of CART over traditional regression models is that it can identify prognostic subgroups that are useful in clinical practice. Because the results of CART analysis are presented as a decision tree, which is intuitive, they can be readily interpreted by medical professionals without any specific knowledge of statistics. The most important consideration is that five variables used in the decision tree were clinical parameters that are readily available by the usual work-up of patients before therapy. Especially, glucose, GGT and LDL-C are simple biochemical markers that are easily measured at a low cost. Using this model, we can rapidly develop an estimate of the response before treatment, which may facilitate clinical decision making.

In conclusion, we built a pre-treatment model for the prediction of virological response in PEG-IFN plus RBV therapy. Because this decision tree model was made up of simple host factors such as steatosis, LDL-C, age, blood sugar and GGT, it can be easily applied to clinical practice. This model may have the potential to support decisions in patient selection for PEG-IFN plus RBV therapy based on the possibility of response against a potential risk of adverse events or costs, and may provide a rationale for treating metabolic factors to improve the efficacy of antiviral therapy.

ACKNOWLEDGEMENTS

THIS STUDY WAS supported by a grant-in-aid from Ministry of Health, Labor and Welfare, Japan. There exist no conflicts of interest.

REFERENCES

1 Strader DB, Wright T, Thomas DL, Seeff LB. Diagnosis, management, and treatment of hepatitis C. *Hepatology* 2004; **39**: 1147–71.
2 Fried MW, Shiffman ML, Reddy KR *et al.* Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N Engl J Med* 2002; **347**: 975–82.
3 Manns MP, McHutchison JG, Gordon SC *et al.* Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *Lancet* 2001; **358**: 958–65.
4 Davis GL, Wong JB, McHutchison JG, Manns MP, Harvey J, Albrecht J. Early virologic response to treatment with peginterferon alfa-2b plus ribavirin in patients with chronic hepatitis C. *Hepatology* 2003; **38**: 645–52.
5 Lee SS, Ferenci P. Optimizing outcomes in patients with hepatitis C virus genotype 1 or 4. *Antivir Ther* 2008; **13** (Suppl 1): 9–16.
6 Breiman LJH, Friedman RA, Olshen CJ, Stone CM. *Classification and Regression Trees*. Calif: Wadsworth, 1980.
7 Averbook BJ, Fu P, Rao JS, Mansour EG. A long-term analysis of 1018 patients with melanoma by classic Cox regression and tree-structured survival analysis at a major referral center: Implications on the future of cancer staging. *Surg* 2002; **132**: 589–602.
8 Garzotto M, Beer TM, Hudson RG *et al.* Improved detection of prostate cancer using classification and regression tree analysis. *J Clin Oncol* 2005; **23**: 4322–9.
9 Zlobec I, Steele R, Nigam N, Compton CC. A predictive model of rectal tumor response to preoperative radiotherapy using classification and regression tree methods. *Clin Cancer Res* 2005; **11**: 5440–3.
10 Jin H, Lu Y, Harris ST *et al.* Classification algorithms for hip fracture prediction based on recursive partitioning methods. *Med Decis Making* 2004; **24**: 386–98.
11 Baquerizo A, Anselmo D, Shackleton C *et al.* Phosphorus ans an early predictive factor in patients with acute liver failure. *Transplantation* 2003; **75**: 2007–14.
12 Valera VA, Walter BA, Yokoyama N *et al.* Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. *Ann Surg Oncol* 2007; **14**: 34–40.
13 Martin MA, Meyricke R, O'Neill T, Roberts S. Mastectomy or breast conserving surgery? Factors affecting type of surgical treatment for breast cancer – a classification tree approach. *BMC Cancer* 2006; **6**: 98.
14 LeBlanc M, Crowley J. A review of tree-based prognostic models. *Cancer Treat Res* 1995; **75**: 113–24.
15 Costanza MC, Paccaud F. Binary classification of dyslipidemia from the waist-to-hip ratio and body mass index: a comparison of linear, logistic, and CART models. *BMC Med Res Methodol* 2004; **4**: 7.
16 Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. *Hepatology* 1996; **24**: 289–93.
17 Kurosaki M, Matsunaga K, Hirayama I *et al.* The presence of steatosis and elevation of alanine aminotransferase levels are associated with fibrosis progression in chronic hepatitis C with non-response to interferon therapy. *J Hepatol* 2008; **48**: 736–42.
18 Segal MR, Bloch DA. A comparison of estimated proportional hazards models and regression trees. *Stat Med* 1989; **8**: 539–50.
19 Miyaki K, Takei I, Watanabe K, Nakashima H, Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. *J Epidemiol* 2002; **12**: 243–8.
20 Akuta N, Suzuki F, Tsubota A *et al.* Efficacy of interferon monotherapy to 394 consecutive naive cases infected with hepatitis C virus genotype 2a in Japan: therapy efficacy as consequence of tripartite interaction of viral, host and interferon treatment-related factors. *J Hepatol* 2002; **37**: 831–6.

21 Poynard T, Ratziu V, McHutchison J *et al.* Effect of treatment with peginterferon or interferon alfa-2b and ribavirin on steatosis in patients infected with hepatitis C. *Hepatology* 2003; **38**: 75–85.
22 Bressler BL, Guindi M, Tomlinson G, Heathcote J. High body mass index is an independent risk factor for nonresponse to antiviral treatment in chronic hepatitis C. *Hepatology* 2003; **38**: 639–44.
23 Romero-Gomez M, Del Mar Viloria M, Andrade RJ *et al.* Insulin resistance impairs sustained response rate to peginterferon plus ribavirin in chronic hepatitis C patients. *Gastroenterology* 2005; **128**: 636–41.
24 Konishi I, Horiike N, Hiasa Y *et al.* Diabetes mellitus reduces the therapeutic effectiveness of interferon-alpha2b plus ribavirin therapy in patients with chronic hepatitis C. *Hepatol Res* 2007; **37**: 331–6.
25 Marchesini G, Avagnina S, Barantani EG *et al.* Aminotransferase and gamma-glutamyltranspeptidase levels in obesity are associated with insulin resistance and the metabolic syndrome. *J Endocrinol Invest* 2005; **28**: 333–9.
26 Fraser A, Ebrahim S, Smith GD, Lawlor DA. A comparison of associations of alanine aminotransferase and gamma-glutamyltransferase with fasting glucose, fasting insulin, and glycated hemoglobin in women with and without diabetes. *Hepatology* 2007; **46**: 158–65.
27 Mazzella G, Salzetta A, Casanova S *et al.* Treatment of chronic sporadic-type non-A, non-B hepatitis with lymphoblastoid interferon: gamma GT levels predictive for response. *Dig Dis Sci* 1994; **39**: 866–70.
28 Villela-Nogueira CA, Perez RM, de Segadas Soares JA, Coelho HS. Gamma-glutamyl transferase (GGT) as an independent predictive factor of sustained virologic response in patients with hepatitis C treated with interferon-alpha and ribavirin. *J Clin Gastroenterol* 2005; **39**: 728–30.
29 Berg T, Sarrazin C, Herrmann E *et al.* Prediction of treatment outcome in patients with chronic hepatitis C: significance of baseline parameters and viral dynamics during therapy. *Hepatology* 2003; **37**: 600–9.
30 Akuta N, Suzuki F, Kawamura Y *et al.* Predictive factors of early and sustained responses to peginterferon plus ribavirin combination therapy in Japanese patients infected with hepatitis C virus genotype 1b: amino acid substitutions in the core region and low-density lipoprotein cholesterol levels. *J Hepatol* 2007; **46**: 403–10.
31 Adinolfi LE, Gambardella M, Andreana A, Tripodi MF, Utili R, Ruggiero G. Steatosis accelerates the progression of liver damage of chronic hepatitis C patients and correlates with specific HCV genotype and visceral obesity. *Hepatology* 2001; **33**: 1358–64.
32 Ortiz V, Berenguer M, Rayon JM, Carrasco D, Berenguer J. Contribution of obesity to hepatitis C-related fibrosis progression. *Am J Gastroenterol* 2002; **97**: 2408–14.
33 Muzzi A, Leandro G, Rubbia-Brandt L *et al.* Insulin resistance is associated with liver fibrosis in non-diabetic chronic hepatitis C patients. *J Hepatol* 2005; **42**: 41–6.
34 Charlton MR, Pockros PJ, Harrison SA. Impact of obesity on treatment of chronic hepatitis C. *Hepatology* 2006; **43**: 1177–86.
35 Di Bona D, Cippitelli M, Fionda C *et al.* Oxidative stress inhibits IFN-alpha-induced antiviral gene expression by blocking the JAK-STAT pathway. *J Hepatol* 2006; **45**: 271–9.
36 Minuk GY, Weinstein S, Kaita KD. Serum cholesterol and low-density lipoprotein cholesterol levels as predictors of response to interferon therapy for chronic hepatitis C. *Ann Intern Med* 2000; **132**: 761–2.
37 Gopal K, Johnson TC, Gopal S *et al.* Correlation between beta-lipoprotein levels and outcome of hepatitis C treatment. *Hepatology* 2006; **44**: 335–40.
38 Agnello V, Abel G, Elfahal M, Knight GB, Zhang QX. Hepatitis C virus and other flaviviridae viruses enter cells via low density lipoprotein receptor. *Proc Natl Acad Sci USA* 1999; **96**: 12766–71.
39 Leiter U, Buettner PG, Eigentler TK, Garbe C. Prognostic factors of thin cutaneous melanoma: an analysis of the central malignant melanoma registry of the german dermatological society. *J Clin Oncol* 2004; **22**: 3660–7.

