

患者中心の癌治療と生存価値の評価尺度

医学統計解析グループ（代表：前谷俊三）

天理よろづ相談所 医学研究所

ランダム化臨床試験により、癌治療の評価は著しく向上したが、理想的な評価尺度はまだ開発されていない。更に多くの患者が治療選択に参加するようになった今（患者中心の医療）、提供される統計学的情報も、素人にも理解でき、かつ患者のアウトカムからできるだけ乖離しないことが必要である。我々はこの視点から現行の3つの生存統計量と平均余命を再検討した。まずログランク検定は治療の有効性を効果サイズで表していないので（ 2 値）、有効か否かの判断はできても、使用に値するだけの有効性があるかどうか分からない。ハザード比は効果サイズを、有効性の差ではなく比で表している。その上リスク比と異なり、繰り返し曝される死亡の危険を各時点で比較するため、全体でどれだけの損得になるか解釈が難しい。さらにログランク検定と同様、早期の死亡に鋭敏に反応するが、追跡期間を延ばしていくと、予後因子のアウトカムに及ぼす影響が低下する傾向がある（比例ハザード性が成立しない）。5年生存率は素人にもわかりやすく、おおまかながら疾患の予後を表す。しかし同時に年齢の影響を受け、また長期のアウトカムを正確に反映するわけではない。平均余命は追跡期間を超えた生存曲線の予測が必要であり、その使用は限られていた。しかしBoagのモデルによるパラメトリックな推定法で予測の精度が進歩し、多くの長所が認められた。例えば、癌やその治療が人生にもたらす恩恵や損失を端的に年数で表し、それは患者の年齢に応じて変化する。また乳癌では癌を根絶する補助化学療法と延命効果しかもたない療法とを限られた追跡期間で識別することができた。パラメトリックな分析はなお発達段階であるが、患者自身が自分に適した治療を選択するためには更なる評価法の進歩が待たれる。

キーワード：癌治療，生存価値，パラメトリックモデル，平均余命，生存統計量

【別刷請求先】

〒632-8552 天理市三島町 200
天理よろづ相談所 医学研究所
前谷俊三

はじめに

癌のような重篤な疾患に対する医療の成果をどのような尺度で評価するかという問題はなお未解決であり、gold standardの確立までには至っていない。我々は1999年以来本誌において繰り返しこの問題に触れ、医療者は未だ患者が求める十分な情報を提供していないことを指摘してきた。¹⁻⁶ 確かに近年ランダム化臨床試験の実施により医療評価に格段の進歩がみられたが、被験者の追跡調査は通常5年か高々10年という制約がある。ところが患者や社会が知りたいのは5年の時点の成績ではない。[終りよければ全てよし]という諺のように、長期的にみた患者の最終的生存価値(long-term survival benefit)を高めるのはどの治療かである。果たして5年生存率や平均ハザード比のように限られた期間の追跡データから導いた生存統計量は長期予後を正しく推測できるのだろうか。更に、たとえ長期の追跡データが得られたとしても、患者が最も知りたい情報を患者のわかる言葉で正確に伝えているのであろうか。今回は過去の報告と一部重複するところもあるが、この問題を取り上げ、患者の視点から再検討することにする。

患者主役の医療

今医療評価はパラダイムシフトを迎えているといわれる。それでは何が変わっているのであろうか。その一つは患者中心の医療、あるいは患者が主役の医療である。これは医療従事者が患者の立場と患者の目線に立って医療を行うことだけではない。患者が積極的に自らの医療に関与することを意味する(shared decision making)。かつて

は患者が十分に理解できなくても患者のためになると思える医療行為を医師が行うことは暗黙裡に容認されていた。これはあたかも父親がわが子のためにする行為に喩えられ、パターナリズム(paternalism)と呼ばれているが、この言葉も今や悪い意味合いで使われている。医師はたとえ最善と思える医療行為をする場合にも、十分な説明により患者を納得させ、その了承(informed consent)を得なければならない。それだけではない。医師は可能性のある治療法を患者に提示し(医療の透明と医療情報の共有)、その中のいずれかを選択するのは患者自身が決めるという考え(informed choice)が医療の理念となっている。

勿論医療技術を行使するのは医師やその他の医療職であることに変わりはない。例えば手術は医師とそのチームに任されるが、手術をするかどうかは決めるのは患者が主となると、ここで重大な疑問が生じる。医師にとっては専門の医療技術を行使することはそれほど困難なことではない。それよりもっと難しいのは、特定の患者に対してどの医療技術を行使すべきかどうかを正しく判断することである。その判断の良し悪しのほうが患者の予後を大きく左右する可能性が高い。その最終決定を生半可な医学知識しかもたない患者にまかせれば、むしろ医療の質を低下させるおそれはないだろうか。これは医師やその他の医療従事者が抱く当然の懸念であらう。しかし他方では、患者の判断を尊重すべき場合も多い。例えば癌手術に続く補助化学療法がどれだけの苦痛と金銭的負担をもたらすかを実感するのは患者側である。むしろ医師は補助化学療法のメリットとデメリット

を患者に伝え、患者にその選択をまかせるべきではないだろうか。問題はいかにわかりやすく、かつ正確に医療情報を患者に伝えるかという医師の説明責任 (professional accountability) にかかってくる。これは医療技術や判断力と同様に医師の備えるべき必須の資質となっている。

アウトカムに基づく医療評価

医療における意思決定がいかに困難かは既にヒポクラテスの金言にもみられる。⁸ しかし患者自身が医療の良し悪しを見分ける手立てがないとはいえない。むしろパターンリズムが患者の啓蒙を阻み、その手立てが十分生かされなかったのかもしれない。その手立てとは結果 (アウトカム) に基づく医療評価である。かつてはある疾患に対する治療法の妥当性を証明するためには、疾患の原因や治療の作用機序を究明することが正道とされてきた。しかしこれを非専門家が理解することは極めて難しい。しかし治療法の良し悪しの決め手になるのは、実際にその治療を行った結果をみるほうがよい場合が多い。これならば素人でもできることである。

