

予測精度に大きく影響する。boosting 数を多くすると、ツリーを構成する変数の数が増え、予測精度も上がるが、学習データだけの予測精度が向上し、他のデータへの汎用性が低下する過学習の可能性も高くなる。このため、学習データを n グループ (n は正数) に分割し、 $(n-1)$ グループのデータで予測モデルを構築し、学習に使わなかった残り 1 グループのデータを予測する n -fold クロスバリデーション(CV)を実施して、汎化能力を評価した。最後にクロスバリデーションで選択した boosting 数を用いて、全学習データを用いて学習モデルを作成し、同時に評価用データの予測も行った。CV 時の予測精度は、10 分割を 5 回実施し、平均値を用いた。boosting 数は 1 から 40 まで変化

させ、ツリーの予測精度としては転移在りを予測するカットオフ値 50% の ROC 下の面積(AUC)を用いた。SVM-FS による変数選択及び ADTree のモデルの構築には標準的な machine learning の手法が多く実装されているフリーウェアである weka[10, 11] (ver.3.5.8)を用いた。SVM-FS を調整する変数 complexity parameter は 1.0 (デフォルト値)を用いた。

C. 研究結果

C-1. リンパ節転移の予測モデル

図 1 にリンパ節転移に関して評価用データを最も高い精度で予測した ADTree の図 1 (SVM-FS の変数選択で 30 変数を選択し、

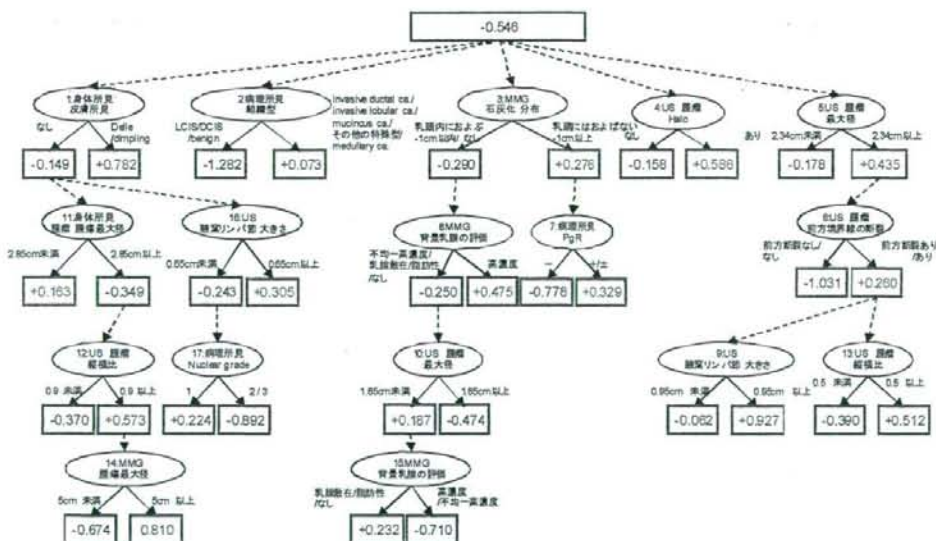


図 1 センチネルリンパ節転移を予測した ADTree。四角の中の数値は ADTree が予測するセンチネルリンパ節のスコアで、楕円はスコアを足し合わせる条件を表す。木の頂点のルートから全ての点線でつながった子ノードにたどり、また、実践でつながった子ノードは条件に合う子ノードのみをたどる。このようにしてたどり着いた四角の中 (スコアがプラスを赤、マイナスを青にしている) のスコアを全て足し合わせた結果、木全体での値がプラスであれば転移あり、マイナスであれば転移なしと予測することを意味する。さらにスコアの合計値の絶対値が 0 から離れるほど予測した状態の可能性が高いことを意味する。

boosting 数 17 で生成) を示す。図 2 に本モデルを用いて全学習データを予測したとき、別施設の評価データを予測したときの ROC 曲線を示す。また図 3 と図 4 に、それぞれ学習データと評価データに関して、本モデルが計算したスコアから転移の予測確率を示す。予測確率は、ADTree が出力するスコアの分布を 0~100% に線形スケールリングして計算した。

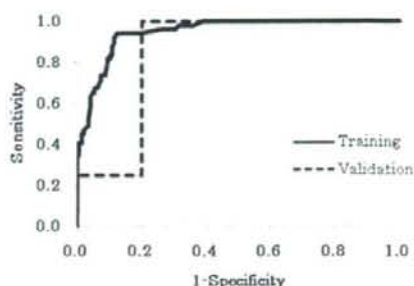


図 2 リンパ節転移の予測モデルの精度。ADTree の boosting 数とリンパ節転移ありを検出する ROC 以下の面積(AUC)との関係。実践は Full training は全ての学習データを用いて予測モデルを作成したときで、点線は評価用データを予測したときの ROC。SVM-FS での変数選択は 30、ADTree の boosting 数は 17 を採用。

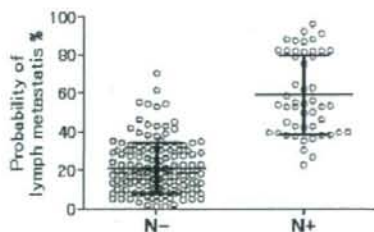


図 3 学習データでの予測確率の分布。

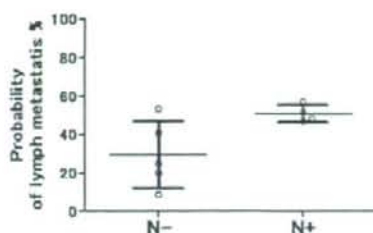


図 4 評価用データでの予測確率の分布。

図 1 の ADTree は、学習データから自動的にリンパ節転移に関連があると予測される因子を含む。ルートノードから子ノードをたどって、全てのスコアを足し合わせる。パスが点線の場合は全子ノードの計算を行い、実践の場合はノードの条件が一致する方の子ノードのみをたどることを繰り返し、全ての到達するノードのスコアを足し合わせる。最終的に木全体で算出したスコアがプラスであればリンパ節転移の可能性が高く、マイナスであればリンパ節転移の可能性が低いと計算する。例えば、ある患者の変数が以下の場合のスコア計算例を記載する。

- 1) 皮膚所見が dimpling: Node 1 (+0.782)
- 2) 組織型が invasive ductal ca.: Node 2 (+0.073)
- 3) MMG 石灰化分布なし、MMG 背景乳腺の評価なし、US 腫瘍最大径が 2.8cm: Node 3:(-0.290)→Node 8 (-0.250)→Node 10 (-0.474)
- 4) US 腫瘍 halo あり: Node 4 (+0.586)
- 5) US 腫瘍最大径 2.98cm、US 腫瘍前方境界線の断裂あり、US 腋窩リンパ節サイズ 13mm、US 腫瘍比率 1.03:Node

5(+0.435)→Node 6(+0.260)→Node
9(+0.927)と Node 13(+0.512)

となり、ルートノードの値(-0.546)と上記の値を足し合わせ、+2.015 となり、プラスの値であることから転移の可能性が高いと予測される。

構築した ADTree の変数間の依存関係を解釈すると以下のルールが得られる。

- 1) 皮膚所見に dele/dimpling がある場合、転移の可能性が高くなる。
dele/dimpling がない場合、身体所見での主流最大径が 2.85cm 未満の場合転移の可能性が高くなる。2.85cm 以上の場合、US 腫瘍の縦横比が 0.9 以上であれば転移の可能性が高くなり、さらに MMG での腫瘍最大径が 5cm 以上で転移の可能性が高くなる。
- 2) 組織型が LCIS/DCIS/良性の場合転移の可能性が低くなり、それ以外の場合は高くなる。
- 3) MMG 石灰化の分布で乳頭内 1cm に及ぶまたは石灰化がない場合、MMG での乳腺背景乳腺の評価が高濃度であれば転移の可能性が高くなる。高濃度以外では腫瘍最大径が 1.85cm 以下、MMG 背景乳腺の評価が乳腺散在/脂肪性で転移の可能性が高くなる。
MMG 石灰化の分布が乳頭に 1cm 以上及ばない場合、PgR が +/- で転移の可能性が高くなる。
- 4) US 腫瘍 halo がある場合転移の可能性が高くなる。
- 5) 腫瘍最大径が 2.34cm 以上の場合、US の前方境界線の断裂がある場合転移の可能性が高くなる。さらに、US 腋窩リ