Ⅱ　岩崎　学

成蹊大学理工学部　情報科学科

厚生労働科学研究費補助金（肝炎等克服緊急対策研究事業）
分担研究報告書

分類木の改良と多変量２値データの空間における構造の探索

岩崎　学（成蹊大学理工学部）

研究要旨：　本研究では，Ｃ型肝炎が有効な患者集団の特定のためデータマイニング手法の一つである分類木 (classification tree) が効果的に用いられた．分類木の作成は統計ソフトによって行われたが，コンピュータの出力である分類木の結果に対しさらなる工夫を加えることによりさらに良い分類木が作成できるかどうかを考察した．

また，分類木の結果は数学的には多変量２値データの空間とみなすことができる．その空間内に治癒率が高い点から低い点への１次元構造を見出すことは重要であり，ここではカテゴリカルデータの射影追跡の考え方を用い，高次元空間の点を低次元（２次元）の空間に射影するための指標を提案し，実際に計算して結果を出力する手法を提案した．

A．研究目的

大量のデータから有益な情報を取り出すための方法論であるデータマイニングにおける主要な手法の一つが分類木 (classification tree) である．近年，この種のマイニング手法は医療分野への応用も広がりを見せつつある．その理由はソフトウェアの充実にあり，専用のソフトウェアの利用によって容易に分類木が構成できるようなってきている．その際，コンピュータの出力そのものを鵜呑みにするのではなく，当該事象に対するより深い考察によってさらによい分類方式が得られる可能性がある．

個々の患者は，性別，年齢，各種検査項目などの大量の背景因子データを持っているが，それらからこの療法が有用となる確率が高いあるいは低い患者集団を見つけ出す必要があり，そこで分類木が用いられている．これまでのロジスティック回帰ではなく分かりやすい指針作りのために分類木を用いる．

分類木における分岐がすべて2分割の場合，分類された結果は多変量2値データとみなすことができる．したがって統計の問題としては，そのように定義された多変量2値データの空間を，有効率という観点からいくつかのサブグループに分けること，さらにはその空間内における治癒率が高いサブグループから低いサブグループへの１次元的なパスを見出すことにある．そのために，多変量２値データのスペクトル分解もしくはカテゴリカルデータの射影追跡などの多変量解析手法を用いることで対応する方策を提案する．

B．研究方法

まず分類木を統計手法とみた場合の性質を整理する．そして分類木の精度をより高く汎用的なものにするための指針を考察する．さらにはモデルの作成と検証をいくつかのシミュレーションによって行い，C型肝炎に対する治療効果に最適なサブグループを分類するツリーを構築する．また，これまで汎用されてきているロジスティック回帰分析との比較も行なう．そして，よりよい分類木を構築するため，近年多くの研究が行われつつある機械学習の分野における３種類の方法の適用可能性を探る．

分類木による分類結果は多変量２値データとみなすことができる．コンピュータによる出力結果を表で示した上で，カテゴリカルデータの射影追跡の考え方に基づく手法により予測因子と治癒率との相関がなるべく高くなるような予測因子の組み合わせを見出す方法を論じる．それにより，コンピュータが単に出力した結果からは見つけられなかったようなより効果的なサブグループを探索する方法によりサブグループを同定することができる．

分類木による空間における治癒率の順序に基づく１次元的な方向の探索を行なう．具体的には３次元もしくは４次元の場合につき，１次元的なパスを図示するための指標を考案する．そして具体的なパスの探索のためのアルゴリズムを導く．実際のデータへの適用だけでなく，コンピュータシミュレーションを用いることにより，そのアルゴリズムの有用性を示す．その方法を実際の肝炎データに適用することによりその有効性を検証する．また，今後さらなる改善が行える可能性についてもさらに論じていく．

厚生労働科学研究費補助金（肝炎緊急対策研究事業）
分担研究報告書

C．研究結果

前述の研究方法で述べた方法につき研究を行なった結果以下の諸点が明らかとなった．まず分類木が二進木である場合には，得られた分類結果は多変量２値データとみなすことができ，その多変量２値データの空間内において治癒率の大小という１次元の構造を探索する手法を開発した．その手法は多変量２値データのスペクトル分解の手法（Bloomfield (1974), Cox (1972), Iwasaki (1992) など）に基づくもので，分類木の各ノードでの分岐を０（左側）もしくは１（右側）で表現した場合，１の個数によって治癒率を表現する方策を考え，１の個数と治癒率との相関が最も高くなるようにした．また０と１の並びから分類木としては得られなかったが有望と思われるサブグループの同定法も試みた．実際にC]型肝炎に対するペグインターフェロン・リバビリン併用療法による肝炎ウィルスの消失の問題に対して適用した結果を評価した．

D．考察

コンピュータソフトウェアが出力する分類木の結果を改良する試みは，当該事象に関する知識を必要とすることから，少なくとも統計の分野ではこれまであまり行われてこなかったが，今回の共同研究により，方法論としても興味深い結果を得，かつ実際のデータに適用したところその有用性が示唆された．

今回の研究成果は2010年度の統計関連学会連合大会にて発表し，現在論文化を進めているところである．

E．結論

医療と統計の共同研究により得られた知見は両分野において今後生かされていくと同時に，私自身としても得難い経験となり，研究はもとより，統計手法の普及啓発においても生かされている（右の「学会発表」の項参照）．今後ともこの種の共同研究の機会があれば積極的にかかわっていきたいと考えている．

参考文献

Bloomfield, P. (1974). Linear transformations for multivariate binary data. *Biometrics*, **30**, 609-617.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics*, **21**, 113-120.

Iwasaki, M. (1992). Spectral analysis of multivariate binary data. *Journal of the Japan Statistical Society*, **22**, 45-65.

G．研究発表

1. 論文発表

Togo, K. and Iwasaki, M. (2010) Sample size re-estimation for survival data in clinical trials with an adaptive design. *Pharmaceutical Statistics*, Online 2010.10.

秋澤忠男, 岩崎　学, ．．．（23人中23番目）(2010) 血液透析患者の腎性貧血に対するKRN321（ダルベポエチンアルファ）静脈内投与の有効性および安全性－KRN321第Ⅲ相臨床試験－．腎と透析，**68**，3，423-435．2010. 5.

秋葉　隆, 岩崎　学, ．．．（30人中28番目）(2010) 保存期慢性腎臓病患者の腎性貧血に対するKRN321（ダルベポエチンアルファ）静脈内投与の有効性および安全性．腎と透析，**68**，3，436-448．2010. 5.

林　晃正, 岩崎　学, ．．．（37人中35番目）(2010) 保存期慢性腎臓病患者を対象としたKRN321（ダルベポエチンアルファ）とエポエチンアルファ製剤の皮下投与における貧血改善効果の同等性に関する検討－KRN321-SC第Ⅲ相臨床試験－．腎と透析，**68**，5，931-945．2010. 7.

2. 学会発表

岩崎　学：医療統計の基礎知識．東京都福祉保健局専門性向上研修．2010. 6.

三上智之・岩崎　学：多変量2値データの空間での1次元方向の探索．統計関連学会連合大会．2010. 9.