その実例は1880年代に遡る。当時、日本軍を悩ませたのは脚気による兵士の死亡であり、その数は戦死者を上回ることもあった。後に海軍の軍医総監となった高木兼寛

はその原因を白米を主食とする和食にあるとみなした。ところが、陸軍軍医正の森林太郎 (鷗外) は最新の設備を備えた検査室での分析の結果、高木の説を否定し、細菌が脚気の原因と主張した。これに対して海軍は全く対称的な評価法を実行した。即ち、遠洋航海の乗組員に洋食を与えて、以前の白米を主食とした乗組員と比べてどちらに脚気が多いか、その結果を比較したのである (表1)。答えは明白で白米で脚気が多発することを確認、これを米麦混合食に代えた。それ以後海軍では脚気が激減したが、白米を継げた陸軍では脚気による多数の死亡が続いた (表2)。⁹ 意外にも当時の医学界は森を「実験法と学理に習熟する」と高く評価し、高木の考えを「統計に拠る学理なき説」と批判したという。しかしもし兵士に表2の結果を見せ、自分で主食を選ばせていれば、たとえ専門的な実験法や理論を知らなくても、もっと賢明な選択をしていたかもしれない。

このようにアウトカムとは病気になるや死亡するように、素人でもそれが理解でき、自分で体験できる結果である。これは悪い結果だけではなく、痛みがとれたや、元気で生きているというよい結果でもよい。これを評価基準にするならば、医療の意思決定には患者も参加できる。いわゆる証拠に基づく医療 (evidence-based medicine、

表1. 遠洋航海における食餌の脚気発生に及ぼす影響

船名	出港日	航海日数	主食	乗組員	脚気発病 (死亡)
龍驤	1882年12月	272日	白米	378人	169人 (23)
筑波	1884年2月	278日	パン	338人	15人 (0)

吉村昭著「白い航跡」より

表2. 陸軍と海軍の脚気発症・死亡の推移

年	陸軍		海軍	
	発症	死亡	発症	死亡
1882	7884	204	1925	51
1883	9939	235	1236	49
1884	10225	209	718	8
1885	6609	63	41	0
1894-1895 (日清戦争)	34783	3944	34	1

吉村昭著「白い航跡」より

EBM)^{5,10,11}でいうエビデンスもこのような結果から得たものである。逆に実験や理論から導いた結果は証拠としての重みは小さい。例えば「検査の数値がよくなった」という類の結果は患者にとってその得失を判断できないので、それだけでは評価の決め手にはならない。これと区別するためには厳密にはアウトカムを臨床的にみて適切なアウトカム (clinically relevant outcome), または患者に立脚したアウトカム (patient-based outcome)¹²と呼ぶべきであろう。勿論このアウトカムと検査の数値の間に高い相関があれば、検査結果を代替のアウトカムとすることはできる。

アウトカムが医療評価のための基準となるためにはもう一つの重要な条件がある。それは異なるアウトカムを比較してどちらが自分にとって損か得かだけでなく、損得の差がどれだけかを患者自身が判断できなければならない。例えば上述の補助化学療法をすべきかどうかを決定する場合、強い苦痛や高い医療費という悪いアウトカムを上回るよいアウトカムが得えられるかどうかを患者自身が実感できなければならない。

表3. 日露戦争における陸軍の戦力と被害

	日本	ロシア
兵力(1905年9月)	87.8万人	100万人
死者	8.4万人	2.8万人

毎日新聞(2005年5月26日)より

い。そのためには例えば「治療をすればしない場合と比べて3年間は長生きできること」など比較可能な具体的なアウトカムを提示する必要がある。しかし、アウトカムは互いに質が異なるだけでなく、それが起きる確率も異なるため、正しい比較は決して容易ではない。アウトカムを要約するためには統計学の助けが必要であるが、それが理解を阻む一因ともなっている。以下にこの問題を述べる。

軍の勝敗と兵士の運命

表3は今から丁度100年前、日露戦争での日本軍とロシア軍の陸軍同士の兵力と戦死者数を比較したものである。戦いに勝った日本軍の兵士の死亡率は約10%であり、敗れたロシア軍兵士の死亡率(3%)の約3

倍となっている。その頃歌人と謝野晶子は旅順攻撃に参戦する弟の身を案じて「君死に給うことなかれ」という詩を明星に発表した。そこで彼女は旅順攻防戦に勝つことよりも、弟が無事帰還することを祈る気持ちを赤裸々に吐露した。しかしその願いがかなう可能性は、平均すればロシア軍兵士のほうが日本軍兵士よりも高かったのである。ちなみにこの攻防戦を指揮した乃木將軍もこの戦いで2人の子息を全て失っている。即ち、兵士のリスクはそれが属する軍の勝敗からは予測できない。

それと似たことがランダム化臨床比較試験(RCT)についてもいえる。RCTによってある疾患に対する2つの治療法の優劣を決めるためには、その治療法を互いに似通った2つの患者群に割り付け、そのアウトカムを比較する。通常のRCT(パラレルデザイン¹³)で比較するのは群対群である。それから導かれた結果は軍の勝敗と同じく、個々の患者の利害を正確に表すとは限らない。EBMの提唱者の一人であるSackettらはEBMを適用する対象を、群ではなく個々の患者に置いている(making decisions about the care of individual patients)¹⁰。この視点は臨床医に共通のものである。ところが彼らのいう強いエビデンスとはRCTから得たものである。果たして彼はこの群と個人の結果の違いを十分認識していたのだろうか。

個々の患者の利害を反映する尺度

ここで問題となるのは「患者の利害をよく反映する尺度は何か」である。もし日露戦争において、軍の勝敗ではなく兵士の死亡率を尺度とすれば、両軍の兵士の運命は

もっと正確に評価できる筈である。同様に急性の重篤な疾患では、どちらの群で死亡率が少ないかを評価基準にすれば、患者にとってよりよい治療法を選べる可能性が高まる。しかしこれで問題は全て解決したわけではなく、むしろ問題はこれからである。

