The Optimal Discovery Procedure:
A Generalization of the Neyman-Pearson Fundamental Lemma to Multiple Significance Testing

野間 久史
情報・システム研究機構　統計数理研究所
2016年10月14日
大分統計談話会第54回大会
e-mail: noma@ism.ac.jp
URL: http://www.ism.ac.jp/~noma/

1

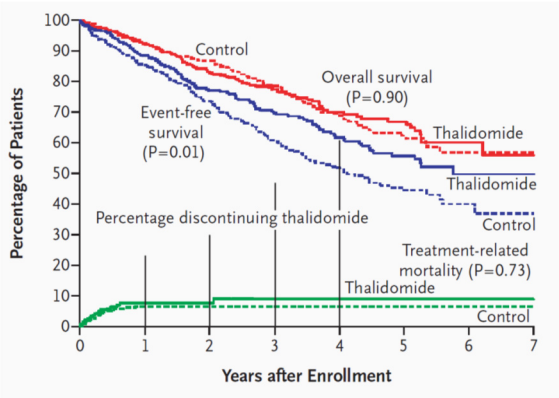
サリドマイド　Thalidomide

- 1957年 西ドイツで鎮静・催眠薬として開発
  - 動物実験で副作用が認められなかったことから、「妊婦や小児が安心して飲める薬」として世界中で使われるように
- 世界中で、妊婦のサリドマイド服用による子どもの四肢欠損症（サリドマイド胎芽症）が報告される
- 日本でも、睡眠薬イソミン（1958年発売）胃腸薬プロバンM（1959年発売）として販売されたが、1962年9月に販売停止された

https://tampoppo.jimdo.com/

2

多発性骨髄腫の治療薬として

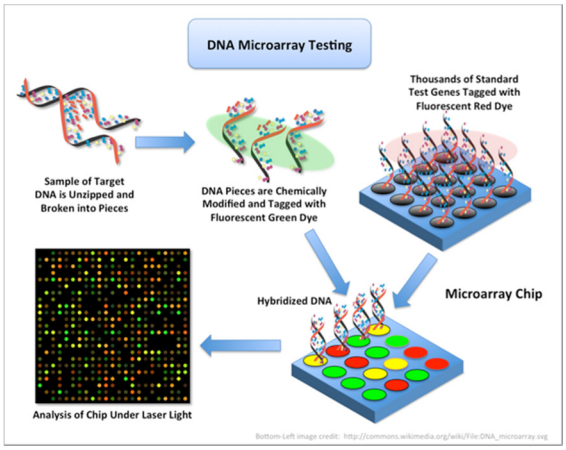


Barlogie et al. (2006 NEJM 354, 1021-30)

- UARK 98-026 TT II 試験
- Melphalanによる高用量化学療法との併用療法が有効かどうかを検証した第III相ランダム化臨床試験
- サリドマイド群（N=323）
- コントロール群（N=345）
- EFSでは有意差あり（P=0.01）
- 一方、OSではあまり顕著な差は認められなかった（P=0.90）

3

DNA Microarray



https://thegenefactor.wordpress.com/final-paper/

- がん細胞における、遺伝子の発現情報を、数万種類以上、同時に測定することができるHigh-Throughput Assay
- 遺伝子ごとの発現レベルが、連続量のデータとして測定される
- 治療効果の予測因子となり得る遺伝子を大規模な遺伝子情報から探索するために、臨床試験等でも広く用いられている

4

UARK 98-026 TT II 試験

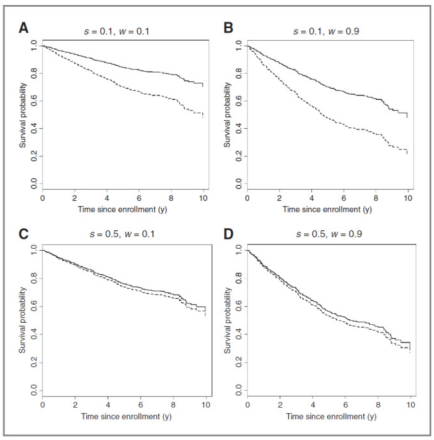


Figure 3. A to D, the predicted survival curves when receiving thalidomide (solid curves) and no thalidomide (dashed curves) for 4 patients with different values of the predictive score (S) and the prognostic score (W); (s, w) = (0.1, 0.1), (0.1, 0.9), (0.5, 0.1), and (0.5, 0.9). s = 0.1 (0.5) represents a high (small) responsiveness to thalidomide, whereas w = 0.1 (0.9) represents a good (poor) prognosis.