阿部貴行・佐藤裕史・山田祥岳・栗林幸夫・岩崎　学：Free-response ROC (FROC) 曲線を用いた画像診断データの解析．統計関連学会連合大会．2010. 9.

岩崎　学：傾向スコア (propensity score) の考え方と実際．統計数理研究所夏期大学院．2010. 9.

岩崎　学：特定保健用食品（トクホ）の信頼を支える統計解析．IBM SPSS Directions Japan 2010．2010.10.

阿部貴行・佐藤裕史・山田祥岳・小川健二・栗林幸夫・岩崎　学：Free-response ROC曲線に基づく統計的推測．日本計算機統計学会シンポジウム．2010.11.

岩崎　学：実験計画と分散分析の基礎．医学統計研究会定例シンポジウム2010．2010.11.

H．知的財産権の出願・登録状況（予定を含む。）

なし

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ
岩崎　学			カウントデータの統計解析	朝倉書店	東京	2010	1-210

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Togo, K. and Iwasaki, M.	Sample size re-estimation for survival data in clinical trials with an adaptive design	. Pharmaceutical Statistics	On-line (2010. 10)	1-7	2010
岩崎　学・堀地晶代	見落とされがちな平均に関する考察	ESTRELA	No. 198	22-27	2010
岩崎　学・堀地晶代	抗うつ薬は有効か？統計的有意性対臨床的有用性	統計	2010.11	54-63	2010

論文紹介　統計教育方法論専門誌“TEACHING STATISTICS”から　第2回

見落とされがちな平均に関する考察

A Closer Look at a Relatively Neglected Mean,
Teaching Statistics, Volume 27 Issue 3, pp.76-80 (August 2005).

Avital Lann and Ruma Falk
(The Hebrew University of Jerusalem, Israel)
訳：岩崎　学（成蹊大学理工学部教授）
　　堀地晶代（成蹊大学情報センター）

『Teaching Statistics』（http://ts.rsscse.org.uk/）は、英国王立統計学会統計教育センター（RSSCSE）内に設立された統計教育公益信託（TST）から、1979年に創刊し、以来、年に3号発行されている統計教育方法論を専門に扱った国際学術誌です。
本記事は、TSTの許可の下、『Teaching Statistics』の掲載論文の要約を意訳し、隔月で紹介するものです。

キーワード：　教育
　　加重平均
　　自己加重平均
　　自己加重抽出

訳者まえがき

何という偶然かつ我が意を得たり。私がこの論文を読んだときの感想がこれである。私は、まさに論文中で例に挙げられている問題を、この論文に接する以前から大学の授業で取り上げていて、平均値というと簡単なものに思えるけれども、実はそう単純なものではないと学生に伝えていたからである。本論文で取り上げられている事柄は、教育的な意味だけでなく、実際の調査の実務に携わる統計家にとっても興味ある内容となっている。是非ご一読いただきたい。

なお、スペースの関係で原論文にはあるが訳出しなかった部分、逆に多少詳しい説明を加えた部分があることを付け加えておく。

＊　＊　＊

1. ひとつの例

次のような例を考えてみよう（vos Savant (1996) pp. 56-58より改変）。A君は自分が進学する大学の選択にあたっていくつかの大学のパンフレットを調査したところ、ある大学では、開講する100ほどの授業での平均クラスサイズは20人であるとしていることを見出し、この大学への入学を目指すことにした。彼は少人数での授業という点を評価したためであった。ところが実際に大学に入学してみると、平均クラスサイズ（受講人数）はそれよりもかなり多いことに気付いた。A君は各クラスを受講している学生に尋ねて平均クラスサイズを計算したのだが、平均は53人を超えていた。A君はパンフレットの記載は虚偽であるとして大学当局に授業料返還を求めることができるであろうか。

上で示した2つの平均値（20人と53人）間に矛盾はない。学生にクラスサイズを聞いた場合には、大人数のクラスの学生は自分のクラスサイズは大きいと答え、そう回答する学生数は多いためである。この場合の平均値は、クラスサイズを重みとした加重平均となっている（次節参照）。それに対し、大学側のいう平均値は各クラスにおける人数の単純平均で、各授業担当者に尋ねた結果の平均値となっている。調査対象が学生か授業担当者かで、平均値は異なる

値となる。このように「平均値」と一言で言っても、それがどのように調査されたデータに基づくどういう種類の平均値であるのかを注意深く吟味しないと、意思決定などで誤りを犯すことにつながる。

2. 定義

一般に、$x_1, x_2, \ldots, x_n$ を n 個の数とし、$w_i \geq 0$ を x_i に与える重みとしたときの重み付き平均（加重平均）W は

$$W = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}$$

によって定義される。特に $w_i = x_i$ とした

$$WS = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{x_1 + x_2 + \cdots + x_n}$$

を自己加重平均（self-weighted mean）という（Lann and Falk（2002）参照）。

3. 平均値間の関係

通常よく用いられる平均は算術（相加）平均（A）、幾何（相乗）平均（G）及び調和平均（H）の3種であり、それらは以下のように定義される（G と H では x の値はすべて正とする）。

$$A = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1 \cdot x_1 + 1 \cdot x_2 + \cdots + 1 \cdot x_n}{1 + 1 + \cdots + 1},$$

$$G = \sqrt[n]{x_1x_2 \cdots x_n},$$

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{\frac{1}{x_1}x_1 + \frac{1}{x_2}x_2 + \cdots + \frac{1}{x_n}x_n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}.$$

定義式からもわかるように、A と H は W の特別な場合である（幾何平均は $n = 2$ のとき加重平均として表現できる：表1参照）。これらの平均値間には以下のような関係がある（たとえば Maor（1977）、Sheldon（2004）などを参照）。

$$x_{min} \leq H \leq G \leq A \leq WS \leq x_{max}.$$

ここで、等号はすべての x の値が同じときに限り成り立つ。自己加重平均が算術平均よりも大きいのは一目瞭然である。値が大きくなればなるほど重みは大きくなるからである（逆に H の場合は値が大きくなれば重みは小さくなる）。

4. 幾何学的表現

値が a と b の2つだけの場合、図式表現により各平均値間の大小関係が容易に理解できる。ここでは台形を用いた表現を示す（Beckenbach and Bellman（1961）p.62 及び p.126、並びに Hoehn（1984）より改変）。$0 < a < b$ とし、上底の長さが a、下底の長さが b の台形を描く。底辺に平行で各斜辺と交わる線分の長さは、何らかの加重平均の値を表している（図1参照）。底辺と平行な線分で作られる2つの台形の高さの比 w_a/w_b が a と b に与えられる重みの比となる。実際、線分の長さは $W = (w_aa + w_bb)/(w_a + w_b)$ となり、上述の各平均値は a と b を特定な値にとることにより加重平均として表現できる（図2参照）。

図1　台形による加重平均の表示

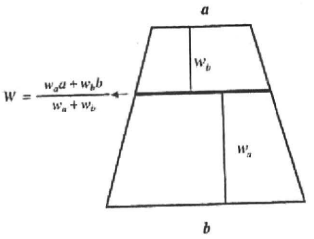


図2　各加重平均のグラフ表示

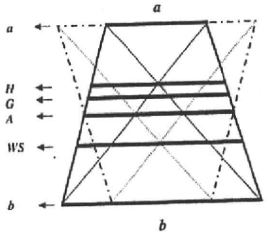


表1にそれぞれの平均値を大きさの順に並べて示している。表には一般のnでの各平均値の公式及び$n=2$の場合の式と重み、並びに台形での表示における線分の定義が与えてある。これらは容易に確認されるであろう。

HとWSは互いに逆の関係にある。すなわち両平均における重みが互いに逆数となっている点に注意されたい。これらは台形をさかさまにすることによって得られる。$n=2$では、それらが算術平均Aから等距離にあることは興味深い事実である。この理由により、実際、Hoehn (1984) は自己加重平均WSを反調和平均 (anti-harmonic mean) としている。