第一の問題は一口に日本軍兵士の死亡率といっても最前線で突撃に加わる兵士と、後方でこれを支援する兵士の間には当然死亡率に違いがあると考えられる。後方支援部隊の死亡率はロシア軍守備隊の死亡率よりも低いかもしれない。もし軍が死亡率の異なる異質な複数の集団からなるならば、全体の死亡率を目安にしてもこれと個々の兵士のリスクには大きい開きが生じ得る。これでは軍の勝敗から兵士の運命を予測するのと大同小異である。医療においても同じ集団の中には重症度や遺伝子変異など、既知や未知の予後因子に違いがあり、それに応じて死亡率にも差が生じるのが普通である。それを十把ひとからげにして「治療AはBに比べて有意に死亡率を減らすというエビデンスがあるので、Aを標準治療とすべきである」といってよいだろうか。このように平均値を評価基準として、その値のよいほうの治療を選択することを期待値基準(expected value criterion)という。しかし中にはBを選んだほうがよかったにも拘わらず、期待値基準に従ったばかりに、反って結果が裏目に出ることもある(患者と治療法間の交互作用⁶)。われわれはこの危険を $P(n=1)$ という確率で表した^{4,5}。これは期待値基準で治療を選択する場合には知っておきたい2つの情報の中のひとつである。もう一つの情報は治療効果(treatment effect)または効果サイズ(effect size)と呼

ばれるもので、比較すべき2治療の効果に平均してどれだけの差があるかを表す数値である。現行のRCT(パラレルデザイン)ではこの効果サイズやその信頼区間は計算できても、 $P(n=1)$ は求められない。つまり期待値基準に従えば反って損をするリスクがどれだけかはわからない。

第2の問題は癌のような慢性疾患の場合には、死亡率だけを比較するだけでは十分とはいえない。生死だけでなく、死亡するまでにどれだけ長く生きたかも治療法を選ぶ重要な要素となるからである。

第3の問題は治療で疾患をどれだけ治したかと、それによって患者の生存価値がどれだけ高まったかは必ずしも一致しないことである。両者はしばしば混同され、どちらを選ぶかで治療の選択に違いが生じえる。われわれは前者を測る尺度を疾患予後基準と呼び、後者を測る尺度を生存価値基準(患者予後基準)と呼んで区別した。³ 両者の違いは医師の立場と患者の立場の相違と似ている。例えば二人の外科医は同じ進行癌をもつ一人の青年と一人の老人に対してそれぞれ根治手術を行った。その結果青年は術後10年後も健在であった。一方、老人は1年後に持病の悪化で他病死亡したが、癌の再発は認めなかった。医師側からみれば、二人の患者は共に目的の癌は全治したので(疾患予後は同じ)、同等の評価を受けるべきと考えてもおかしくはない(2人の技量に優劣はつけられない)。しかし患者側からみて重要なのは、癌治療の成否ではなく、治療が患者の生存にどれだけの恩恵(survival benefit)をもたらしたかである。この恩恵は二人の患者の間で大きい違いがある。もう一つの例を挙げると、ある検査法

は感度99.9%、特異度99.9%でエイズの診断が可能であった。この値をみると検査が陽性とであれば間違いなくエイズと誤解する者がいるかもしれない。しかし感度や特異度とは検査の性能を測る尺度であり、患者がエイズである可能性を表す尺度ではない。後者を測る尺度は陽性予測値と陰性予測値である。日本人のエイズ感染者を5000人に1人とした場合、この検査で陽性と出た被験者がエイズである可能性は17%に過ぎない。⁵ これは誤った尺度を使用すると判断を誤る例である。同様の理由により、患者にとってよい治療を選ぶためには疾患予後ではなく、生存価値を尺度とすべきであるが、従来の生存統計量はどちらの尺度が明らかでない。

第4の問題は限られた追跡期間のデータから長期の結果を予測する場合にどれだけの誤差が生じるかである。以下にこれらの問題を念頭に入れ現在使われている3つの評価基準と平均余命を再吟味する。

Mantel-Haenszel 検定 (ログランク検定)

1966年Mantelは所謂Mantel-Haenszelの²検定法¹⁴を生存データに適用し、群間の生存パターンを比較するための新しい方法を発表した。¹⁵ この方法では、もし2群の最終死亡率が同じでも、死亡時期に早いか遅いかの違いがあれば群間に差が生じるようになっている。表4にその1例を示す。A、Bの2群はいずれも総数は50例で中途打ち切りはない。死亡は2時点で置き、A群には早期死亡が多く、B群には後期の死亡が多い。表4-1では両群とも合計38例が死亡し、最終死亡率は同じであるが、 $\chi^2 = 5.53$ とA群が有意に悪い。この結果自体は特に

表 4-1. Mantel-Haenszel 検定 (その1)

	A群 (治療A)	B群 (治療B)
総数	50	50
早期死亡数 (死亡順位1)	35	10
後期死亡数 (死亡順位2)	3	28
死亡総数 (死亡率)	38 (76%)	38 (76%)

ログランク 検定: $\chi^2 = 5.53$ $P = 0.02$

表 4-2. Mantel-Haenszel 検定 (その2)

	A群 (治療A)	B群 (治療B)
総数	50	50
早期死亡数 (死亡順位1)	35	10
後期死亡数 (死亡順位2)	2	28
死亡総数 (死亡率)	37 (74%)	38 (76%)

ログランク 検定: $\chi^2 = 4.43$ $P = 0.05$

問題とはならないが、次の表 4-2 の結果は医師や患者に Mantel-Haenszel 法に対する疑問を起こさせる。この表では A 群の後期死亡数を 1 例だけ減らし、全体の死亡率は B 群より少なくした。にも拘わらず早期死亡の少ない B 群が A 群より有意に良好な結果である。つまり彼の方法は全体の死亡率にもまして早期の死亡を重視し、生存率からみた結果とは逆になっている。Mantel は彼の方法が群の生存価値を反映すると考え、survival value function と呼んだ。果たしてこの検定結果が生存価値を正しく反映すると考えてよいのだろうか。答えは否定的である。

何故ならば、もし後期死亡者が生存者と変わらないくらい長生きしたのであれば、後期死亡の多い B 群が A 群に勝るとみなしてもよい。しかし本検定では各患者の生存期間から、長さという情報を取り除き、全体で何番目に早く死亡したかという死亡順

位のみを統計量の算出に使っている(順位統計量)。表 4 の後期死亡というのは生存期間の長さにして 1 年であっても 10 年であっても計算結果に変わりはない。Mantel が「その方法はより早期の死亡に、より大きい重みを付与する (It gives greater weight to earlier deaths)」¹⁵ といっているが、重み付けの妥当性に疑問がある。

しかしこの検定法はログランク検定として今も広く使用されている。その理由は群間のハザード比(後述)が一定の場合(比例ハザード性¹⁶の成り立つ場合)に本法の検出力が高く、ハザード比によって群間の治療効果を比較するのに適しているからである。こうして本法の意味付けが変わり、死亡時期に違いのある 2 群ではなく、むしろハザード比が一定の 2 群の比較に有用な方法となった。それはともかくこの検定法のもう一つの問題点は結果が χ^2 値か P 値

でしか表されない(効果サイズでは表されない),素人には難解である.また効果に差があるかどうかはわかって,どれだけの差があるかはわからない.これを解決するためには次に述べるハザード比が使用されている.いずれにせよMantelが治療効果の比較において生存値関数を念頭に入れたことは注目すべきであろう.