ンパ節の大きさが 5mm 以上、US 腫瘍の縦横比が 1.03 以上だとさらに転移の可能性が高くなる。

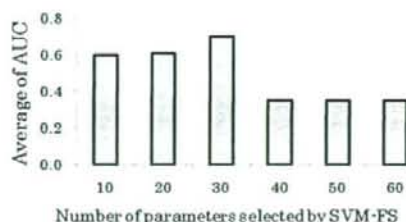


図 5 SVM-FS による選択した変数の数と予測精度の違い。横軸は変数の数、縦軸は予測精度 (ROC 曲線以下の面積で、バーが長いほど精度が高い) を示す。

図 5 に SVM-FS で選択した変数の数と、その変数を用いて構築した ADTree の予測精度の違いを示す。図より変数 30 までは変数の増加と共に徐々に精度が向上するが、40 を超えると極度に精度が落ちることが分かる。

図 6 に ADTree 以外の各手法の予測精度を示す。全てのデータを学習データとした場合と CV を行った場合 ADTree と C45 は類似した傾向、(前者で高い精度、後方で低い精度) を示したが、評価用データでは ADTree が高い値を示し、C45 は最も低い値を示した。Logistic 回帰と ANN は、全データ使用時、クロスバリデーション、評価用データと精度が減少の傾向にあり、SVM は全般的に低い予測精度であった。

以上のことより、SVM-FS と ADTree を組み合わせることによって作成したモデルでは、他の手法に比べて新規データに対して性能の劣化が小さく汎化能力の高い予測ができることがわかる。

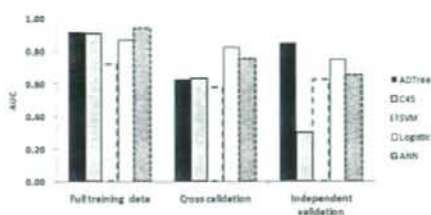


図6 手法による予測精度の違い。左は全てのデータを学習データとして用いた場合、中央は10-fold CVを実施した場合、右側は別施設の評価試験を行った場合の結果。それぞれの、左からADTree、C45、SVM、Logistic回帰、ANNの予測精度を示す。縦軸はAUCを示す。

C-2. 術前化学療法の効果予測モデル

術前化学療法の効果は、都立駒込病院と国立病院大阪医療センターから、アンストラサイクリン系(FEC)4クールに続きドセタキセル(DOC)4クールを投与された91症例の患者に対し、効果予測モデルの構築を行

った。

図7にADTreeの予測モデルを示す。また、図8に本モデルによって腫瘍サイズが25%以上、または25%未満と予測される確率の分布をクロスバリデーションで求めた結果を示す。図7のツリーより、ERと超音波での前方境界線の断裂が常に影響因子であることがわかる。ER+の場合、PgRの影響も加味される。一方、超音波の前方境界線の断裂がなく、haloがある場合は、nuclear gradeの影響が出てくる。その代わりに、haloがない場合は、石灰化の範囲が影響因子となる。その他の影響因子としては、乳頭腫瘍間距離、乳頭分泌など皮膚所見に関するものも多い。

予測精度は、クロスバリデーション時にAUC=0.75と高い(全症例を使用時はAUC=1.0となり全症例を分離できる)が、図より25%以上のサイズの症例は正確に予

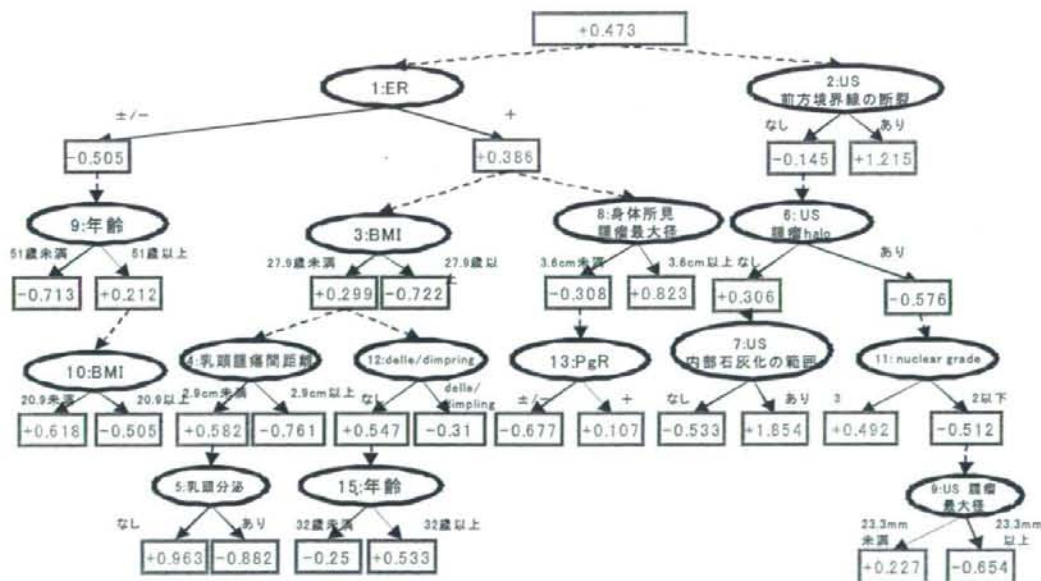


図7 術前化学療法による腫瘍縮小率(25%以上/未満)を計算するモデル。トータルスコアが、+の場合25%以上、-の場合25%未満と予測。

測できていることが多いが、25%以下の症例に関しては、逆に予測しているものも含まれる。つまり、効果がないことのほうが予測を行いやすい結果であった。

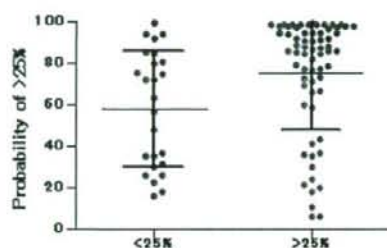


図 8 術前化学療法後、腫瘍サイズが 25%未満になる群と 25%以上になる群の予測。

また、腫瘍の縮小率を直接重回帰分析による予測モデルを構築した。図 9 に駒込病院の 36 症例、図 10 に国立病院大阪医療センターの 55 症例、図 11 にこれらを合わせた全 91 症例を用いて学習した結果を示す。それぞれ相関係数は $R=0.59$ 、 $R=0.67$ 、 $R=0.58$ となり、症例数は多くなるものの、異なる施設の症例を合わせた場合必ずしも相関係数が高くない。これらのモデルに使われている変数は、昨年度までに開発した、欠損値が多いデータセットでも変数選択が可能な CFS (Correlation-based Feature Selection) に、大域的最適化の手法である遺伝的アルゴリズム (GA: Genetic Algorithm) を組み合わせた方法 (以下、CFS+GA と記載する) で選択している。