表1　各加重平均の定義

平均	n個の正の数 $x_1, x_2, \ldots, x_n$	2つの正の数 a と b, $a<b$			
	平均の定義	加重平均の定義	w_a	w_b	台形の底辺に平行な線分の定義
最小	$x_{min}=\min(x_1, \ldots, x_n)$	$a=\frac{1\cdot a+0\cdot b}{1+0}$	1	0	上底に一致
調和平均	$H=\frac{\frac{1}{x_1}x_1+\cdots+\frac{1}{x_n}x_n}{\frac{1}{x_1}+\cdots+\frac{1}{x_1}}$	$H=\frac{2ab}{a+b}=\frac{\frac{1}{a}a+\frac{1}{b}b}{\frac{1}{a}+\frac{1}{b}}$	$\frac{1}{a}$	$\frac{1}{b}$	対角線の交点
幾何平均	$G=\sqrt[n]{x_1\cdots x_n}$	$G=\sqrt{ab}=\frac{a\sqrt{b}+b\sqrt{a}}{\sqrt{b}+\sqrt{a}}$	$\sqrt{b}$	$\sqrt{a}$	2つの相似形に分ける
算術平均	$A=\frac{x_1+\cdots+x_n}{n}$	$A=\frac{a+b}{1+1}$	1	1	中心線
自己加重平均	$WS=\frac{x_1^2+\cdots+x_n^2}{x_1+\cdots+x_n}$	$WS=\frac{a^2+b^2}{a+b}$	a	b	回転させた台形の対角線との交点
最大	$x_{max}=\max(x_1, \ldots, x_n)$	$b=\frac{0\cdot a+1\cdot b}{0+1}$	0	1	下底に一致

5. 実際の場面での自己加重平均

冒頭に述べた例では、学生たちに尋ねた数値（クラスサイズ）がその学生たちの人数に比例していた。したがって、回答にはその回答の値だけの重みがかかっていることになる。それゆえこの調査法は自己加重抽出 (self-weighted sampling) とも呼ばれる。そしてその調査結果に基づいて計算されたクラスサイズの平均値（自己加重平均WS）は、各学生にとっての平均値であり、大学側の求めたクラスサイズの単純な算術平均Aよりも大きな値となる (Hemenway (1982)、Madsen (1981) などを参照)。このような調査対象や調査法の違いによる計算結果の違いは「標本抽出バイアス」あるいは「調査の矛盾」などとも呼ばれる (Stein and Dattero (1985))。

自己加重抽出は、実際問題の多くの場面で現れる。次の例を考えてみよう。バス停でバスを待つとき、バスがなかなか来ずにいらいらした経験は多くの人が持っているに違いない。バスの時刻表にはバスは1時間当たり3本の割合(すなわち20分に1本) で運行しているとあっても、実際にバスを待つ時間の平均値は10分よりは長くなってしまう。実際、待ち時間の平均を計算してみたら16分になったとしよう。いったいどうなっているのだろう。バス会社に文句を言うべきであろうか。たとえばバスの間隔が36分であるものが50%、4分であるものが50%であるとしよう（訳注：バスはよく団子になってくるものである。寺田寅彦の随筆にあるように)。あなたのバス停への到着がランダムであるとすると、36分間隔のバスの平均待ち時間は18分であり、そのバスに遭遇する確率は36/(36+4)=36/40である。一方、4分間隔のバスの場合は平均待ち時間2分であるが、それに運よくめぐり合う確率は4/40に過ぎない。

したがって平均待ち時間は $18 \times (36/40) + 2 \times (4/40) = 16.4$ 分となる。間隔の長いバスに遭遇する確率が高いことから、このような長めの待ち時間となるのである。これは待ち時間をその長さに比例する重みをかけて加える自己加重平均の例であり、待ち時間の平均はその単純な算術平均（このバスの例では $(18+2)/2 = 10$ 分）よりも大きくなる。van Dijk（1997、p.27）も指摘するように、待ち時間というものは一筋縄ではいかないものである。たとえ自己加重平均の計算法を知っていたとしても、いらいらすることに変わりはない。

もう1つの例として、高速道路に自動車の速度を自動測定するレーダーがあり、その地点を通過するすべての自動車の速度を一定時間測定しているとしよう。様々な速度の車がその地点を通過する場合、速度の速い車が計測される頻度は、遅い車より高い。ある車がレーダーに記録される確率は、その車が一定時間内に走行した距離、すなわちその車の速度と比例する。よって、記録された速度の頻度はその速度に比例することになり、記録された速度の平均値は自己加重平均になる（Falk et al.（2005）及び Stein and Dattero（1985）などを参照）。

同様の例として、健康診断における疾病のスクリーニング検査では、進行の遅い腫瘍のほうが発見される確率は高い。遅効性の腫瘍は長い時間存在しているからである（Zelen and Feinleib（1969））。したがって、様々な腫瘍間の比較では進行の遅い腫瘍がより多く報告されることになり、この種の偏りを調整する必要が出てくる。

このように自己加重抽出は様々な分野、たとえば人口統計、医療、遺伝、マネジメント、オペレーションズ・リサーチなどで多く遭遇する（Patil et al.（1988））。したがって、どのようなデータが自己加重抽出によって得られたのかの見極めが必要となり、これは、データ解析の結果を解釈する側及びデータ解析の結果を提供する側の双方にとって重要なこととなる。

6. 自己加重抽出と直観

加重平均の理解に関するこれまでの研究は、平均を重み付きで計算すべきであるときに、それを行わない例が多いことを示唆している。Pollatsek et al.（1981）は、多くの学生がサンプルサイズの異なるデータセットから計算した標本平均を統合して全体として1つの標本平均を計算する際に、それぞれの集団から求められた各標本平均を単に足して集団数で割ることにより求めてしまうという誤りを犯しがちであると報告している。この場合、標本平均の値そのものには、それを計算したサンプルサイズの情報が明示的には含まれていないため、加重平均をすべきであることが見過ごされてしまうのである。

我々は、ヘブライ大学の統計学の知識のあまりない学生に次のような課題を出した。「子供の数が1人、2人、3人、4人の家庭がそれぞれ同数ずつあるとする。各子供に自分の家の子供数を尋ねた場合、得られる平均はいくらであろうか」。冒頭のクラスサイズの例で示したように、この問題の答えは自己加重平均となる。話を単純化するために、上述の子供数の家庭が1つずつあったとしよう。各家庭を訪問して子供数を尋ねたとすると、その平均値（A）は $(1+2+3+4)/4 = 2.5$ となる。ところがこれらの子供10人全員が公園で遊んでいたとして、各子供一人ひとりに兄弟（姉妹）数を聞いたとすると、たとえば「兄弟（姉妹）数は3人」と答える子供が3人いるわけで、この場合の平均

値（WS）は$(1\times1+2\times2+3\times3+4\times4)/10=3.0$となり、これがこの問題への正解となる。ところが大多数の学生（72.2%）は誤って算術平均の2.5人と答えたのである。

同じくヘブライ大学の学生に、高速道路で時速100km/hの車と時速60km/hの車が同数いた場合、1時間中に速度測定のレーダーに記録される車のスピードの平均値を答えさせたところ、65.4%の学生が自己加重平均の85km/hではなく算術平均である80km/hと答えた。

すべての値に同じ重みをかけて算術平均を計算してしまうという性向は、Tversky and Kahneman(1974)のいうところのヒューリスティック原理に通い合うものがある。人々は複雑な問題を考える際、自らの直感に支配されて簡単なものと捉えてしまい、結果として誤った答えに到達しがちである。ナイーブな直感が誤りを生む例である。特に、重み付きで計算しなくてはいけないときに、重みをすべて同じとしてしまうことがよくある（Albert (2003)）。それは、わからないときはすべてのものを同等として捉えてしまうことにつながる。たとえば起こり得るいくつかの結果に対し、それらの生じる確率を同じとしてしまったりするなどは、その例である。

7. 教育的解説

自己加重抽出の場合であっても、学生は簡単な算術平均を求めてしまうという誤りを犯しがちであることに、教師は注意しなくてはいけない。実際、我々は、自分たちの学生が加重平均を計算すべきときに単純な算術平均を求めてしまう誤りを経験した。正しく加重平均を求めた学生は、子供の数の平均の場合は18%、高速道路でレーダーに記録された車の速度の問題では35%のみでしかなかった。これらの例は、平均を求めるという一見単純な問題であっても、学生たちに対し、どのような平均を求めるべきかを正しく考える教育及び訓練の必要性を示している。