ハザード比(相対ハザード)

ハザードとは別名瞬間死亡率ともいい,一種の死亡率である.これはある時点まで生きていた被験者の中でその時点で死亡するものの割合である.従って分子をその時点の死亡数とすると,分母は患者総数ではなく,それぞれの死亡が起きる直前の生存者の数である.ハザードは時間と共に変動する.通常の癌では最初の数年間は癌による死亡率は高く,この間にピークがみられ

る.その後死亡率は急速に下降するが,今度は加齢のための死亡が時と共に尻上がりに上昇する(図1).このようにハザード自体は一定の値をとらないので,その値によって各群の予後や治療効果を表すことはできない.ただ2群A,Bのハザード H_A, H_B を同一時点で比較するとその比 H_B/H_A は時点の選び方に拘わらず比較的一定となるので,平均的なハザード比は予後指標として使用されている.もしこの値が1であれば群間の予後に差がなく,1より小さいほどAに比べてBの予後が良好なことになる.表5はその例であり,B群のハザードは各時点で変動するが,同一時点でAと比較すると,その比は大よそ0.5である.つまりB群の瞬間死亡率はAの約半分である.しかし両群の生存価値を比較する場合,ハザード比は誤解をきたしやすい尺度である.第1に,リスク比(相対リスク)と同

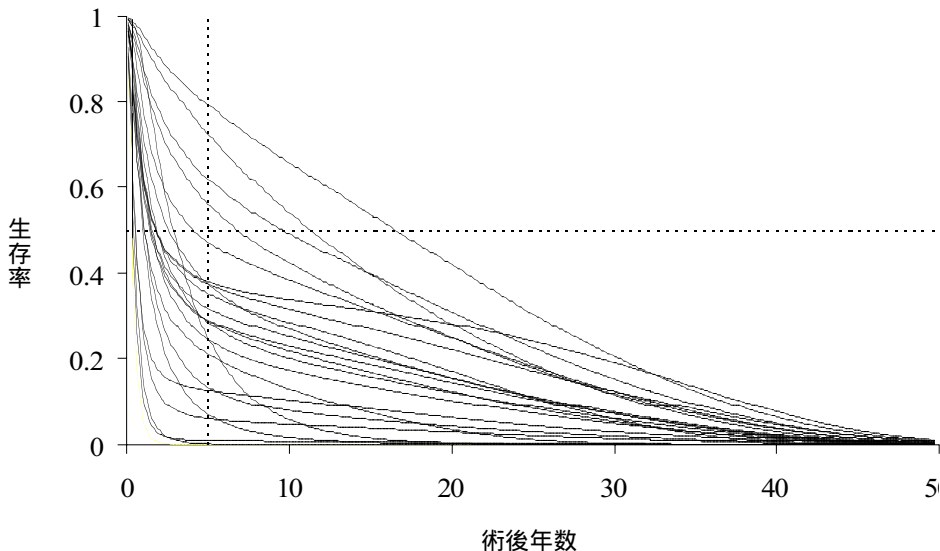


図1.胃癌患者の種々の群の生存曲線
生存曲線と垂線との交点は各群の5年生存率を表す.曲線と水平線の交点は各群の生存期間中央値を示す.

表5. ハザード比

順位	A群 (治療A)			B群 (治療B)			ハザード比 (B/A)
	生存数	死亡数	ハザード	生存数	死亡数	ハザード	
1	50	40	80.0%	50	20	40.0%	0.50
2	10	6	60.0%	30	9	30.0%	0.50
3	4	2	50.0%	21	6	28.6%	0.57
全体	50	48	96.0%	50	35	70.0%	0.73

平均ハザード比=0.53

様、この指標は群間の違いを差ではなく比で表している点にある^{1,2,4}。比は同じ値でも生存価値に及ぼす影響は色々である。例えば治療Aに対する治療Bのリスク比は0.5としよう。これはBにより死亡のリスクが半減したことを意味する。慌て者ならば高い金を余分に支払っても、治療Bを受けるかもしれない。しかし、もし治療Aの死亡リスクが1%だとすれば、Bのリスクはその半分の0.5%に減るだけであり、両群のリスクの差はわずか0.5%(1% - 0.5%)に過ぎない。従って治療Bの恩恵を受けるのは200人(=1/0.005)中わずか1人となり、残りの199人は治療Aを受けるのと変わりはなく、無駄金を払うことになる。これに対しても治療Aでは全員死亡するとすれば、死亡リスク = 100%、その半分の50%はBで助かることになる。その恩恵に浴するのは2人(1/0.5)の中の1人であり、これなら高額の治療費を払う値打ちがある。このように「利益を得るのは何人中の一人か」という形に変えるのがNNT (number needed to treat)¹⁷であり、リスク比の問題点が浮き彫りになる。

第2にはリスク比ではなく、ハザード比に特有のものである。リスク比の場合は危

険が発生するのはまとめて1回であるが、ハザード比の場合は一回目の危険を免れたとしても、第2、第3の危険が待っている。ハザード比が0.5というのはこの1回当たり(または単位時間当たり)のリスク比が0.5という意味であり、全体とすればこれより1に近いリスク比となるのが普通である。表5では3回死の危険があり、1回毎にみれば、治療BはAに比べてリスクが平均0.53倍である。しかし全体を一括してみれば最後の行に示すように、リスク比は0.73倍である。言い換えればリスクは27%(1 - 0.73)だけしか減少していない。ところがこれを平均ハザード比から計算して、「治療Bは死亡のリスクを47%(=1 - 0.53)減少する」と記載した論文をみかける。これはBの治療効果を誇張することになる^{4,18-22}。