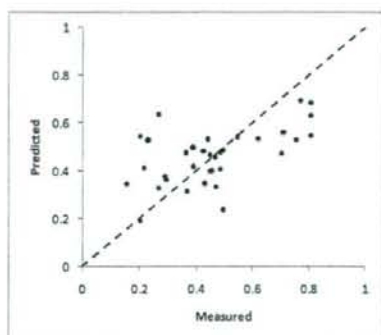


図 9 $n=36$ のデータを用いて予測した腫瘍縮小率

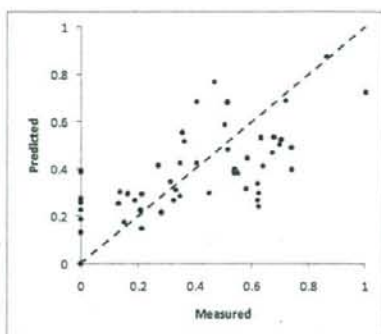


図 10 $n=55$ のデータを用いて予測した腫瘍縮小率

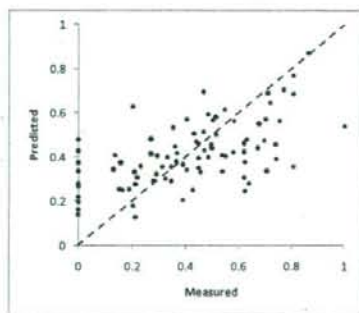


図 11 全データ ($n=91$) を用いて予測した腫瘍縮小率

表 1 腫瘍縮小率予測モデルで選択された変数

Training data		重回帰分析		ADTree	
		学習データ	評価データ	全データ	全データ
年齢					○
BMI	BMI	0.08	-0.29	-0.29	○
家族歴	家族歴	-0.18			
腫瘍最大径	腫瘍最大径	0.15	0.31	0.42	
腫瘍深度				-0.48	
della/dmp	della/dmp	0.14		0.16	
リンパ管浸潤			-0.38		
乳頭牽引				0.05	
乳頭腫瘍間質					○
乳頭分泌					○
MMG 横径の広れ			0.06		○
MMG 皮厚の異常所			0.15	0.11	
US 腫瘍最大径					○
US 乳腺石灰化の付	US 乳腺石灰化の付随	-0.47			
US リンパ門の有無	US リンパ門	-0.16	0.15		
US 内部エコー			0.51		
US 乳管内浸潤			0.14		
US 腫瘍形状				0.13	
US 血流信号				-0.13	
US 乳管内浸潤				0.13	
US 腫瘍 前方境界線の所取					○
US 腫瘍 halo					○
US 内部石灰化の有					○
ER	ER	-0.08	0.10	0.13	○
PgR			-0.27	-0.14	○
nuclear grade					○
組織型	組織型	0.23			○
定数項	定数項	0.87	-0.08	0.68	○

それぞれのモデルで選択された変数と、図 7 に示した ADTree で選択した変数を表 1 に示す。BMI は全モデルで選ばれており、これは単に欠損値が少ないこともが寄与している懸念もあるが、同様に欠損値のない年齢は全モデルで選択されているわけではなく、欠損値の少なさに関係なく影響因子としてきちんと選択されていると考えて妥当である。そのほか、PgR、nuclear grade、ER などは ADTree では全て包含しているが、CFS+GA の選択法では必ずしも含まれているわけではない。CFS+GA では腫瘍縮小率と相関が高く、お互いに相関の低い独立な変数セットを選択するために、腫瘍縮小率に重要と考えられる変数であっても、それらの変数間に相関係数が高ければどちらかが欠落する可能性もある。一方 ADTree では予測を 2 値でしか行えないため、縮小率を連続値として予測することはできないが、変数をそれぞれの依存関係も含めて選択することが出来ている。今回、縮小率の閾値として 25% を用いたが、Rouzier らは、

乳房保存療法が可能かどうかの境界線である 3cm を閾値としている[12]。しかし、単に、乳房保存療法適応可能性の予測だけでなく、縮小率を定量的に予測できれば、個人ごとに必要最小限の投薬量の計算など、よりきめ細かい医療の個別化を達成することが可能になる。このためには、より多種のレジメンの症例データから定量性の高いモデルを構築してゆく必要がある。また予測手法そのものも、現段階では、定量値の予測よりも、2 値予測の方が、従来から影響変数ではないかと考えられる重要変数を取り込むことに成功している。重回帰分析以外にニューラル・ネットワークなどの適応も考えられるが、本手法も同様に多重共線性の問題があり、入力として使える変数としては同じものである。このため、依存関係のあるような変数を入力変数として連続値を予測する手法の開発、または、予測対象とする症例を、予測する時点までに得られている情報で絞りこみを行い、それぞれグループ化された症例ごとに異なるモデルを構築して予測精度を向上させるなどの対策が必要となる。例えば、最初は全症例でモデルを構築し、ER が変数に入っていない場合、ER+ の症例群と ER- の症例群に対して個別にモデルを構築する。このようにして、問題を分割して、個々のモデルを構築するには、個々のモデルを構築するために十分な症例数が必要で、かつ新たなマーカーの導入などが実施された場合、マーカーの診断により分類が大きく変動するため、継続的に新しいデータを取り込んで、最新の予測モデルをアップデートしていくようシステム化を行う必要がある。

C-3. リンパ管侵襲の予測モデル

リンパ管侵襲は、術前化学療法を実際された症例と、実施されていない手術先行の症例に分けて予測モデルの構築を行った。予測モデルの構築方法は、リンパ節転移の予測モデルの構築と同じく、SVM-FS によって変数選択を行い、ADTree によって予測を行った。それぞれ駒込病院の症例を用い、全データを学習データとした場合と、10-fold クロスバリデーションを実施した場合を実施した。

図 12 に化学療法を実施した症例のリンパ管侵襲の予測モデルを示す。また、図 13 に本モデルを用いたときの ROC 曲線を示す。全データを学習したときの AUC は 0.95、クロスバリデーション時は 0.66 であった。

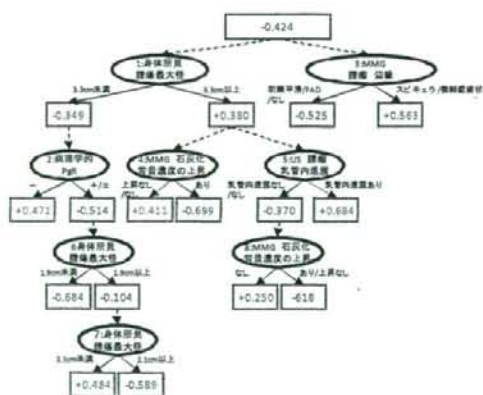


図 12 術前化学療法の症例に対するリンパ管侵襲予測モデル。
スコアが+の場合転位あり、-の場合転位なしと予測。(n=48)

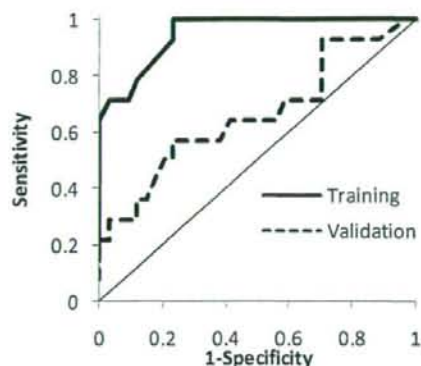


図 13 術前化学療法の症例に対するリンパ管侵襲予測モデルの ROC 曲線。Training は全症例を学習データとした場合で、Validation はクロスバリデーションによって得られた結果。

また、図 14 に手術先行症例の予測モデル、図 15 には ROC 曲線を示す。手術先行の場合、全データを学習データとしたときの AUC は 0.97、クロスバリデーション時は 0.62 であった。

術前化学療法時と手術先行時では、手術先行時の方がツリーのサイズが大きくなり、より多くの影響変数が取り込まれている。術前化学療法時では、腫瘍最大径と、マンモグラフィ(MMG)での腫瘍辺縁が常に影響因子となっており、PgR、石灰化による背景濃度の上昇、乳管内進展などが含まれる。手術先行時では、センチネルリンパ節転移の結果、組織型、背景乳腺の評価、delle/dimpli.、MMG での構築の乱れなどが含まれる。