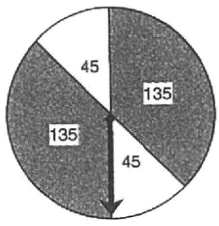


図3 自己加重ルーレット（針の止まったところの金額がもらえる）

教室で簡単な例題を出すなどして、単なる算術平均では駄目で、適切な重みを用いた加重平均を求めなくてはいけないという経験を積ませなくてはならない。それを助けるひとつの例として、図3のようなルーレットが有用であろう。ルーレットは135度、45度、135度、45度の4つの領域に分けられ、針が止まった領域に示されている数字に相当する金額のお金がもらえるとする。

4つの金額の算術平均は$A=(135+45+135+45)/4=90$である。しかし実際の期待値は自己加重平均であるところの$WS=(135\times135+45\times45+135\times135+45\times45)/360=112.5$である。小学校の子供でさえ、このルーレットで得られる金額の期待値は90よりも大きいことに気付くであろう。このわかりやすい例により、大きな数字となる確率が大きいことの説明が容易に見て取れ、その際求めるべき平均値は単なる算術平均ではないことも理解されるであろう。

＊　＊　＊

訳者あとがき

如何であろうか。本論文中で言及された事柄は実際問題でよく見られるタイプのものであり、もしかしたら思い当たる節のある読者諸氏もいるのではないかと推察する。同じような例でいかにも日本的なものとして、元東京理科大学教授の芳賀敏郎先生はかつて「流しそうめん

をはしでつまんでその長さを測る」を挙げられた。確かにその通りであるといたく感心した覚えがある。

本文中では各平均値間の関係を台形を用いて説明していたが、図示による直感的な証明の例がNelson (1993) に集められていて、きわめて興味深い。実際、私はこの本を大学1年生のセミナーで取り上げたのであるが、高校で習った証明よりはるかに分かりやすいと学生には評判がよかった。また、関連した話題を扱った最近の論文としては、同じ著者によるLann and Falk (2006) あるいはGraziani and Veronese (2009) などがある。また、Kadane (2008) も飛行機の込み具合という類似の話題を取り上げ、算術平均と自己加重平均の比 A/WS が $\boldsymbol{l}=(1, \cdots, 1)^T$、$\boldsymbol{x}=(x_1, \cdots, x_n)^T$ として

$$\frac{A}{WS}=\frac{x_1+\cdots+x_n}{n}\frac{x_1+\cdots+x_n}{x_1^2+\cdots+x_n^2}$$

$$=\frac{(x_1+\cdots+x_n)^2}{n(x_1^2+\cdots+x_n^2)}=\frac{(\boldsymbol{l},\ \boldsymbol{x})^2}{\|\boldsymbol{l}\|^2\cdot\|\boldsymbol{x}\|^2}$$

となり、これは l と x に関するコーシー・シュワルツの不等式に他ならないという興味深い事実を示している（これより $A/WS\leq 1$ が示される）。コーシー・シュワルツの不等式は相関係数の絶対値が1より小さいことの証明にも使われ、ここに平均値の比と相関係数との間の関係も出てくるのである。興味は尽きない、というべきであろうか。

本稿作成に当たっては、文部科学省の私立大学戦略的研究基盤形成支援事業の支援を受けた。

＊参考文献（[17]～[20]は訳者追加）

[1] Albert, J. H. (2003)：College students' conceptions of probability：*The American Statistician*, 57(1), pp.37-45.

[2] Beckenbach, E. and Bellman, R. (1961)：*An Introduction to Inequalities*：Washington, DC:The Mathematical Association of America.

[3] Falk, R., Lann, A. and Zamir, S. (2005)：Average speed bumps:four perspectives on averaging speeds：*Chance*, 18 (1), pp.25-32.

[4] Hemenway, D. H. (1982)：Why your classes are larger than 'average'：*Mathematics Magazine*, 55 (3), pp.162-164.

[5] Hoehn, L. (1984)：A geometrical interpretation of the weighted mean：*The College Mathematics Journal*, 15 (2), pp.135-139.

[6] Lann, A. and Falk, R. (2002)：An average with unimaginative weights: when the weights equal the values：in A. D. Cockburn & E. Nardi (eds), *Proceedings of the 26th Annual Conference of the International Group for the Psychology of Mathematics Education*：Norwich:University of East Anglia, volume 1, p. 288.

[7] Madsen, R.W. (1981)：Making students aware of bias. *Teaching Statistics*, 3 (1), pp.2-5.

[8] Maor, E. (1977)：A mathematician's repertoire of means：*Mathematics Teacher*, 70, pp.20-25.

[9] Patil, G. P., Rao, C. R. and Zelen, M. (1988)：Weighted distributions：in S. Kotz and N. L. Johnson (eds), *Encyclopedia of Statistical Sciences*：New York:Wiley, volume 9, pp.565-571.

[10] Pollatsek, A., Lima, S. and Well, A. D. (1981)：Concept or computation: students' understanding of the mean：*Educational Studies in Mathematics*, 12, pp.191-204.

[11] Sheldon, N. (2004)：The generalized mean：*Teaching Statistics*, 26 (1), pp.24-25.

[12] Stein, W. E. and Dattero, R. (1985)：Sampling bias and the inspection paradox：*Mathematics Magazine*, 58 (2), pp.96-99.

[13] Tversky, A. and Kahneman, D. (1974)：Judgment under uncertainty: heuristics and biases：*Science*, 185, pp.1124-1131.

[14] van Dijk, N. M. (1997)：To wait or not to wait: that is the question：*Chance*, 10 (1), pp.26-30.

[15] vos Savant, M. (1996)：*The Power of Logical Thinking*：New York:St Martin's Press.

[16] Zelen, M. and Feinleib, M. (1969)：On the theory of screening for chronic diseases：*Biometrika*, 56 (3), pp.601-614.

[17] Graziani, R. and Veronese, P. (2009)：How to compute a mean? The Chisini approach and its applications：*The American Statistician*, 63, pp.33-36.

[18] Kadane, J. B. (2008)：Load factors, crowdedness, and the Cauchy-Schwarz inequality：*Chance*, 21 (1), pp.33-34.

[19] Lann, A. and Falk, R. (2006)：Tell me the method, I'll give you the mean：*The American Statistician*, 60, pp.322-327.

[20] Nelson, R. B. (1993)：Proofs Without Words: Exercises in Visual Thinking：*The Mathematical Association of America*, Washington DC.

抗うつ薬は有効か？(Do antidepressants work?)

統計的有意性 対 臨床的有用性(Statistical Significance versus clinical benefits)

日本統計協会は、英国王立統計学会(RSS)の許可を得て、その季刊誌 Significance から本誌読者に興味のあると思われる記事を、統計専門家による翻訳と解説を加えて本年５月号から掲載しています。本稿は、Significance の2008年 Vol.5、No.2の"抗うつ薬は有効か？統計的有意性対臨床的有用性"（執筆 B. T. Johnson and I. Kirsch)から、岩崎学氏(成蹊大学教授)、堀地晶代氏(成蹊大学情報センター)が適宜、翻訳の上、解説を加えたものです。

翻訳・解説　岩崎 学・堀地 晶代

〈翻訳〉

我々の世代はプロザック世代と言われている。ところが今年（2008）の初め「抗うつ薬は効かない」という見出しが英国の新聞の一面を賑わした。本論の著者たちによる複数の臨床試験の再分析（メタアナリシス）の結果、プロザックのような抗うつ薬は効果がなくしかも処方が多すぎると報道されたのである。しかし、Blair T. Johnson と Irving Kirsch が以下に述べるように、実際はより複雑でしかもなおさらに興味深いものである。

訳注：本論で取り上げられているプロザック Prozac は日本では未承認である。

うつ病は、世界中で何百万人もの人々を苦しめている深刻な病気である。それは一時的に症状が悪化するものからさらに長期に渡って患者を苦しめるものまで様々な様相を呈し、病状の程度もごく軽いものからきわめて重度のものまでかなりの広がりをもっている。症状には、何らかの特異的なものが単独で現れることもあるが、妄想や幻覚といった精神的機能不全を伴うこともあり、そうなった場合は日常生活の営みも困難もしくは不可能になる。うつ病への対処法は、心理セラピーをはじめオメガ３脂肪酸サプリメント摂取など様々なものがある[1]。