第3には治療の評価に拘わる重要なことであるが、あまり知られていない。一般に平均ハザード比やログランク統計量は、病気を完全に根治する治療法よりは、一時的な延命効果しかもたない治療法に鋭敏に反応する²⁰⁻²²。延命効果を見落とさないこと自体は別に問題ではないが、これと比べて根治療法の検出力が弱いことは由々しい問題である。というのは二つの治療法をRCT

で比較すると、保存療法が根治療法よりよく効くという誤った結果が出るおそれがあるからである。これは特に追跡期間が短い場合にみられる。

第4に平均ハザード比を効果サイズとする背景には、「ハザード比が全期間を通して一定である」という前提（比例ハザード性）がある。しかし、これが成立するとしても、一定期間内にしか当てはまらないことは医学的常識からみれば肯ける。例えば石綿の暴露による中皮腫の発生や、C型肝炎ウイルス陽性の輸血による肝癌などを対照群と比較する場合、初期には群間に差がなく、ハザード比はほぼ1である。差が出るのは何十年という先のことである。逆に治療後の悪性腫瘍患者では、術後長く生きるほど腫瘍関連予後因子の生存に及ぼす影響が薄れ²³⁻²⁴ 平均ハザード比は1に近づく傾向がある。³ これと裏付ける所見としては、毎年の実測生存曲線から比例ハザードモデルに基づきそれ以後の生存曲線を予測すると、実測曲線が平均から偏移する群ほど予測曲線の偏移が誇張される。²⁵ 言い替えばいくら例数を増やしても、限られた期間の追跡結果から未来を予測すると偏った結論を導くことになる。

第5に、比較の基準群の選び方によって結果が逆転することである。例えば、もしハザード比でA、Bの治療群を直接比較したところ、Aのほうが有効であった（A対Bのハザード比が1より小）とする。ところが今度はC群を基準としてA、B群を別々に比較したところ、Bのほうが有効B対Cのハザード比がA対Cのハザード比より小）というパラドックスが起こりえる。

5年生存率とその関連尺度

5年生存率は上述のログランク検定やハザード比と比べると、素人にもわかりやすく、疾患の予後や治療法の良し悪しを大まかに評価する上では有用な指標として使われてきた。しかし厳密に言えば以下に述べるような問題を抱えている。

第1に病気の進行度や治療法を同じに揃えても、患者が高齢になるほど5年生存率は低くなる。例えば日本人の生命表から計算すると1990年に20歳であった一般男性が5年後に生きている確率は、99.6%であるが、80歳の男性ならばこれが60.5%に低下する。従って80歳の男性患者が治療によって病気が完治したとしても、毎生存率でみれば20歳の患者よりも遥かに悪い成績となる。このような場合には5年生存率は疾患の予後よりはむしろ生存価値の違いを表すとみなされる。

この偏りを避け原疾患だけの予後を見たい場合には、患者の生存率を、患者と同性、同年生まれの一般人の生存率で割って求めた相対生存率を使用する。これは後述する疾患関連生存曲線やBoagのモデルに基づく生存曲線²⁶ とほぼ同じ形となる。悪性腫瘍ではこの曲線の勾配は次第に緩やかになり、水平となったときのX軸からの高さが全治率（cure rate）²⁶ である。これは疾患予後基準となる。

第2は5年生存率はログランク検定や平均ハザード比と比べるとまだ後期に重みを置いた評価法といえるが、それでも後二者と同様の問題が存在する。即ち、治療法の中には一定の割合の患者が全治する療法がある一方、再発時期を後にずらすだけの一時しのぎの療法もある。後者の5年生存率

が前者を上回り、あたかも姑息療法が根治療法よりも優れているかのような誤解を与えることがある。例えばある癌の姑息療法により5年生存率が3%上昇したとしよう。もし5年以内に打ち切り例がなければ、これは新たに3%の患者の生存時間を5年未

満から5年以上に延長したことを意味する。一方、根治手術により3%の患者が新たに全治したとすれば、彼らのほうが大きい恩恵を受ける筈である。ところが根治手術で全治するのは比較的限局した早期の癌に偏ることが多い。そこでたとえ根治手術をしな