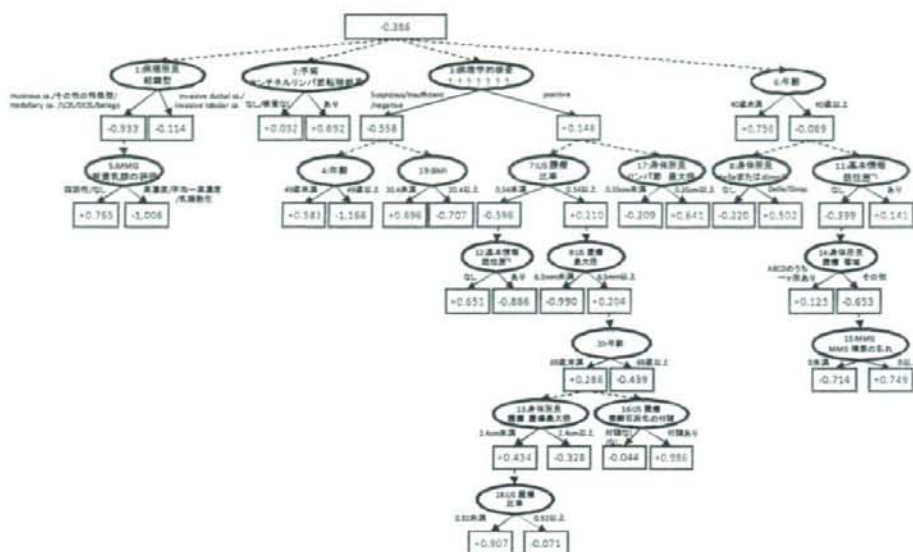


図 14 手術先行の症例に対するリンパ管侵襲予測モデル。スコアが+の場合転位あり、-の場合転位なしと予測。(n=172)

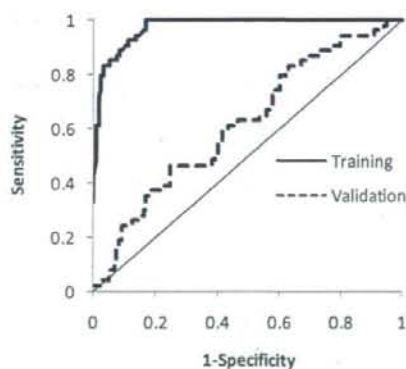


図 15 手術先行の症例に対するリンパ管侵襲予測モデルの ROC 曲線。Training は全症例を学習データとした場合で、Validation はクロスバリデーションのよって得られた結果。

C-4. 静脈侵襲の予測モデル

静脈侵襲の予測は、リンパ管侵襲のモデルと同様に、術前化学療法を実施した症例と手術先行の症例に分けてモデルの構築を試みた。しかし、既に構築していたデータベ

ース内で手術先行症例では、静脈侵襲の値の欠損値などから症例数が少なく、クロスバリデーション時に AUC が 0.5 に達成するモデルが構築できなかったため、ここでは術前化学療法の症例の結果のみを示す。

図 16 に予測モデル、図 17 に ROC 曲線の図を示す。症例数は 172 症例で、全データを学習データとして用いた場合、AUC は 0.91、クロスバリデーション時 0.76 と高い予測精度を得た。

予測因子としては、MMS の腫瘍中心濃度、nuclear grade、リンパ節の数が常に関係している。これらの変数に依存して、さらに PgR、腫瘍領域、形状、MMG 構築の乱れ、多発など、腫瘍そのものの特徴に関連する変数が多い結果となった。

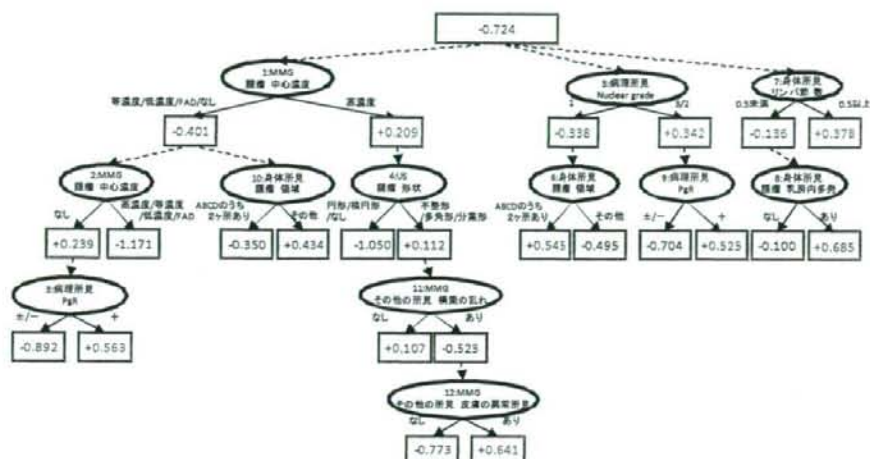


図 16 術前化学療法の症例に対する静脈侵襲予測モデル。スコアが+の場合侵襲あり、-の場合侵襲なしと予測。(n=172)

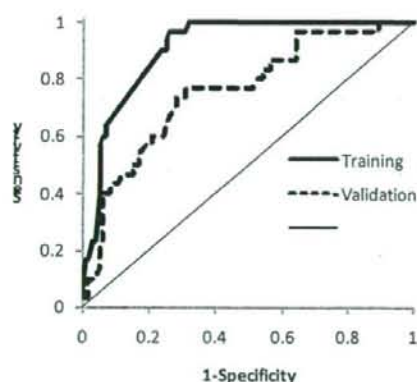


図 17 術前化学療法の症例に対する静脈侵襲予測モデルの ROC 曲線。Training は全症例を学習データとした場合で、Validation はクロスバリデーションによって得られた結果。

D. 考察

本研究では、術前療法に焦点を絞り、診療の意思決定に大きく関与するリンパ節の転移、術前化学療法の効果など予測するためのモデルの構築を行ってきた。

類似する研究例として、腋窩リンパ節転移の予測に関しては、医学的知見から選択された変数から多変量解析[13, 14]や変数を組み合わせるスコアリングの方法[15]を用いて予測される研究が行われてきた。また、センチネルリンパ節(SLN)に転移がある患者に対して、非センチネルリンパ節(non-SLN)の予測する研究例も同様の方法の開発[16-22]とその評価報告[23-27]が多い。しかし、このような方法では単変数解析でリンパ節転移の有無との直接の相関が高いものが選ばれるが、本来相関が低いと考えられており既存のモデルに用いられて来なかった変数や、他の変数との階層的または条件分的な組み合わせなど、複雑な影響因子や、その組み合わせによるルールをモデルに反映させることができない。情報工学的な視点からは、患者から得られる病態、家族歴、各種検査結果など極力多くの情報から、機械学習を用いて客観的に変数を選択し、予測モデルを作成すれば、精度