うつ病の薬物療法の開発には40年以上の歴史があるが、1987年に米国食品医薬品局（FDA）がフルオキセチン fluoxetine を承認し、その製造元である Eli Lily 社が商品名プロザック Prozac として最初に市場に出したのが最初である。その後他の製薬会社も独自の抗うつ薬を開発し FDA に承認申請を行ってその後を追った。これらの薬剤の承認によって抗うつ薬の人気は増大し、精神科医や総合診療医によって日常的に処方されるようになった。現在、抗うつ薬は世界中で最も多く処方される薬のひとつである。

その人気とは裏腹に、抗うつ薬は人々が想像するほどうつ病に対して有効でないかもしれないと示唆する調査研究もある[2,3]。最近我々は[4]、フルオキセチン（プロザック）を含むヴェンラファキシン（エフェクサー）、ネファゾドン（サーゾーン）とパロキセチン（パキシル）の４つの新世代抗うつ薬を選び、FDA に提出された35の臨床試験結果を再解析した。FDA の審査では、申請資料を提出した製薬会社に対し、臨床試験によって得られたすべての結果の提供を要求している。米国連邦情

報公開法により我々はFDAに提出されたこれら臨床試験すべての結果を入手することができた。すべての結果を用いることにより、我々の用いたデータが臨床試験成績の恣意的な選択によって抗うつ薬の効果をいたずらに誇張したりあるいはしなかったりしている[5]のではないことが保証される。このように多くのデータが入手できたことにより、各臨床試験結果には相当程度のばらつきがあることや他の種々の有益な情報たとえば投薬開始時のうつ病の程度などの重要な知見が得られ、統計解析を行う上で大きなメリットとなった。

製薬会社からFDAに提出された個々の臨床試験は主として、抗うつ薬とプラセボ（偽薬：外見上は薬剤と同じであるが有効成分が含まれないもの）を設定し、うつ病患者のそれらに対する反応の差異を比較したものであった。多くの場合4週間から8週間の試験期間が設定され、試験は二重盲検すなわち患者も医者も投与されたのが薬剤であるかプラセボであるかが分からないようにして行われた。これら臨床試験データの再解析により次の事柄が明らかになった。全般に、抗うつ薬を投与された患者はプラセボを投与された患者よりも有意にうつ状態が改善している。しかるにこの結果をもってFDAは各薬剤を市場に出すことを承認するに至ったのである。加えるに我々のメタアナリシスの結果は、この全般的な結果よりもさらに詳細な結果を導いている。すなわち、抗うつ薬の効果は患者のうつ状態の程度によって変わり、投薬開始時のうつの段階が軽い患者にはあまり効き目がみられないといったことが明らかとなった。個々の試験は比較的重症のうつ患者に対する効果の判定のために計画されたものであったことから、これら軽度のうつ患者に対する結果は見過ごされてきたのであった。

残念なことに、この研究の最終的なレポート以前にマスコミは我々の研究結果を「抗うつ薬は効かない」という形で取り上げたのである。しかしそれらは、我々の研究結果のかなり微妙な結論を誤って表現したものであった。さらに、この微妙な解析結果は臨床現場に携わる人たちの持つ抗うつ薬に対する印象とも異なるものであった。

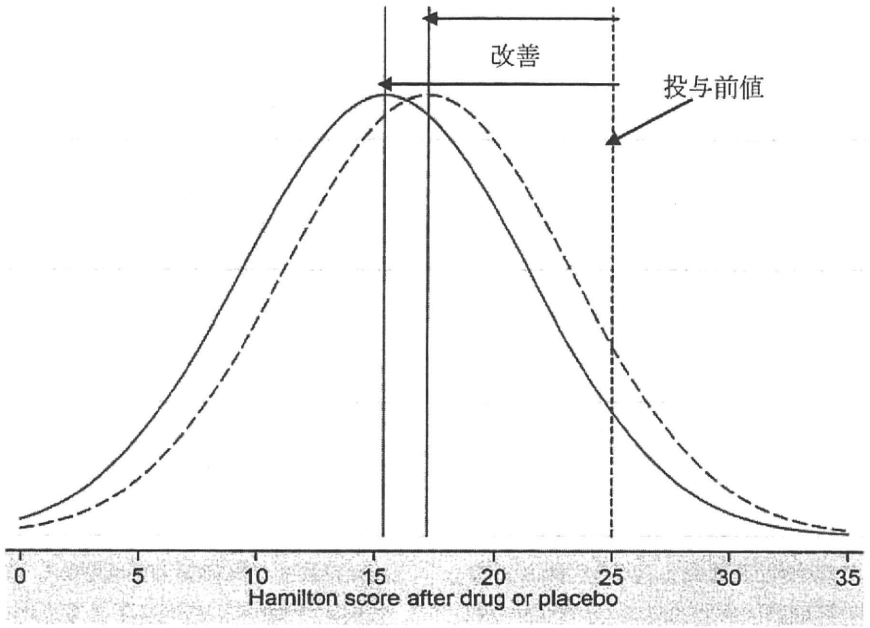
実際上意味を持つためにはどの程度の改善が必要か？

我々の分析結果の解釈を巡っての論争は、分析結果の統計的有意性と臨床的有意性との違いを中心に行われた。もちろんそれらはまったく同じというわけでないのはいうまでもない。後述するように、統計的に有意な結果であってもそれが患者一人ひとりの臨床症状に及ぼす効果はあまり大きくないといったことは十分にあり得る。

まず統計的有意性について考えてみよう。ここで論じる問題では実薬とプラセボ間の症状改善の程度の比較である。患者をランダムに実薬群とプラセボ投与群とに分けて臨床試験を実施することを（仮想的に）多数回行ったとしよう（図1）。比較の結果はP－値、すなわち両群間で効果に差がないという帰無仮説の下で現実に観測された差以上に両群間での差が大きくなる確率、によって表される。

我々は独立に実施された35本の臨床試験の結果を統合してメタアナリシス[6-8]を行なった。統合によってデータ数が格段に多くなったことから、我々の研究では実薬群とプラセボ群間の差の検出のための統計的検定の検出力が個々の研究それぞれの検出力よりもはるかに大きくなった。この結果、プラセボ群に比べ実薬群のほうが症状の有意な改善が見られたことは特に驚くべきことではない。さらに前述のように我々の分析結果は、抗うつ薬はうつ症状の重い患者にはより有効で、うつ症状が比較的軽い患者にはあまり有用ではないこ

図1　分析した臨床試験における投与後のハミルトンスコアの分布
（実線：実薬群、破線：プラセボ群）



とを見出した。

次に、臨床的あるいは実質上の有意性を考える。統計的な有意性に対し臨床的な有用性をどのように考えたらいいのだろうか。平均的な患者に対し抗うつ薬はどの程度有効であればいいのだろうか。イギリスのNICE(National Institute for Health and Clinical Excellence) は、この種の研究で通常用いられている[9]ハミルトンうつ病評価尺度Hamilton rating scale of depression(HRSD)での３ポイントの改善を臨床的に意味のあるものとしている。この評価尺度は０点から52点までの範囲のスコアからなるもので、そこでの３ポイントの差異は比較的小さいものであるように思われるであろう。実際、睡眠障害のみを取り上げても６ポイント動いてしまうことがある。統計学的な表現では、３ポイントの差異は標準化された効果の大きさ（エフェクトサイズ）ｄが0.5であることに相当している。偶然にもこの値は統計学者Ｊ・コーエンの「中程度のエフェクトサイズ」に対応している[10]。

コーエンはエフェクトサイズを両群での平均値の差を標準偏差で標準化して表現し、小（0.20)、中（0.50)、大（0.80）という基準を示した。コーエン自身は必ずしもこれらの基準を強く推奨したわけではなかったが、これらの値は臨床試験等における結果の表現において広く用いられるようになった。コーエンの意図は、注意深い観察者がすぐに見て取れるほどの効果の違いを「中程度のエフェクトサイズ」とする点にあった。この観点からすると、抗うつ薬の臨床的有意性は、医師や医療関係者のような注意深い観察者が、実薬群の患者がプラセボ群の患者より見た目で明らかにうつ状態が軽減した、と認めることができる程度ということができる。この点でNICEの３ポイントすなわちエフェクトサイズｄ＝0.5という基準は合