Matsui et al. (2012 Clin Cancer Res 18, 6065-73)

- DNA Microarrayによる遺伝子発現データから、治療への反応と関連する遺伝子を探索し、それによって層別したKaplan-Meier曲線の推定値
- 全集団では、OSで差が出なかったものの、遺伝子発現情報によって規定される治療効果が大きく期待できるサブグループがあることが示唆された

5

治療効果予測因子の同定の問題

- 大きな治療効果が期待できるサブグループがある場合、それを規定する治療効果予測因子を、大規模な遺伝子発現データから正確に同定するには？
  - UARK 98-026 TT II試験で使われてている Affymetirix U133Plus2.0 microarrayも、54,675遺伝子のデータが得られている
- 大多数の遺伝子は治療効果予測因子ではない
- 大規模なノイズを含むデータの中から、正確に治療効果予測因子となる遺伝子を選択するためには？？

Matsui, Noma, Qu et al. (2017 Biometrics, DOI: 10.1111/biom.12716.) 6

Multiple Significance Testing

- Cox Proportional Hazard Regression
  - $h(t) = h_{0j}(t)\exp(\beta_j x_{ij})$
  - x_{ij}: i番目の対象者のj番目の遺伝子の遺伝子発現データ
  - β_i: i番目の遺伝子の対数ハザード比
- OS, EFSなどのエンドポイントと遺伝子データの関連を "$\beta_i = 0$" の検定によって評価
- 数万規模の多重検定によって、関連遺伝子を探索することに
- 5%水準の検定では、平均5%の頻度（5万個であれば2500個）のFalse Positiveが必ず生じることに！！

7

False Discovery Rate (FDR)

- Familywise Error Rate (FWER)
  - FWER = Pr(V ≥ 1)
  - V: Number of false positive
  - S: Number of significant tests
- False Discovery Rate (FDR)
  - FDR = $E[V/S]$
- 探索的な大規模多重検定問題には有効な方法であり、FDRの原著論文Benjamini and Hochberg (1995) の引用回数は、既に3万4千回以上（Google Scholar; 2016年10月現在）

8

The Neyman-Pearson Fundamental Lemma

- The well-known Neyman–Pearson (NP) lemma provides an optimal procedure for performing a single significance test when the null and alternative distributions are known (Neyman and Pearson, 1933). Given observed data, the optimal testing procedure is based on the likelihood ratio

$$\frac{\text{probabaility of data under alternative distribution}}{\text{probabaility of data under null distribution}}$$

The null hypothesis is then rejected if the likelihood ratio exceeds some prechosen cut-off. This NP procedure is optimal because it is 'most powerful', meaning that for each fixed type I error rate there does not exist another rule that exceeds this one in power.

Storey (2007) 9

Generalization to Multiple Significance Testing

- Although a single-hypothesis test involves forming a statistic, deriving a set of significance regions, and determining the type I error rate for each region, these components can conceptually be broken down into two major steps when testing multiple hypotheses:
  - (a) determining the order in which the tests should be called significant and
  - (b) choosing an appropriate significance cut-off somewhere along this ordering.
- The first step can also be thought of as the process of ranking the tests from most to least significant.

Storey (2007) 10

Storey (2007)'s Formulation

- EFP (Expected False Positive)

$$\mathrm{EFP}(\Gamma) = \int_{\Gamma} f_1(\boldsymbol{x})d\boldsymbol{x} + \cdots + \int_{\Gamma} f_{m_0}(\boldsymbol{x})d\boldsymbol{x}$$

- ETP (Expected True Positive)

$$\mathrm{ETP}(\Gamma) = \int_{\Gamma} g_{m_0+1}(\boldsymbol{x})d\boldsymbol{x} + \cdots + \int_{\Gamma} g_m(\boldsymbol{x})d\boldsymbol{x}$$