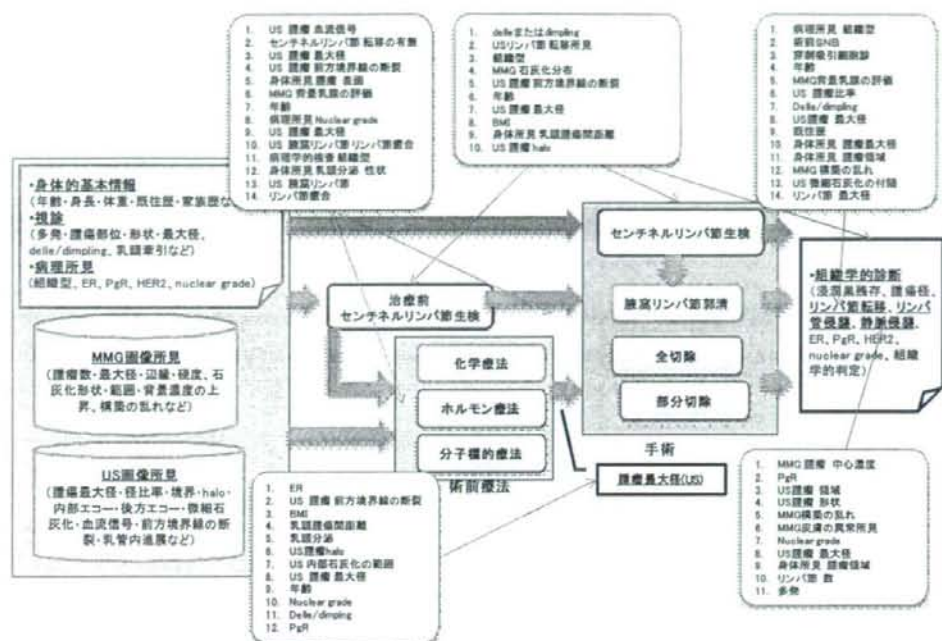


図 18 乳癌の術前療法における診療のフローと予測モデルの影響変数の一覧

の向上や従来知られていなかった変数が探索できる可能性があると考えられる。

術前化学療法についても、NeoAdjuvant Online[12]など、Nomogramを用いたモデルの研究例があるが、少ない変数で腫瘍サイズが一定の大きさ以下になる確率を計算するものである。リンパ節転移のモデル同様により多くの変数を考慮し、かつ、直接予測する変数との相関は低くとも、複合的に複数の変数からルールを抽出できることが重要である。このような目的だけでなく、本研究で扱うデータの特徴として欠損値が多いことにも対処できる方法としてADTreeによるモデルの開発を行ってきたが、これにSVM-FSによる変数選択を組み合わせて予測精度、汎化能力ともに高いモデルを開発することができた。本方法をリンパ節転移の予測と術前化学療法の効果予

測以外に適応したリンパ管侵襲[28-30]と静脈侵襲[31]は、どちらも非センチネルリンパ節の転移の予測因子であるため、これらの変数が術前に正確に予測することができれば、手術における切除範囲など意思決定支援にツールになり得る。

図 18 に各予測モデルに含まれている影響変数を列挙した。従来の診療をこれらの低侵襲に置き換えることができれば個人への負担も低下できるだけでなく、診療にかかるコストを低下させられることも可能である。しかし、図 18 に示した通り、各ルールに含まれる変数だけでも数が多く、これらをツリー上で示すと、複雑な構造になるため、全体の俯瞰と各症例の病態がその中でどのようなルールに当てはまりどのように予測されるかを可視化するためのシステム化が必須である。また、学習データは適

宜最新のものにアップロードし、自動的に学習モデルを再学習させる機能を実装し、新しい診療ガイドラインの変遷や新しいメーカーの導入にも耐えうる構造を持たせることが必要となる。本プロジェクトでは方法論の開発が中心として進め、実際にリンパ節転移などで高精度な予測に成功はしたが、これらシステム化によって実用化が可能であると考えられる。

E. 研究発表

研究発表

論文発表

1. Sato S, Arita M, Soga T, Nishioka T, Tomita M.,
Time-resolved metabolomics reveals metabolic modulation in rice foliage. *BMC Syst Biol*. 2008;2:51.
2. Shinoda K, Tomita M, Ishihama Y.,
Aligning LC peaks by converting gradient retention times to retention index of peptides in proteomic experiments. *Bioinformatics*. 2008;24(14):1590-5.
3. Ohno H, Naito Y, Nakajima H, Tomita M.,
Construction of a biological tissue model based on a single-cell model: a computer simulation of metabolic heterogeneity in the liver lobule. *Artif Life*. 2008;14(1): 3-28.
4. Shinoda K, Sugimoto M, Tomita M, Ishihama Y.,
Informatics for peptide retention properties in proteomic LC-MS. *Proteomics*. 2008;8(4):787-98.

5. 内藤泰宏, 細胞・組織シミュレーションの現状 薬効・毒性の予測にむけて *ファルマシア*, 44(9), 885-889, 2008

F. 知的財産権の出願・登録状況
なし

参考文献

1. Carter, C.L., C. Allen, and D.E. Henson, *Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases*. *Cancer*, 1989. 63(1): p. 181-7.
2. Gipponi, M., et al., *Sentinel lymph node as a new marker for therapeutic planning in breast cancer patients*. *J Surg Oncol*, 2004. 85(3): p. 102-11.
3. Goldhirsch, A., et al., *Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007*. *Ann Oncol*, 2007. 18(7): p. 1133-44.
4. Ge, G. and G.W. Wong, *Classification of premalignant pancreatic cancer mass spectrometry data using decision tree ensembles*. *BMC Bioinformatics*, 2008. 9: p. 275.
5. Ayers, M., et al., *Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer*. *J Clin Oncol*, 2004. 22(12): p. 2284-93.

6. Freund, Y. and L. Mason. *The alternating decision tree learning algorithm*. in *Proceeding of the Sixteenth International Conference on Machine Learning*. 1999. Bled, Slovenia.
7. Liu, K.Y., et al., *Boosting alternating decision trees modeling of disease trait information*. BMC Genet, 2005. 6 Suppl 1: p. S132.
8. Zhou, X., J. Xu, and Y. Zhao, *Machine learning methods for anticipating the psychological distress in patients with Alzheimer's disease*. Australas Phys Eng Sci Med, 2006. 29(4): p. 303-9.
9. Qu, Y., et al., *Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients*. Clin Chem, 2002. 48(10): p. 1835-43.
10. Witten, I.H. and E. Frank, *Data mining: practical machine learning tools and techniques*. 2nd ed. Morgan Kaufmann series in data management systems. 2005, Amsterdam; Boston, MA: Morgan Kaufman. xxxi, 525 p.
11. Frank, E., et al., *Data mining in bioinformatics using Weka*. Bioinformatics, 2004. 20(15): p. 2479-81.
12. Rouzier, R., et al., *Breast-conserving surgery after neoadjuvant anthracycline-based chemotherapy for large breast tumors*. Cancer, 2004. 101(5): p. 918-25.
13. Olivotto, I.A., et al., *Prediction of axillary lymph node involvement of women with invasive breast carcinoma: a multivariate analysis*. Cancer, 1998. 83(5): p. 948-55.
14. Promish, D.I., *Prediction of axillary lymph node involvement of women with invasive breast carcinoma: a multivariate analysis*. Cancer, 1999. 85(5): p. 1201-3.
15. Laura, S., et al., *Tumour size as a predictor of axillary node metastases in patients with breast cancer*. ANZ J Surg, 2006. 76(11): p. 1002-6.
16. Barranger, E., et al., *An axilla scoring system to predict non-sentinel lymph node status in breast cancer patients with sentinel lymph node involvement*. Breast Cancer Res Treat, 2005. 91(2): p. 113-9.
17. Kohrt, H.E., et al., *New models and online calculator for predicting non-sentinel lymph node status in sentinel lymph node positive breast cancer patients*. BMC Cancer, 2008. 8: p. 66.
18. Sachdev, U., et al., *Predictors of nonsentinel lymph node metastasis in breast cancer patients*. Am J