理的なものであるといえる。

図２は我々が行った分析の結果の一つを表している。図では、各臨床試験における投薬による改善は実線の回帰線のまわりの三角形で表し、プラセボでの改善は破線の回帰線のまわりの丸印で表されている。図形の大きさはその臨床試験での被験者数を表している（すなわち、被験者数の多い試験では大きな形となっている）。図中の影の部分は実薬群対プラセボ群でNICEの臨床的有意性すなわちエフェクトサイズｄ＝0.5を達成した部分を表している。図から見て取れるように、この領域はハミルトン尺度のスコアで28以上の範囲となっている。ハミルトン尺度では23以上が「非常に重いうつ」と見なされていることから、我々は、抗うつ薬は投薬開始時におけるうつの状態が「非常に重いうつ」より軽い場合には有意な効果を示さないが、「非常に重いうつ」以上の患者に対しては臨床的に有意な結果をもたらすものと結論付けたのである。

図２で示した分析は、各臨床試験においてプラセボ投与群に対し抗うつ薬投与群がどのような改善を示したのかを直接評価したものではない。我々の分析はハミルトンの評価尺度上での改善傾向を調べたものである。図３は、被験者が非常に重いうつ状態にない場合には臨床的有意性であるところの３ポイント改善という基準の達成が難しいことを示している。それに対し、もっと重いうつ状態の被験者に対しては抗うつ薬の効果が臨床的有意性の基準に到達しやすいであろうことを示唆している。

患者や医師たちの評価

表面的には、我々の分析結果は臨床現場での印象と異なるように見える。医師や患者は共に抗う

図２　投与前のうつの重症度（横軸）と改善の程度（縦軸）。
（実線および三角印：実薬群、破線および丸印：プラセボ群）

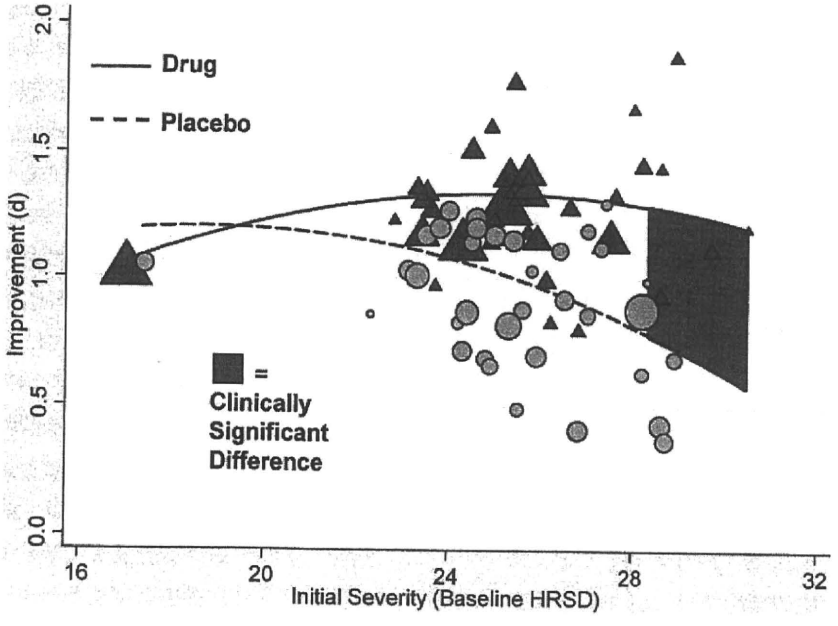
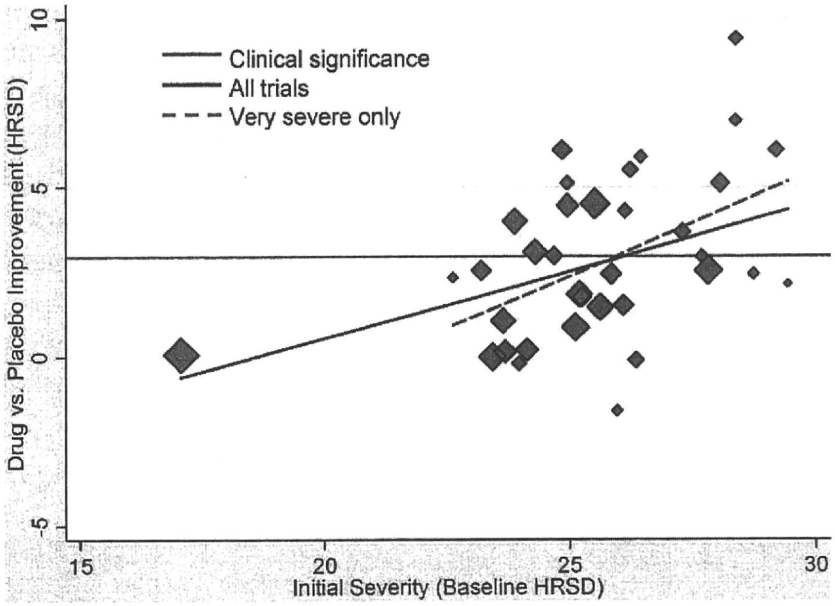


図３　実薬群とプラセボ群の投与時での重症度（横軸）と改善度の差（縦軸）
縦軸の改善度＝３がNICEによる臨床的有用性の基準



つ薬は効果があると主張している。しかし通常の医療現場では患者にプラセボが処方されることはないことに注意すべきである。臨床試験のように患者は有効成分が入っているかいないか分からないようなものを処方される、ということはない。すなわち、医師も患者も有効成分が何もないときに「これによって病気が治るかもしれない」という単なる期待（いわゆるプラセボ効果）だけによって病状が軽快することがどの程度あるのかを知る立場にはない。それに対し抗うつ薬の臨床試験では、実薬もしくはプラセボが被験者も医師も分からないようにして無作為に投与される。プラセボの投与により、単なる期待のみによってどの程度病気が治るのかが試されるのである。すなわち臨床試験は、薬剤の真の有効性をその有効成分の寄与のみとして客観的に評価するものと位置づけられる。通常の臨床現場では、プラセボの投与がないことから臨床症状の改善の中に単なる期待によるプラセボ効果が幾分かは入ってしまっている。そしてそれは程度の軽いうつの場合により著しいのであり、それ故薬剤の投与によって病気が治ったとされるのである。しかし当然のことながら実薬はプラセボより効果があるので、特に非常に重いうつの場合には実薬の効果によってプラセボ効果を凌駕する改善がみられるのである。

実際、我々の実施したメタアナリシスにおける実薬のデータのみをみると、薬剤の投与によって明らかに大きな改善が得られている。概していえば、薬剤によりハミルトン尺度で平均9.6ポイントの改善が得られている。これはNICEの臨床的有意性の基準３ポイントを劇的に上回るものである。しかし問題は、プラセボを処方された患者群でも改善が得られているという点である。プラセボ群でもハミルトン尺度で平均7.8ポイントの改善が

得られていて、これもまたNICE基準に照らし合わせると大きな改善といえよう。共に処置前後では効果があるといえるが、本来の目的は実薬とプラセボとの比較であり、その比較を行うと薬剤の効果は限定的なものとなってしまう。両群での改善の差は9.6−7.8=1.8ポイントに過ぎず、この差は臨床的有意性の基準を満たしてはいない。ハミルトン評価尺度での２ポイントの改善は、一人の患者に対する診断の最中に手で髪の毛を弄んだかどうかということを意味する程度に過ぎない。この程度の改善では、注意深い観察者であってさえどの患者が実薬を与えられどの患者がプラセボを与えられたかを見分けることができないに違いない。しかし、実薬かプラセボかは分からないものの投与の前後でのうつ状態が改善したとの判断はされるであろう。重症のうつ病患者以外では、改善するという期待（プラセボ効果）が大きく影響してくる。軽度のうつ病患者に対する薬剤の効果は、有効成分の働きというより薬剤を投与したということおよびそれによって症状が改善するであろうという期待によるものであるともいえる。うつは精神的な不調であるので、このような心理学的な要因が強く働くとしても驚くことではない。

プラセボが処方されない実際の医療現場では、医師も患者もプラセボ効果の程度が不明であるため、患者の症状の改善が処方した薬剤のおかげだと考えるのも無理のないことである。プラセボによってもたらされた効果は当然実薬が処方された患者たちにも同じく現れるであろう。人々は通常薬は効くものだと期待しているがそのことがうつ病には大きなプラセボ効果をもたらすのである。このことはこれまでの研究によっても立証されている[11,12]。同様に、磁気共鳴画像装置（MRI）による研究では、プラセボで症状が改善したうつ病患者は抗うつ薬が有効であった患者と同様の神経的な変化がみられることを示されたという[13]。プラセボがうつ病の治療に対してこのような大きなインパクトがあるという事実はうつ病の症状そのものを考え直すことにつながるかもしれないが、いずれにせよプラセボがうつ病の症状に何がしかの影響を与えることは否定できない。たとえば、改善するという希望によって患者がまわりの人たちに微笑みかけ、微笑を受けた人たちが患者に微笑み返すことが患者のうつ状態を軽減する手助けになるかもしれないのである。