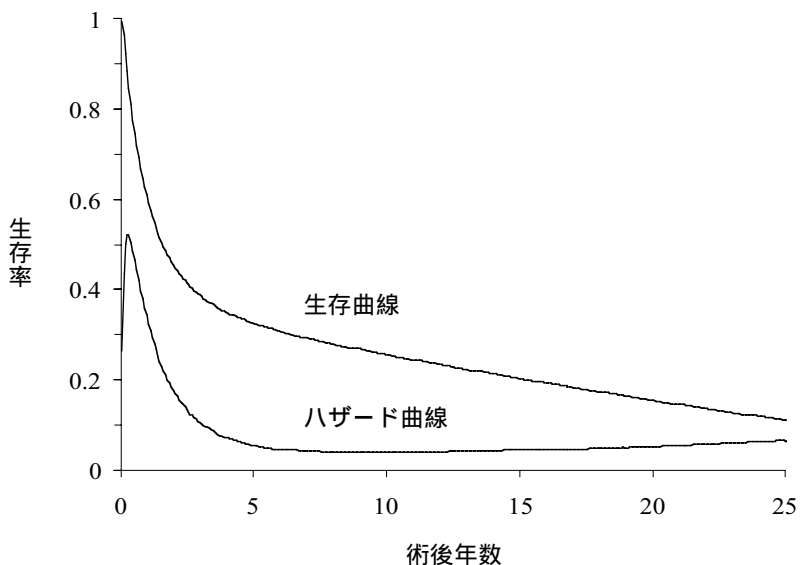


図2．胃癌患者の生存曲線とハザード曲線

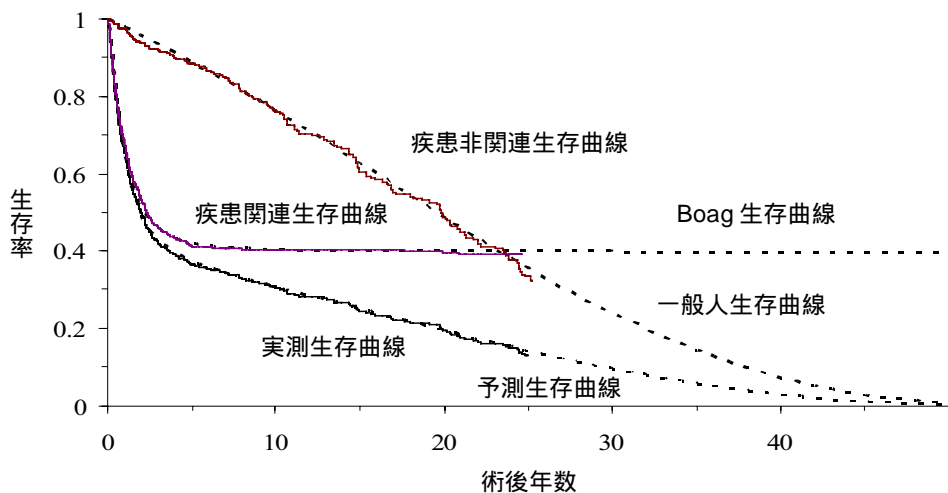


図3．競合リスクモデルによる予測生存曲線の求め方

くても彼らが再発死するのは5年以後だと仮定すれば、彼らが根治手術後いくら長生きしたとしても、5年生存率は根治手術を受けない場合と変わらない。このパラドックスを回避するためには上述の全治率を基準とするか、後に述べる平均余命に拠らねばならない。

上記とは対照的にログラंक値やハザード比は、患者の死亡時点が少し動いても反応する。但しもし観察期間が5年であれば、反応はその期間内の変動に対してのみである。それを超えて生存期間がいくら延長しても結果は変わらない。

図2は天理よろづ相談所病院で30年以上前に手術した約1000人の胃癌患者を種々の因子で群分けし、24群の生存曲線を数理モデルにより50年先まで予測したものである。横軸の術後5年を通る垂線と各生存曲線との交点はその群の5年生存率を表す。生存曲線同士が交叉し、時間によって生存率が逆転することも稀ではない。これをみると、「果たしてこの生存曲線上の1点だけからその群の予後を正しく評価できるだろうか」という疑問が生じる。これに対して1点ではなく、全生存曲線から導いた指標が次にのべる平均余命である。

平均余命

図2において群間の比較のためには特定の時点の生存率ではなく、全生存曲線の下面積の大きさを尺度にしてはどうだろうか。実はこの面積は一人一人の患者が生きた年数を足してその人数で割った値、即ち平均余命である。言い替えれば一人一人の患者が生きた1年は、患者の能力、年齢、時代を越えて同じ価値をもつと仮定すれば、

この面積がその群の生存価値を表すこととなる。その計算のためには限られた追跡期間のデータから、その先の生存曲線まで予測する必要がある。

1. 予測生存曲線

図4に競合リスクモデルに基づく予測生存曲線の求め方を示す。¹⁸ まず通常のKaplan-Meierの実測全生存曲線を死因により二つの成分に分ける。その一つは原疾患に関連した死亡が起きたときだけ生存曲線が下降する疾患関連生存曲線である。他の一つはこれと逆で、原疾患以外で死亡したときだけ生存曲線が下降する疾患非関連生存曲線である。こうして得られた2つの曲線は実際の追跡期間内でしか描けず、いわば尻切れトンボである。そこで前者にはBoagのモデルに基づく曲線を肩代わり(simulate)させ、後者は患者と同性同年生まれの一般日本人の生存曲線を肩代わりさせる。その結果2つの近似曲線は追跡期間を超えて将来に(図では50年まで)延長することができる。そこで2つの曲線を再合成すると(掛け合わせると)十分な長さの予測生存曲線が得られる。平均余命はその曲線の下面積を測れば求められる。

2. Boagのパラメータ²⁶

以上の計算中Boagの3個のパラメータが求まる。ここで言うパラメータとは、生存曲線の特徴を表す数値である。Boagは癌患者を治療で全治する割合 c と残りの非治癒($1-c$)にわけた。更に非治癒例が癌死するまでの時間(対数)は正規分布をすと仮定し、その平均を m 、標準偏差を s とした。こうして求めたBoagモデルのパラメータ c 、 m 、 s はその群の疾患予後をよく表し、この3つの値から疾患関連生存曲線が再現できる。