- Surg, 2002. 183(3): p. 213-7.
19. Rouzier, R., et al., *Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer*. J Clin Oncol, 2005. 23(33): p. 8331-9.
 20. Van Zee, K.J., et al., *A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy*. Ann Surg Oncol, 2003. 10(10): p. 1140-51.
 21. Pal, A., et al., *A model for predicting non-sentinel lymph node metastatic disease when the sentinel lymph node is positive*. Br J Surg, 2008. 95(3): p. 302-9.
 22. Katz, A., et al., *Nomogram for the prediction of having four or more involved nodes for sentinel lymph node-positive breast cancer*. J Clin Oncol, 2008. 26(13): p. 2093-8.
 23. Lambert, L.A., et al., *Validation of a breast cancer nomogram for predicting nonsentinel lymph node metastases after a positive sentinel node biopsy*. Ann Surg Oncol, 2006. 13(3): p. 310-20.
 24. Ponzzone, R., et al., *Comparison of two models for the prediction of nonsentinel node metastases in breast cancer*. Am J Surg, 2007. 193(6): p. 686-92.
 25. Soni, N.K., et al., *Evaluation of a breast cancer nomogram for prediction of non-sentinel lymph node positivity*. Eur J Surg Oncol, 2005. 31(9): p. 958-64.
 26. Specht, M.C., et al., *Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram*. Ann Surg Oncol, 2005. 12(8): p. 654-9.
 27. Klar, M., et al., *The MSKCC nomogram for prediction the likelihood of non-sentinel node involvement in a German breast cancer population*. Breast Cancer Res Treat, 2008.
 28. Anan, K., et al., *Axillary lymph node metastases in patients with small carcinomas of the breast: is accurate prediction possible?* Eur J Surg, 2000. 166(8): p. 610-5.
 29. Ito, M., et al., *Significance of pathological evaluation for lymphatic vessel invasion in invasive breast cancer*. Breast Cancer, 2007. 14(4): p. 381-7.
 30. Cetintas, S.K., et al., *Factors influencing axillary node metastasis in breast cancer*. Tumori, 2006. 92(5): p. 416-22.
 31. Weigand, R.A., et al., *Blood vessel invasion and axillary lymph node involvement as prognostic indicators for human breast cancer*. Cancer, 1982. 50(5): p. 962-9.

原発性乳癌の治療に関連するバイオマーカーの探索

主任研究者 戸井 雅和
京都大学医学部付属病院 乳腺外科教授
分担研究者 黒井 克昌
東京都立駒込病院 臨床試験科・外科部長

研究要旨

研究の開始から2年間にわたるプロテオミクス手法を用いた原発性乳癌患者血清でのバイオマーカー探索を行ってきた。

バイオマーカーとして候補となるいくつかのタンパク質、ペプチドピークを SELDI システムによって得られたので、その精製・同定を試みた。

1. 術前ホルモン療法応答性に関するバイオマーカーの探索
2. 非浸潤性乳癌に関するバイオマーカーの探索について報告する。

1. ホルモン療法応答性に関するバイオマーカーの探索

A-1 研究目的

プロテオミクス手法で探索絞り込みされたホルモン療法応答性原発性乳癌マーカー候補の同定を目的とする。

B-1 研究方法

B-1-1 対象

術前ホルモン療法臨床試験 CAAN スタディにおけるアロマターゼ阻害剤術前投与患者よりインフォームドコンセントが得られ研究可能な乳癌症例。

術前ホルモン療法前、ホルモン療法開始後3日、術前ホルモン療法終了時（治療開始後約12週）、手術終了後8～9週の血清試料を使用。

- ・Letrozole 治療群血清13例 50%奏効で群分けし奏効6例、非奏効7例
- ・Exemestane 治療群血清9例 50%奏効で群分けし奏効5例、非奏効4例

B-1-2 サンプル測定

適正な採取・保存を行った血清サンプルを Urea 変性後、ペプチド、タンパク質の PI 値による性質を利用した陰イオン交換樹脂での分離を4段階で行ったものを陽イオン交換プロテインチップ [CM10]（結合・洗浄条件 pH4、pH7）、金属イオン（銅イオン）修飾チップ [IMAC30] に添加した。
エネルギー吸収分子として SPA（シナピン

酸）を添加した。

ProteinChip リーダーを使用してタンパク質やペプチドの質量数と相対発現量について飛行時間型質量分析計測定を行った。

B-1-3 データ解析

データ解析は Protein Chip Data Manager version 3.0 を用いて行った。質量数校正、イオン計測正規化処理を測定条件毎に実施し、同一の正規化処理を行ったデータ群間で比較解析を実施した。

B-1-4 バイオマーカー候補の抽出

解析対象はシグナル・ノイズ比 (S/N) > 2.5 以上のピークとし、Permutation 検定、ロジスティック回帰分析、経時変動パターン解析により総合的に、また特異性により判断、抽出した。

B-1-5 二年に亘る解析結果でのマーカー候補比較再現性

Letrozole 治療群血清13例 と Exemestane 治療群血清9例での重複マーカー候補を対象に分析を実施する。Permutation 検定 ($p < 0.05$) で絞り込まれたピークを変動パターンによる分類で特徴づけ、これを基に候補選択を行った。
経時変動パターンは薬剤投与前 (d0)、投与3日後 (d3)、手術前 (op)、術後35

日 (op35) のそれぞれの時点で0 (2群間に有意差なし) と1 (2群間に有意差あり) で表現する。

・薬剤投与前には有意差がなく、投与直後から奏効群-非奏効群間の有意差が手術時まで持続するパターン (0110、0111) : 治療の効果を判定する指標になるマーカーの候補。

・薬剤投与前に有意差があり、手術時には有意差がなくなるパターン (1000、1001) : 治療の奏効性を予測する指標になるマーカーの候補。

・薬剤投与前に有意差があり、投与後手術時まで有意差が持続するパターン (1110、1111) : 治療の奏効性を予測し同時に効果の判定も行えるマーカーの候補。

2006年度の候補ピーク (Letrozole 治療群血清より得られたもの) に対して、2007年度の解析結果 (Exemestane 治療群血清より得られたもの) 中に近似分子量 (m/z) のピークが29個あり、この中から①類似の変動パターンを示すもの、②分子量の近似性、によってマーカー候補を3つまで絞った。

第一候補 m/z 9713
Letrozole における変動パターン 1000
Exemestane における変動パターン 1011

第二候補 m/z 26517
Letrozole における変動パターン 1001
Exemestane における変動パターン 1000

第三候補 m/z 3252
Letrozole における変動パターン 0110
Exemestane における変動パターン 0010

上記第一候補の精製・同定を試みる。

B-1-6 マーカー候補の精製

陽イオン交換樹脂を想定したCM10プロテインチップと、陰イオン交換樹脂を想定したQ10プロテインチップへの吸着条件がイオン交換樹脂への吸着条件を、ま

た塩濃度によるプロテインチップからの溶出条件がイオン交換樹脂からの溶出条件をシミュレーションできることを利用して、精製条件を検討した。

本シミュレーション結果に基づき、目的のピークの相対量が多いサンプル血清を使用し、目的のピークができるだけ多く残り、別のペプチド、タンパク質が除かれるよう精製を進めた。

さらにTricine-SDS PAGEを行いゲル上に目的タンパク質を分離した。

Serum 150 μ l
↓
Q spin column
↓ 33.3% Isoplopanol/
↓ 16.7% Acetonitrile /0.1% TFA

C4 column COSMOSIL 4.6×150mm 5C8-MS
↓ Acetonitrile 40%
↓
Tricine-SDS-PAGE 16.5%

A-2-7 マーカー候補の同定

Tricine-SDS-PAGEによって分離された目的ピーク分子量付近のゲル中に目的ピークのタンパク質が含まれているかどうかを、ゲルからタンパク質を抽出後にプロテインチップで確認した上で、同部分のゲルをLC-MS/MS解析して同定を試みた。目的のピークが含まれたゲル部分をTrypsin消化し、LC-MS/MS装置によって測定解析した。

断片化されたペプチドは、アミノ酸の配列によって固有の組み合わせ (分子量) を持つことから、多くの消化断片データを解析ソフトに入力することでゲル中に存在するペプチド、タンパク質名称候補が推定できる。