抗うつ薬の臨床的有意性の基準

「科学は、真実の程度をもたらすのではなくそれを計測する道具を提供するのみである[14]」ことから、統計的有意性の基準のみでは十分なものではなく、臨床的もしくは実際的な有意性の基準もまた必要であり、さらにはそれら基準に対する何らかの解釈が必要である。米国FDAは抗うつ薬の承認に対し統計的有意性をその根拠としている。しかしNICEのような機関のもたらした客観的な臨床的有意性もまた考慮されるべきだと我々は信じている。

重要なことは、薬剤の効果は抗うつ薬とプラセボとの比較によって得たものだけでは測れないという点である。この観点からは、臨床試験において観測された非常に重度のうつ状態にある患者層に対する中程度のエフェクトサイズでさえ重要ではないとみなされかねないし、確立した臨床的有意性の基準が十分厳格であるかどうかとの疑問もわいてくる。私たちは抗うつ薬が最も重度のうつ病患者だけのための臨床的有意性に限ることに満足すべきなのであろうか。

薬剤投与時のうつ状態がハミルトン尺度で「とても重いうつ」を示すおよそ25ポイントの患者におけるプラセボもしくは実薬でのうつ状態の変化

の大きさを考える。実薬が投与された患者の平均的なスコアの改善度は9.6ポイントであったので、患者は薬剤投与後に平均してスコアが15.4ポイントになっている。ハミルトン尺度での15.4ポイントは「中程度のうつ」とみなされる。通常は7ポイント未満で「正常」もしくはうつ状態ではないとされる[15]。米国精神医学会 American Psychiatric Association (APA) と NICE は共にスコア8～13が「軽度」、19～22が「重度」、以上が「非常に重度」としている（図1参照）。スコアが30以上の患者は妄想、幻覚症状といった精神病的特徴に悩まされている傾向がある（こういった患者は我々が分析した臨床試験における除外基準により除外されている）。我々が調べた臨床試験はひとつを除いては平均スコアが23と30の間すなわち「非常に重度のうつ」の患者たちのものであった（「重度のうつ」未満の患者に対する臨床試験はほとんどなく、そのことが我々の結論を軽度のうつの患者たちにまで適用することの妨げになっている）。これらの事実は臨床試験の計画に由来するものではあるが、この種の情報の提供は実際に医療現場での投薬前後におけるうつの程度の差の評価の手助けになるであろう。

問題は治療後のスコアが15となったとしても依然としてまだ強いうつの症状を有しているということにある。臨床的に有意な改善がもたらされたとしても、患者たちのうつ状態をさらに軽減するためになすべきことはまだあるのである。プラセボを投与された患者たちは平均的にはハミルトン尺度で16.5程度であり、プラセボであっても実薬群に比較してそう極端に悪いわけではない。この種の比較のロジックは Jacobson and Truax[16]によってさらに議論されている。我々の議論にはコネチカット大学のステファニー・ミラン博士の協力を得ている。抗うつ薬を投与された患者については、平均的には臨床的に有意な改善は見られたというものの、多くの患者はまだ高度のうつ状態にある。要するに、投与されたのがプラセボであれ実薬であれ、患者は平均して「非常に重いうつ」から「中程度のうつ」になったのみであり、まだかなり改善の余地が残されたままの状態にあるといえよう。

ハミルトン尺度値がおよそ30の患者層では、プラセボと比べて抗うつ薬が大きな治療効果をもたらしてはいるというものの、事態はそう大きく改善しているとは言い難い。抗うつ薬の投与によりこれらの患者が平均12.8ポイントの改善が得られたとしても依然としてスコアは平均17.2である。プラセボの投与では平均して8ポイントの改善でスコアは平均22ポイントになるのみである。これらの群間差は注目に値する。しかしながら、両群ともほとんどの患者が臨床上望ましくないとみなされるような状態のままであることは間違いない。

これまでの議論では主として平均的な改善度を扱ってきて、平均のまわりでのデータのバラつきを考慮してこなかった（図1）。臨床的有意性に関する今ひとつの評価基準として、うつ状態が「正常」となった患者の比率がある。ハミルトン尺度で6以下となった患者の比率はどの程度であろうか。実薬の投与前にスコアが25ポイントであった患者のうちスコアが6以下となったのは5.86%であった。それに対しプラセボ群では、スコアが6以下となったのは3.09%であった。ハミルトンスコアの平均が30で開始した試験の実薬群の患者の正常化率は3.10%であったのに対し、プラセボ群では0.38%であった。個々の臨床試験での再発率を考慮していないため我々のメタアナリシスでは患者の長期にわたる予後を調べることは出来なかった。別のデータソースを基にした議論で、専門家たちのなかには「正常」であることの基準はも

っと厳格であるべきというものもある。ハミルトン尺度で2になった患者たちであれば再発率が低くなっているとの指摘もある[9]。当然ながら、判定基準を厳格にするほど臨床的な有意性には到達しにくくなる。

これまで述べてきた比較により、抗うつ薬の臨床的な有意性は「非常に重いうつ」の患者に対してでさえ限定的であることを示唆している。しかし、臨床的有意性についてのひとつの異なる見解は、「たとえ小さな効果であったとしてもそれが患者が問題なく活動できるための壁を超えるには十分であるかもしれない」というものである。抗うつ薬によく反応する患者層もいることは事実である。そうであっても、抗うつ薬以外の治療法には、抗うつ薬に高頻度で出現する副作用や有害事象（それ自体がうつの源となる）のような不都合なものが潜在的に少なめであるようなものもあるであろう。

私たちの行ったメタアナリシスはただひとつの結果変数すなわちハミルトン尺度の変化量に的を絞ったものである。薬剤投与前のうつの程度が、投薬治療により副作用などとの絡みにおいてどのように変化していくのかについては今後の研究を待たねばならない。さらには、治療にかかる費用的な問題も避けては通れないものである。

上述したように、医師としては患者に対してプラセボを処方することは出来ないので、実際上うつ病患者は何ヶ月あるいは何年もの間抗うつ薬治療を続けることになる。これらの薬は後発品でない限りかなり高価なものである。それ故、他のうつ病治療法の中の費用対効果が高いものへの切り替えあるいは併用の可能性が生まれる。いずれにせよ、うつを軽減するためにどんな解決法を試みたとしてもその費用対効果は切り離して考えることはできない。我々の研究は、抗うつ薬のもたらす恩恵は一般に広く信じられているほどは大きくはないことを示したものであるといえよう。

＜キーワード解説＞

ここでは、記事に現れたいくつかの用語について説明を加える。

・新薬の承認

医薬品は、一般用医薬品と医療用医薬品に大別される。前者はOTC（Over The Counter）とも呼ばれ、医師の処方箋がなくても一般の薬局、ドラッグストアで購入可能なものをいう。後者は医師の処方箋に基づいて薬局などで処方されるもので、その発売に関しては規制当局（日本では厚生労働省）の承認（認可）を必要とする。承認を得るためには、各製薬メーカーは当該医薬品の規格や安定性という物理化学的な性質、動物実験の結果、ヒトを対象とした臨床試験の結果を取りまとめた膨大な資料を提出し、医薬品の有効性・安全性に関する厳正な審査を受ける。審査を担当するのは、米国では食品医薬品局（Food and Drug Administration = FDA）であり、日本では医薬品医療機器総合機構（Pharmaceuticals and Medical Devices Agency = PMDA）である。審査ではこれら機関内および外部の医薬統計家が大きな役割を果たしている。

・無作為化二重盲検群間比較試験

医薬品などの効果を客観的に立証するためには比較が必須である。新薬の開発では、新規医薬品をプラセボ（下記参照）もしくは既存の薬剤と比較した臨床試験が実施される。比較にあたっては、比較の対象物の違い以外の条件は両群で均一で偏りのないのが望ましい。それを達成するため、被験者への薬剤の割付けをランダムに行い、さらには処方されたのがどちらのものかが投与された被