- 一般性を失うことなく、m個の検定のうち、はじめの m_0個が帰無仮説が正しいものとする；$\boldsymbol{x}_j \sim f_j(\boldsymbol{x})$ $(j = 1,2,\ldots,m_0)$，$\boldsymbol{x}_j \sim g_j(\boldsymbol{x})$ $(j = m_0 + 1,\ldots,m)$
- Γ は、棄却域を表す（$\boldsymbol{x}_j$'sは共通の確率空間を持つものと仮定する）

11

Storey (2007)'s Lemma

- The multiple-testing procedure that maximizes ETP for each fixed EFP among all STPs (single-thresholding procedure) is defined by the following significance thresholding function:

$$S_{\mathrm{ODP}}(\boldsymbol{x}) = \frac{g_{m_0+1}(\boldsymbol{x}) + g_{m_0+2}(\boldsymbol{x}) + \cdots g_m(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_{m_0}(\boldsymbol{x})}$$

- Storey (2007) named this optimal testing procedure as "optimal discovery procedure (ODP)."
- James-Stein推定量と同じく、検定間の情報を "Borrowing" することで、全体の検出力を上げることができる

12

Comparison of NP and ODP Testing Approaches

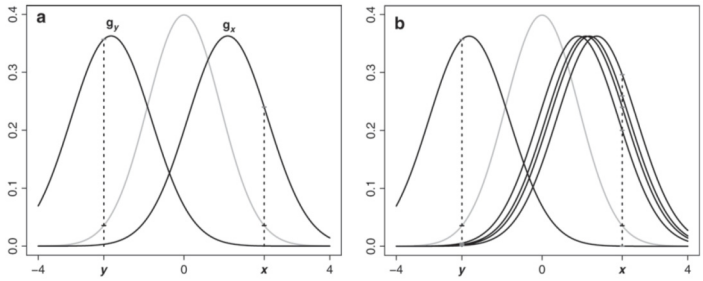


Fig. 1. Plots comparing the NP testing approach to the ODP testing approach through a simple example. (a) NP approach. The null (gray) and alternative (black) probability density functions of a single test. For observed data x and y, the statistics are calculated by taking the ratio of the alternative to the null densities at each respective point. In this NP approach, the test with data y is more significant than the test with data x. (b) ODP approach. The common null density (gray) for true null tests and the alternative densities (black) for several true alternative tests. For observed data x and y, the statistics are calculated by taking the ratio of the sum of alternative densities to the null density evaluated at each respective point. In this ODP approach, the test with data x is now more significant than the test with data y because multiple alternative densities have similar positive means even though each one is smaller than the single alternative density with negative mean. A color version of the figure is given in the supplementary material available at *Biostatistics* online, Figure 8.

Storey et al. (2007)

13

Storey (2007) のODPの問題点

$$S_{\mathrm{ODP}}(\boldsymbol{x}) = \frac{g_{m_0+1}(\boldsymbol{x}) + g_{m_0+2}(\boldsymbol{x}) + \cdots g_m(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + \cdots + f_{m_0}(\boldsymbol{x})}$$

- 理論上、最適な検定方式は構成できたものの、個々の検定の分布 $f_1(\boldsymbol{x}), \ldots, f_{m_0}(\boldsymbol{x}), g_{m_0+1}(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x})$ は未知であり、実装上はデータから推定しなくてはいけない
  - 数万個のCox回帰モデルの推定結果のPlug-in
  - $S_{\mathrm{ODP}}(\boldsymbol{x})$のPlug-in推定量のVariationはかなり大きなものに
- 分子・分母に、True Null, True Alternativeの組を分ける必要がある
  - それがわかっていれば、検定などする必要がない…！！

14

Empirical Bayes Formulation

- 点推定の問題では、James-Stein推定量の解釈として、Efron and Morris (1972, 75) によって経験ベイズ法としての定式化が行われた
- 経験ベイズ法の枠組みで、Storey の Optimal Criterion を達成する検定方式はどうなるか？
  - $\boldsymbol{x}_i \sim f(\boldsymbol{x}|\boldsymbol{\theta}_i, \boldsymbol{\psi}_i),\ (i = 1,2,\ldots,m)$
  - Null True: $(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i) \sim G_0(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{\xi}_0)$
  - Alternative True: $(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i) \sim G_1(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{\xi}_1)$
  - それぞれの検定の間に、パラメータの交換可能性を仮定したモデル