タンパクは精製後も単離はできていないこと、プロテインチップシステムで確認できない (イオン化されない) ペプチド、タンパク質を含むこと、分解されたタンパク断片が存在すること、これらの理由から、目的のピークタンパク質を推定するために、精製等で得られたピークの性質 [分子量、等電点 (PI)、親水性、疎水性] 等によって選択した。

B-1-8 倫理面への配慮

本研究に使用した患者血清は全て患者本人の研究使用への承諾を受けたものであり、別の目的での使用は行わない。ま

た患者の個人識別ができないよう検体番号の暗号化等の配慮を行った。

C-1. 研究結果

C-1-1 精製ピークの Exemestane 治療群血清 9 例の m/z9713 についてのデータ分布

奏効群 (E)、非奏効群 (U) と時間経過薬剤投与前 (d0)、投与 3 日後 (d3)、手

術前 (op), 術後 35 日 (op35)

変動パターン 1 0 1 1 : 治療の奏効性を予測する指標になるマーカーの候補となる。

奏効群として振り分けた 92 番のサンプルは奏効群のはずれ値となっているが、奏効率は 50% でありカットオフのボーダーラインとなっているので、非奏効群としての扱いの適応も考えられる。(図 1)(図 2)

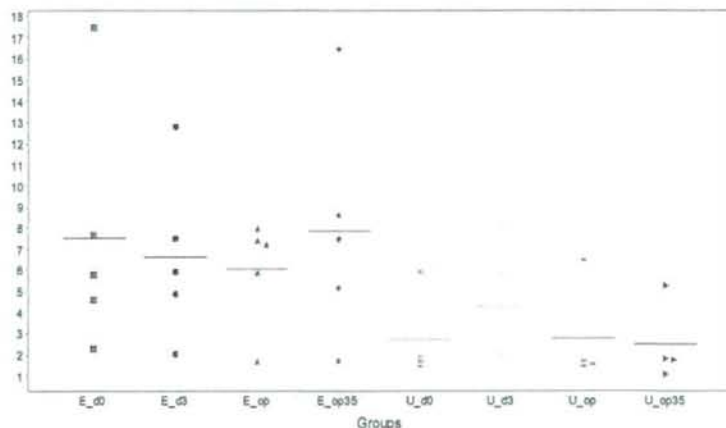
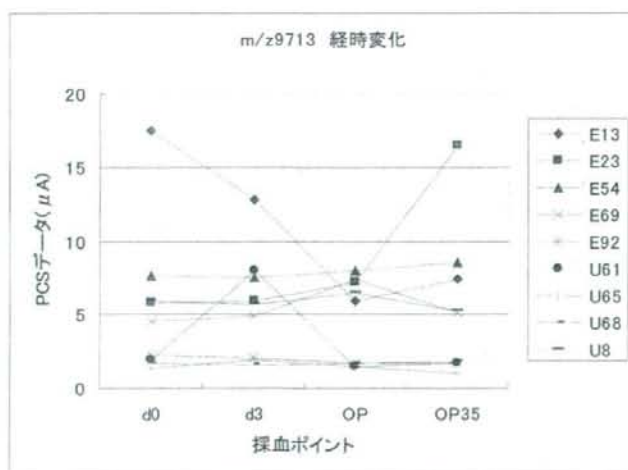


図 1 Exemestane 術前補助療法患者血清の ProteinChip リーダー測定経時変化効果群別分布 m/z9713



E: 奏効群 U: 非奏効群

図 2 Exemestane 術前補助療法患者血清の ProteinChip リーダー測定経時変化 m/z9713

C-1-2 精製過程

血清からQ スピンカラム（陰イオン交換）で有機溶媒画分に溶出したものをHPLCのC4カラムでさらに分離した。

(図3)

C-1-2 精製過程

Qカラム精製後COSMOSILでフラクション分離した。

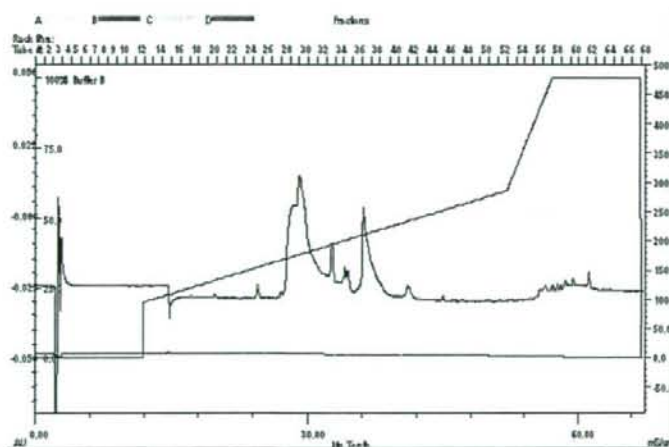


図3 HPLC分離像

フラクションに含まれるタンパク質、ペプチドをProteinChipリーダーで確認した。(図4)

目的のピークが含まれているフラクションを精製サンプルとした。

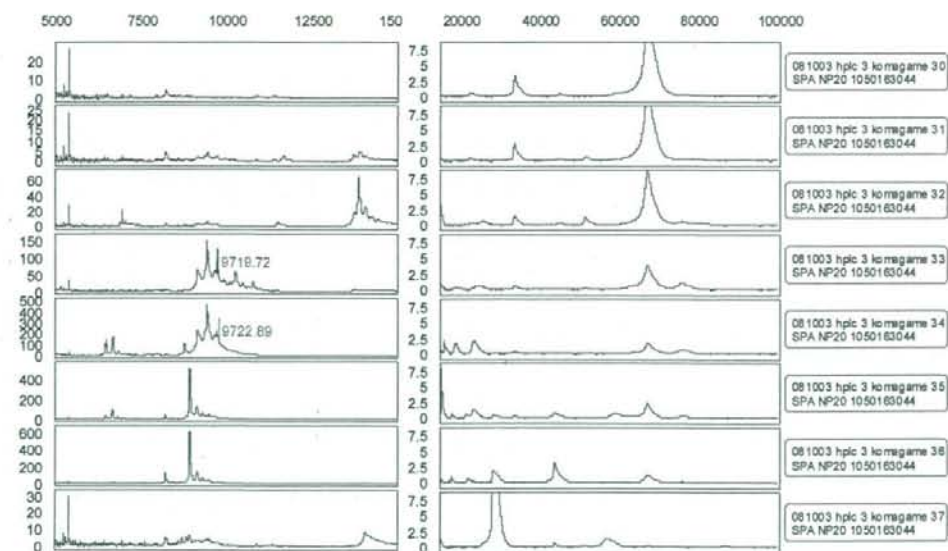


図4 ProteinChipリーダー測定 HPLC フラクション

Tricine Buffer を使用して低分子用電気泳動 (Tricine-SDS-Page) を精製サンプルに対して行った。

泳動ゲルにバックグラウンドが白く染まる Negative Gel Stain を施し染色され

ない部分を切り出した。(図5) 切った部分に目的のタンパク質が含まれていることを ProteinChip リーダー測定で確認した。(図6)

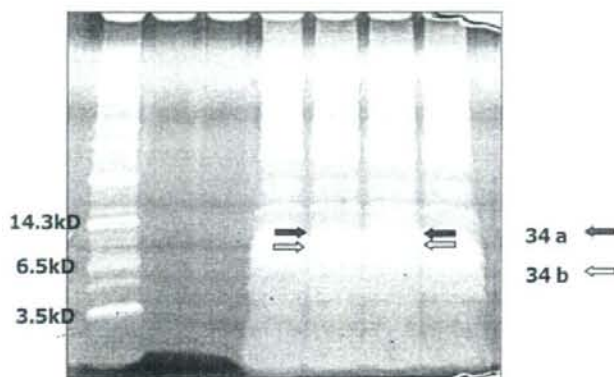


図5 Tricine-SDS-Page 染色ゲルとゲルカット位置

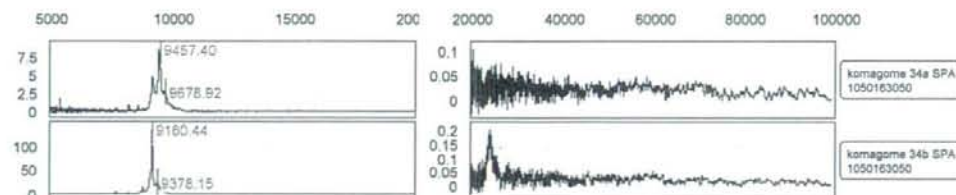


図6 切り出したゲルから抽出した画分の ProteinChip リーダー測定 上 (34a) 下 (34b)