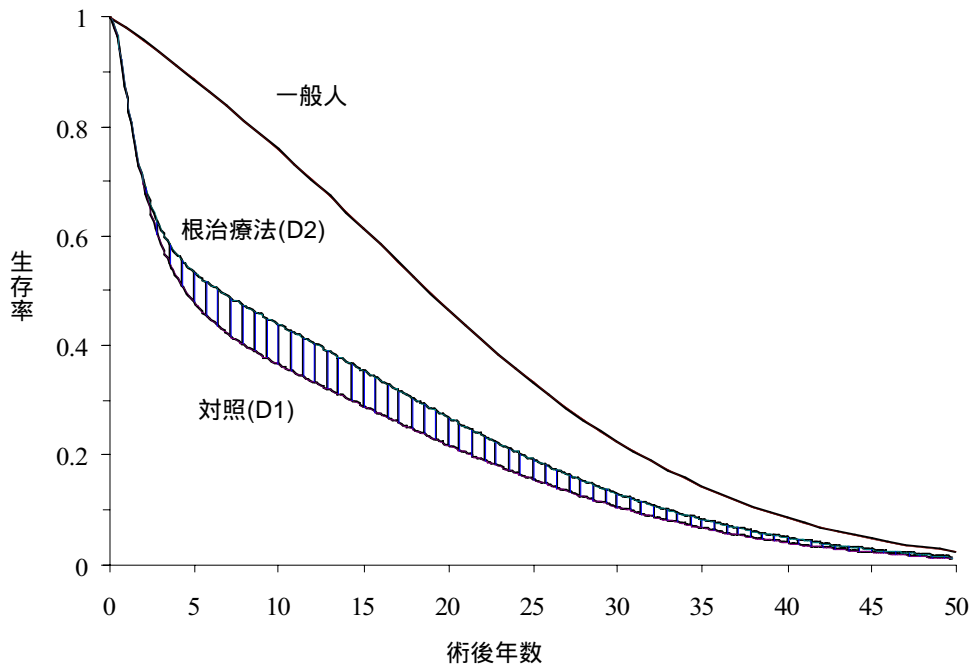


図4．オランダでの胃癌に対するD1対D2手術群の生存曲線(近似)
 模様面の面積は、両群の平均余命の差を表す．両曲線の間隙は最初狭く、次第に開大する．

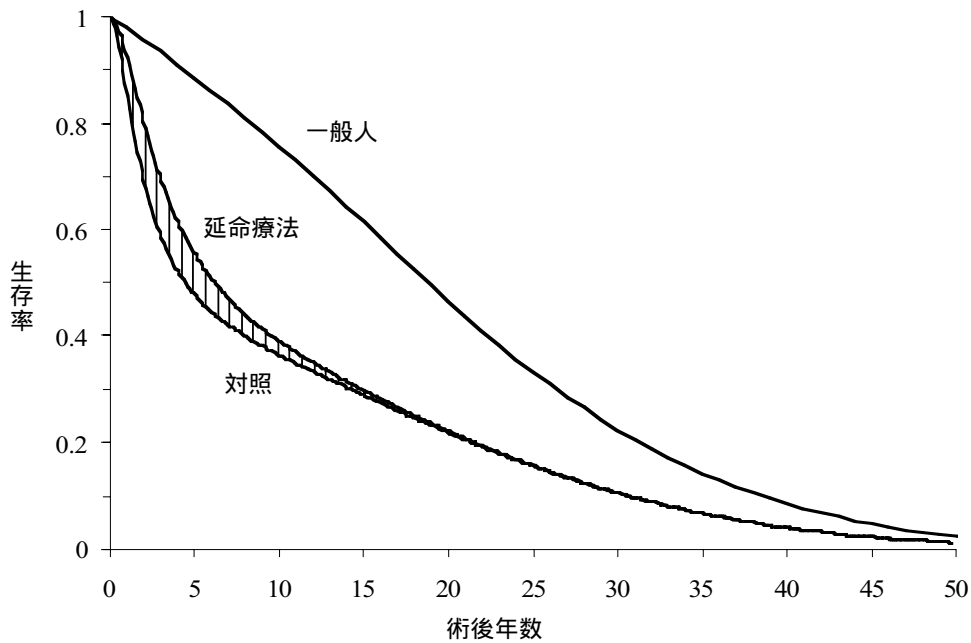


図5．延命効果しか達成できない治療群対対照群の生存曲線(架空)
 模様面の面積は、両群の平均余命の差を表す．両曲線は早く開き、早く閉じる．

3. 平均余命の意義

こうして得られた平均余命は生存価値基準としての利点の外に、従来の統計量と比べて以下のようなメリットがある。

まず「疾患により平均して人生の何年が失われ、治療によりそのどれだけが、取り戻せるか」という素人にとっても分かりやすい情報が提供できる。しかもそれが数値だけでなく、グラフ上の面積として表され、一般人と比較してどれだけの人生を失うかが視覚的に捉えやすい。図4はオランダでの胃癌に対する通常リンパ節郭清術(D1)と拡大リンパ節郭清(D2)を比較した成績である。²⁷ 患者と同性同年生まれの一般人の平均余命を19年とし、かつ手術関連死がないと仮定した場合に、縞模様面積で示すようにD2患者がD1患者に比べて約1.5年長く生きることになる。

一方、治療がもたらす survival benefit の大きさは、患者の年齢と治療の根治性に大きく左右される。例えばGamelは乳癌のステージIIに対する補助化学療法の効果の評価する複数のRCTデータを再解析した。その結果延命効果しかもたないレジメンと、全治例を増やすレメンのあることを確認した。²⁸ もし後者のレメンを実施すれば、手術後宿主に残存して再発死をきたす癌細胞が死滅するとみなされるので、化学療法のもつインパクトは大きい。従来の分析ではこの根治療法と保存療法の違いをみわけられないか、むしろ保存療法を過大評価する傾向がみられた。図5は図4と対照群が同じで、これと治療群はとの差は延命のみと仮定している(Boagのパラメータの中でmのみを対照より大きく設定している)。一方、図4のD2群では全治率cが対照より約10%

高く、平均余命(縞模様)も長いことがわかる。ところが両者を5年生存率で比べると、保存療法のほうが根治手術を受けた群より高くなっている。生存率によって根治手術の優位を確かめるためには5年では不十分なことがわかる。一方、平均余命や全治率でみればどちらが望ましい治療かが確かめられる。それだけでなく、どのような患者がその化学療法のレスポンスかも平均余命から推測できる。その結果平均余命は年齢によって大きい差があり、若い患者ほど長期生存の可能性が高いことがわかった。我々のモデル¹⁸によれば個々の患者の平均余命はBoagのパラメータと、患者と同性同年生まれの一般人の平均余命から計算できるので、患者はより自分に近い群の生存情報を得ることができる(personalized medicine)。

平均余命のもう一つの有用性は、癌化学療法などに要する医療費が大幅に高騰する時代を迎えて、それぞれの診断や治療がどれだけcost-effectiveかを見極めるてだとなることである。これは費用対効果比(1年寿命を延ばすためにどれだけ費用がかかるか)を算出するために不可欠の数値である。

かつてPeto等10名の世界的統計学者は「平均生存時間は他の統計量よりも遥かに悪く、記載すべきでない(Average survival times can be far more worse, and should almost never be cited)」と述べた。²⁹ また追跡期間を超えて生存曲線を延長することは、観察データのない領域にまで回帰曲線を外挿するのと同じ過ちを犯すとみなされていた。ところが50年を超える胃癌患者の追跡データから求めた実測生存曲線と、5年の

追跡データから拡張 Boag モデルを使って予測した生存曲線を比較すると、例数さえ増やせば予想以上に両者はよく一致することがわかった。¹⁸ 更に Cox の回帰モデルと同様に、Boag のパラメータを多変量解析により、治療その他の予後因子から推定することが可能となっている。^{25,32} パラメトリックな平均生存分析はなお発展段階であり、改良の余地が残されているが、平均余命の意義を認める研究者は次第に増加している。^{19, 28,30,31} 患者自身が各自に適した治療をきめ細かく選択するためには、評価法の一層の進歩が待たれる。