15

Derivation of ODP (1)

- It is sufficient to show that, for any Γ' such that $\mathrm{EFP}(\Gamma') \geq \mathrm{EFP}(\Gamma)$, $\mathrm{EFP}(\Gamma') \leq \mathrm{EFP}(\Gamma)$. Equivalently, for any Γ' such that $\mathrm{EFP}(\Gamma')/m_0 \geq \mathrm{EFP}(\Gamma)/m_0$, $\mathrm{ETP}(\Gamma')/m_1 \leq \mathrm{ETP}(\Gamma)/m_1$.

$$\frac{\mathrm{EFP}(\Gamma)}{m_0} = \int_\Gamma \underline{\int f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\psi}) dG_0(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{\xi}_0) d\boldsymbol{x}}$$

$$\frac{\mathrm{ETP}(\Gamma)}{m_1} = \int_\Gamma \underline{\int f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\psi}) dG_1(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{\xi}_1) d\boldsymbol{x}}$$

- The optimization problem is equivalent to the NP lemma!!

16

Derivation of ODP (2)

- Therefore, the significant regions obtained by the marginal likelihood ratio statistic

$$\Lambda(\boldsymbol{x}) = \frac{\int f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})dG_1(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_1)}{\int f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})dG_0(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_0)}$$

achieves the Storey's optimality criterion.

17

Noma and Matsui (2012)'s ODP

- The optimal discovery procedure that maximizes ETP among all STPs is defined by the following significance thresholding function under the empirical Bayes formulation:

$$R_{\mathrm{ODP}}(\boldsymbol{x}) = \frac{E_{G_1}[f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})]}{E_{G_0}[f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})]}$$

$$= \frac{\int f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})dG_1(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_1)}{\int f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})dG_0(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_0)}$$

18

Noma and Matsui (2012)'s ODP

$$R_{\mathrm{ODP}}(\boldsymbol{x}) = \frac{\int f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})dG_1(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_1)}{\int f(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\psi})dG_0(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_0)}$$

- 個々の検定の分布 $f_1(\boldsymbol{x}),\dots,f_{m_0}(\boldsymbol{x}),g_{m_0+1}(\boldsymbol{x}),\dots,g_m(\boldsymbol{x})$ を推定する必要はなく、事前分布 $G_0(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_0),G_1(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_1)$ のみを推定すればよい
- 分子・分母に、True Null, True Alternativeの組を分ける必要はなく、事前分布 $G_0(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_0),G_1(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_1)$ さえうまく推定できれば、それをPlug-inすればよい

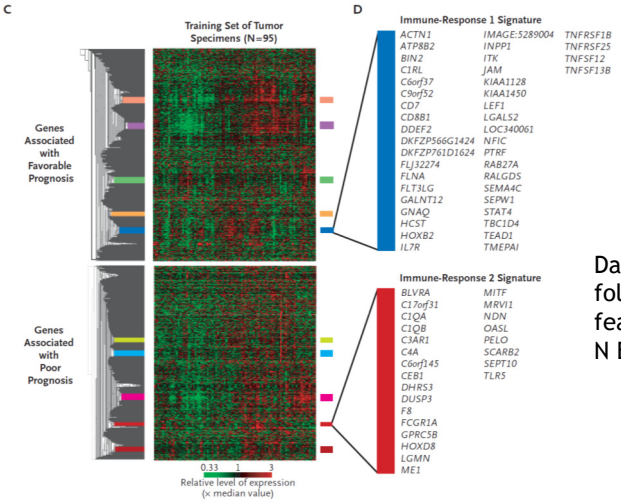
19

Estimation of the Hyperparameters

- Two-Component Hierarchical Mixture Model
  - $\boldsymbol{x}_i \sim f(\boldsymbol{x}|\boldsymbol{\theta}_i,\boldsymbol{\psi}_i),\ (i=1,2,\dots,m)$
  - $(\boldsymbol{\theta}_i,\boldsymbol{\psi}_i) \sim \underbrace{\pi_0 G_0(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_0)}_{\text{Null}} + \underbrace{\pi_1 G_1(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\xi}_1)}_{\text{Non-Null (Alternative)}}$
- 2000年前後から発展した、Efron and Tibshirani (2002), Efron (2008) のMixture ModelのEmpirical Bayes理論、Storey (2002, 2003) の FDR のベイズ流の定式化と同じ
- Noma and Matsui (2012, 2013ab, 2015) では、EMアルゴリズムによる最尤推定法を採用している

20

Analysis of Lymphoma Clinical Study



Dave et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N Engl J Med. 2004 Nov 18;351(21):2159-69.