C-1-3 同定

ゲルをプロテアーゼ消化し LC-MS/MS 解

析により得られたペプチド、タンパクの候補から目的分子名を推定した。

条件
同定した領域が m/z 9740 以下のもの
血清にあるもの
m/z 9742 付近にあるもの (内部補正)
34b のみで検出される物を除く
34b に検出される物を除く
プロセッシングによって m/z がマッチするもの

表1. LC-MS/MS 結果による目的タンパクの推定条件1

以上のような選択作業の結果、候補として13のタンパク質が選択された。

現在抗体を使用して同定を進めている。

2. 非浸潤性乳癌に関するバイオマーカーの探索

A-2. 研究目的

プロテオミクス手法で探索絞り込みされた非浸潤性乳癌マーカー候補の同定を目的とする。

B-2 研究方法

B-2-1 対象

非浸潤性の乳癌と診断された治療前の患者でインフォームドコンセントが得られた20例の血清(b群)、対象群として浸潤性乳癌血清18例(a群)、良性乳腺疾患血清10例(c群)。

治療前1ヶ月以内(T1)、治療(手術)後3週間(T2)、治療(手術)後3ヶ月(T3)の3ポイントでの採血を行った。

B-2-2 サンプル測定

適正な採取・保存を行った血清サンプルをUrea変性後、ペプチド、タンパク質のPI値による性質を利用した陰イオン交換樹脂での分離を4段階で行ったものを陽イオン交換プロテインチップ[CM10](結合・洗浄条件 pH4)、逆相プロテインチップ[H50](結合・洗浄条件 pH7)、金属イオン修飾プロテインチップ[IMAC30](銅イオン)に添加した。

エネルギー吸収分子としてCM10、IMAC30にはSPA(シナピン酸)を、H50にはCHCAを添加した。

ProteinChipリーダーによりタンパク質やペプチドの質量数と相対発現量について飛行時間型質量分析計測定を行った。

B-2-3 データ解析

データ解析はProtein Chip Data Manager version 3.0を用いて行った。質量数校正、イオン計測正規化処理を測定条件毎に実施し、同一の正規化処理を行ったデータ群間で比較解析を実施した。

B-2-4 バイオマーカー候補の抽出

解析対象はシグナルノイズ比(S/N) > 2.5以上のピークとし、Permutation検定、ロジスティック回帰分析、経時変動パターン解析により総合的に、また特異性により判断、抽出した。

経時変動パターンは、治療前1ヶ月以内、治療(手術)後3週間、治療(手術)後3ヶ月をそれぞれ0(各2群間に有意差なし)と1(各2群間に有意差あり)で表現する。

マーカーとして有用なパターンとして以下が考えられる。

- ・ 治療(手術切除)前に有意差があり治療後有意差のなくなるパターン(100、110): 診断のマーカーとなり得る。
- ・ 治療(手術切除)前に有意差がなく治療後有意差が出るパターン(011): 治療効果判定のマーカーとなり得る。

これらを考慮に入れ、マーカー候補を3つに絞った。

第一候補 m/z10250 (変動パターン 100)
第二候補 m/z4643 (変動パターン 110)
第三候補 m/z4658 (変動パターン 110)

第一候補のm/z10250ピークの抽出、同定を行った。

B-2-5 マーカー候補の精製

目的ピークの高い患者血清を用いて、競合するペプチド、タンパク質をできるだけ取り除く。

Serum 50 μ l

↓

Agilent カラムを用いて高含有量タンパク除去

↓

Q spin column

↓ (pH9 Buffer でスルー)

Tricine-SDS-PAGE 16.5%

B-2-6 マーカー候補の同定

Tricine-SDS-PAGEによって分離された目的ピーク分子量付近のゲル中に目的ピークのタンパク質が含まれているかを、ゲルからタンパク質を抽出し、プロテインチップで確認した上で、同部分のゲル

をLC-MS/MS解析して同定を試みた。目的のピークが含まれたゲルをTrypsin消化し、LC-MS/MS装置によって解析した。

断片化されたペプチドはアミノ酸の配列によって固有の組み合わせ（分子量）を持つことから、消化断片データを解析ソフト（UniProt/SwissProt）に入力することでゲル中に存在するペプチド、タンパク質名称候補が推定できる。

タンパク質は精製後も単離はできていないこと、プロテインチップシステムで確認できない（イオン化されない）ペプチド、タンパク質を含むこと、分解されたタンパク断片が存在すること、これらの理由から、目的のピークタンパクを推定するために、精製等で得られたピークの性質[分子量、等電点 (PI)、親水性、疎水性]等によって選択した。

B-2-7 倫理面への配慮

本研究に使用した患者血清は全て患者

- a 群：浸潤性乳癌 T1: 治療前1ヶ月以内
 b 群：非浸潤性乳癌 T2: 治療（手術）後3週間
 c 群：良性乳腺疾患 T3: 治療（手術）後3ヶ月

本人の研究使用への承諾を受けたものであり、別の目的での使用は行わない。また患者の個人識別ができないよう検体番号の暗号等の配慮を行った。

C-2 結果

C-2-1 精製ピークの3群における経時変化分布

浸潤性乳癌群と非浸潤性乳癌群、良性乳腺疾患群を完全に区別するバイオマーカー候補ピークは抽出されなかったが、有意差のある特徴的なピーク (m/z10250) をマーカー候補として特定した。

浸潤性乳癌群のデータ分布は限局された範囲に留まるが、非浸潤性乳癌群、良性乳腺疾患群では優位に高値のデータが散見され（図7）、また手術で患部を切除した後に低値に移行する挙動を示した。（図8）

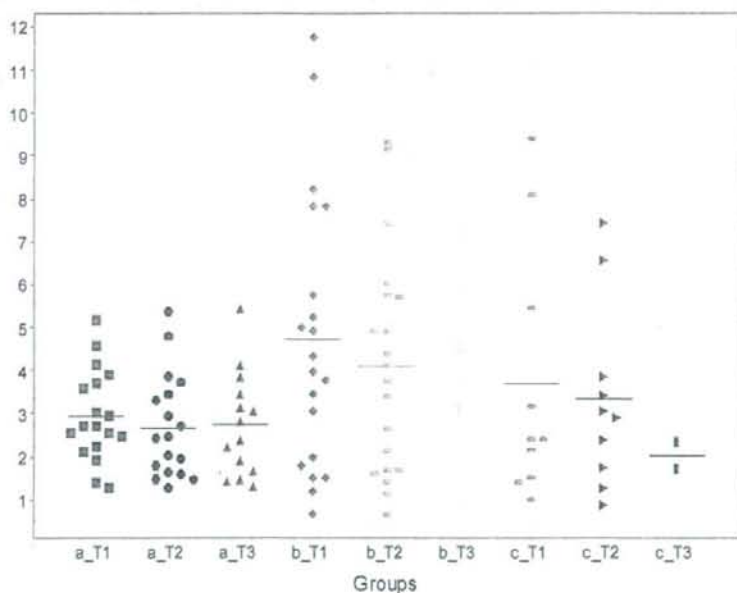


図7 非浸潤性乳癌患者血清のProteinChipリーダー測定経時変化効果群別分布 m/z10250