参考文献

1. 天理医学統計解析グループ．情報の価値からみた医学統計学の再検討．天理医学紀要 1999;2:138-153.
2. 天理医学統計解析グループ．証拠に基づく医療 (evidence-based medicine) その妥当性と限界．天理医学紀要 2000;3:138-153.
3. 天理医学統計解析グループ．臨床側からみた癌の生存分析：Cox モデルから新しい生存モデルへ．天理医学紀要 2001;4:128-129.
4. 天理医学統計解析グループ．Neyman-Pearson 統計学から新しい臨床統計学へ：エビデンスよりは説明責任を．天理医学紀要 2002;5:100-115. <ダウンロード可能>
5. 天理医学統計解析グループ．医療におけるエビデンスと P 値．天理医学紀要 2003;6:54-74. <ダウンロード可能>
6. 天理医学統計解析グループ．根治療法は両刃の剣か？患者治療間の交互作用．天理医学紀要 200;7:78-103. <ダウンロード可能>
7. 矢崎義雄．医療評価のパラダイムシフトに向けて．日本医師会雑誌 2005; 134: 24-27.
8. Colder R. Medicine and man: story of the art and science of healing. 佐久間昭訳．物語：人間の医学史．東京：平凡社；1996.
9. 吉村昭．白い航跡．東京：講談社；1991．
10. Sackett DL, Rosenberg WM, Gray JAM, et al. Evidence based medicine: what it is and what it isn't. Brit Med J 1996; 312: 71-72.
11. 福井次矢．新しい診療概念：EBM．日医雑誌 2001; 125: 1178-1130.
12. 福原俊一．患者立脚アウトカム．日本医師会雑誌 2005; 133: 118-120.
13. Bailar JC, Mosteller F. Medical uses of statistics. Boston, the Massachusetts Medical Society; 1996.
14. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Nat Cancer Inst 1959; 22:719-748.
15. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966; 50: 163-171.
16. Cox DR. Regression models and life-tables. J R Stat Soc 1972; B34:187-220.
17. Laupacis J, Sackett DL, Roberts RS. An assessment of clinically useful measure of the consequence of treatment. N Engl J Med 1988; 318: 1728-1733.
18. Maetani S, Nakajima T, Nishikawa T. Parametric mean survival analysis in gastric cancer patients. Med Decis Making 2004; 24: 131-141.
19. Tan LB, Murphy R. Shifts in mortality curves saving or extending lives? Lancet 1999; 554: 1378-81.
20. 前谷俊三．ハザードに基づく生存データ群間比較の問題点．癌生時研誌 1984; 5: 47-51.
21. Gamel JW, Vogel RL, McLean IW. Assessing the impact of adjuvant therapy on cure rate for stage 2 breast carcinoma. Br J Cancer 1993; 68: 115-118.
22. 前谷俊三．Log-rank 検定は何を比較しようとするのか．癌生時研誌 1994; 14: 5-13.
23. Langlands AO, Pocock SJ, Kerr GR, et al. Longterm survival of patients with breast cancer: a study of the curability of the disease. Br Med J 1979; 2: 1247-1251.

24. Gamel JW, McLean IW, Greenberg RA. Interval-by-interval Cox model analysis of 3680 cases of intraocular melanoma shows a decline in the prognostic value of size and cell type over time after tumor excision. *Cancer* 1988; 61:574-579.
25. Maetani S, Nakajima T. 30- to 50-year follow-up of gastric cancer patients: are the results predictable 5 years after surgery? In Brennan MF, Karpeh MS, Jr eds. 4th International Gastric Cancer Congress. Bologna: Monduzzi Editore; 2001.
26. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J Roy Statist Soc* 1949; 11:15-53.
27. Bonenkamp JJ, Hermans J, Sasako M, et al. Extended lymph node dissection for gastric cancer. *N Engl J Med* 1999; 340: 908-914.
28. Gamel JW, Bonnadona G, Valagussa P, et al. Refined measurement of outcome for adjuvant breast carcinoma therapy. *Cancer* 2003; 97: 1139-1146.
29. Peto R, Pike MC, Armitage P, et al. Design and analysis of a randomized clinical trials requiring prolonged observation of each patient: II. analysis and examples. *r J Cancer* 1977; 35:1-39.
30. Wright JC, Weinstein MC. Gain in life expectancy from medical interventions: standardizing data on outcome. *N Engl J Med* 1998; 339: 280-286.
31. Haybittle JL. Life expectancy as a measurement of the benefit shown by clinical trials of treatment for early breast cancer. *Clin Oncol* 1996; 1998; 10: 92-94.
32. Gamel GW, McLean IW. A stable, multivariate extension of the log-normal survival model. *Computers And Biomedical Research* 1994; 27: 148-155.

Patient-centered cancer treatments and measure of survival benefit

Tenri Medical Statistical Group (representative: Syunzo Maetani)

Tenri Institute of Medical Research

The use of randomized controlled trials (RCT) has greatly improved the measurement of cancer treatment efficacy. However, an ideal measure of survival benefit from treatments has yet to be developed. Furthermore, with more patients participating in the medical decision-making process (patient-centered health care), the statistical information provided to them must be easily comprehensible, and tailored to the individual. We have examined three conventional survival statistics (the logrank test, hazard ratio and 5-year survival) and also parametric mean survival analysis. Each of the conventional statistics has inherent limitations such as failure to consider the size of an effect, ratio rather than difference in treatment effect between groups, excessive sensitivity to early deaths compared with late outcomes, confusion of patient outcome with disease outcome, and favoring palliative treatment over curative treatment. Although estimation of the mean survival time requires extrapolation of the survival curve beyond the follow-up period, it has been estimated with increasing accuracy and a number of advantages have followed. In particular, it is readily appreciated by patients, showing how many years of life are lost as a result of disease and how many are gained by treatment. It is able to distinguish between curative and palliative chemotherapy based on limited follow-up. However, parametric survival analysis is still in its developmental stage and needs further improvements in order that patients can select their best treatment options by themselves.

Keywords: cancer treatment, survival benefit, parametric model, mean survival, survival statistics