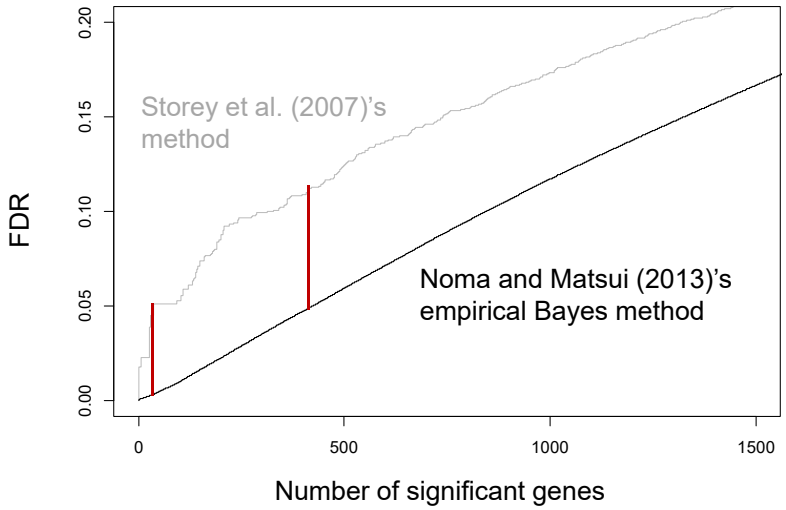
21

Analysis of Lymphoma Clinical Study

- 191名の非ホジキンリンパ腫患者の生検標本の遺伝子発現情報を Affymetrix HG-U133A,B Microarraysで解析
- 44,928遺伝子についての多重検定
- 予後のアウトカムについて、2グループ比較
  - Poor prognosis (5年以内の死亡; 51名)
  - Good prognosis (5年以上生存; 109名)
- 非ホジキンリンパ腫患者の予後に関連する遺伝子を同定し、生物学的な機序についての分析，予測アルゴリズムを構築することが目的

Dave et al. (2004) 22

Results of the ODP analyses



23

Concluding Remarks

- 米国でオバマ大統領によるPrecision Medicine Initiativeが出されたことにより、遺伝情報の医療への応用は、近年、また大きな話題を呼んでいる
- 臨床試験，臨床研究における効率的なバイオマーカー開発等のために、有効な多重検定の方法論は、実務上も有用なアプローチとなると考えられる

24

References

- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289-300.
- Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1-22.
- Efron, B., and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *Journal of the American Statistical Association* **67**, 130-139.
- Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**, 311-319.
- Efron, B., and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70-86.
- Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289-337.

25

- Noma, H., and Matsui, S. (2012). The opitimal discovery procedure in multiple significance testing: an empirical Bayes approach. *Statistics in Medicine* **31**, 165-176.
- Noma, H., and Matsui, S. (2013a). Bayesian ranking and selection methods in microarray studies. In *Statistical Diagnostics for Cancer*, F. Emmert-Streib, and M. Dehmer (eds), 57-74. Weinheim: Wiley-VCH.
- Noma, H., and Matsui, S. (2013b). An empirical Bayes optimal discovery procedure based on semiparametric hierarchical mixture models. *Computational & Mathematical Methods in Medicine* **2013**, 568480.
- Noma, H., and Matsui, S. (2015). Univariate analysis for gene screening: Beyond the multiple testing. In *Design and Analysis of Clinical Trials for Predictive Medicine*, S. Matsui, R. Simon, and M. Buyse (eds), 227-251: Chapman and Hall/CRC.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479-498.
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013-2035.

26

- Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B* **69**, 347-368.
- Storey, J. D., Dai, J. Y., and Leek, J. T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* **8**, 414-432.