., Scanlon, E.F. (1979): Symptomatology as an indicator of recurrent or metastatic breast cancer. *Cancer* 43, 956-960

Windeler, J. (1992): Argumentationsstrukturen bei der Verteidigung nicht wissenschaftlich begründeter Verfahren in der Medizin. In: Köbberling, J. (Hrsg.): *Die Wissenschaft in der Medizin*. Schattauer, Stuttgart - New York; 83-114

Wolf, R.F., Cohen, A.M. (1997): The minuscule benefit of serial carcinoembryonic antigen monitoring after effective curative treatment for primary colorectal cancer. *J. Am. College of Surgeons* 185, 60-64

Wood, C.B., Ratcliffe, J.G., Burt, R.W., Malcolm, A.J.H., Blumgart, L.H. (1980): The clinical significance of the pattern of elevated serum carcinoembryonic antigen (CEA) levels in recurrent colorectal cancer. *Br. J. Surgery* 67, 46-48