特集：EBMとEBH

良質の根拠を生む randomization の本質
－科学研究者としてのセンス－

丹 後 俊 郎

Essence of randomization creating the best quality of evidence

Toshiro TANGO

1. はじめに

日本の臨床では無作為化臨床試験（RCT, Randomized Controlled Trial）に対する研究上の関心とその重要性への認識が少ないため，きちんと試験プロトコールを決めて行われたRCTが極めて少ない．そのためか最近の「根拠に基づく医療」（EBM, Evidence-Based Medicine）の根拠（Evidence）が無作為化比較試験（RCT, Randomized Controlled Trial）にあることは以外と知られていない．日本の臨床でよく行われる過去の治療歴を利用した非実験的かつ記述的な後ろ向き比較研究にも充分な根拠があると錯覚している臨床医も少なくない．

本小論では，母集団の無作為標本とは言えない来院患者等の被験者を対象とした比較研究の結果が良質な根拠の一部として蓄積されるための必要条件として無作為化（randomization, random allocation）が果たす役割とその重要性を解説する．

2. 統計学の基本

まず，統計学の基本をおさらいすることから始めよう．統計学の原点は研究対象集団（母集団）からのrandom samplingにある．random sampleの平均値は母集団の平均値のバイアスのない不偏推定量となる．しかし，random sampleでなければその平均値はバイアスのある推定量となり，一般に，その大きさ，方向はわからない．この未知のバイアスが研究結果の信憑性に大きく影響を与えることになる．極端に言えば，バイアスの大きさと方向が全くわからなければ，その研究結果は評価できないのである．例えば、皆さんがよく知られているStudentのt検定

$$T=\frac{\overline{X}_A-\overline{X}_B}{\sqrt{\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\left(\frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A+n_B-2}\right)}} \quad (1)$$

による推測プロセスを再確認してみよう．重要なポイントは，計算されたT値がt分布の97.5パーセント点である1.96より大きければ，両側検定で有意（statistically significant, two-tailed $p < 0.05$）と推測する点である．

なぜ，t分布を参照するのだろうか？ データがrandom samplingされていれば，帰無仮説の下でStudentのt検定統計量がt分布するからであり，逆に，random sampleでなければt分布する保証はどこにもない．したがって，極言すれば，nonrandom sampleに対しては統計学的に有意であるとか有意でないとかの推論は不可能となることにまず注意していただきたい．たとえば，2，3年前からの診療記録を整理してt検定を適用して… と安易に考えるべきではないということです．

前節では，母集団の縮図としての標本をとるには「無作為抽出」という抽出確率がどの個体も同じとなる無作為化の手続きの重要性を解説した．「比較」を基本とする医学研究においては，後の2.5節で述べるように，地域特性を調査する研究を除いては，標本の無作為抽出は一般に不可能で，それより,「作用因子の無作為割り付け（random allocation）」という無作為化の手続きが重要となる．それは

1．評価したい因子以外の結果に影響を及ぼす潜在的な交絡因子（confounding factor, confounders）の影響を少なくし，
2．比較する群のデータのバラツキをほぼ等しくする

という重要な役目がある．これは統計解析を非常に簡単にし，結果の解釈を単純明瞭にさせる．したがって，これが可能か否かでその研究結果の信憑性，データの解析方法，解釈のしかたが大きく異なるのである．それでは，研究の種類別に無作為割り付けの役割とその重要性を解説していこう．

3. 動物実験

ここでは，2種類の薬剤A,Bを投与して3時間後の反応を観察して比較するラットを使用した動物実験を考えてみよう．使用するラットはそれぞれ10匹ずつである．以下は実験状況の記録である．

1．まず，薬剤Aの実験を最初に実施する計画を立てた．実際に実施した日はどんよりとした曇り空の極めて寒

国立公衆衛生院
附属図書館・疫学部理論疫学室

い日で，実験者の体調もすぐれなかったが変更するわけにいかず，室内の温度を高めに設定して窓を締め切って行った．その際，実験に使用したラットは薬剤Bを投与する予定のもう一方の群のラットに比較すると体重の重いものが多かったが気にしなかった．

2．薬剤Bの実験を行った日は快晴で暖かい日であったので，窓を全開して行った．体調も良かったので実験に要した時間も前回の実験よりも短時間で終了した．この原因としては実験に対する慣れもあるかも知れない．

3．両群の実験結果のデータをStudentの両側t検定で検定したところ有意にB薬剤の反応効果が高かったので薬剤BはAに比較してより効果の大きいものであると結論した．

さてこの種の実験では，「観察された反応の差が薬剤だけの効果を表しているだろうか？」という疑問に明快に解答できなけらばならない．しかし，この実験ではあまりにも実験環境が違っている．

・実験者の体調の違い
・実験順序の違い（時間的要素）
・天候の違い（温度，湿度，光）
・体重の違い（個体差）

これでは，観察された差が薬剤の効果を表しているという結論ははなはだ疑問である．少なくとも，「実験者の技能，光，熱，湿度」，などの因子は反応に影響を与える最も基本的な攪乱因子，潜在的な交絡因子であることは多くの種類の実験で知られているわけで，これらの因子が異なる実験環境で測定された実験結果はもはや比較できないのである．さらに，動物，ヒトという生体を対象にする場合は更に，「時間（日内変動・日間変動），個体差」などの因子が加わる．したがって，実験では処理以外に結果に影響するかもしれない因子を事前に検討し「同一環境に制御できるものは設定する（光，熱，湿度など），できないものは処理を無作為に割付ける（時間，個体差，など）」ことが重要となる．つまり無作為割り付け（random alloction）によって，制御不可能な要因の影響を「確率的に均一化」して実験誤差（偶然変動）の中に組み入れることができるのである．特に重要な点として強調したいことは，「現在の知識では分からない未知の因子までも誤差に組み込める」点が素晴らしい！ということである．こうすることにより，「A処理群とB処理群との差が処理A,Bの他には偶然だけでしかない」
という比較可能性（comparability）を保つことができる．この方法が，Fisherによって提唱された実験計画法（design of experiment）であり，そのための統計手法が分散分析である．なお，Studentのｔ検定は2群の一元配置分散分析，Studentの対応のあるｔ検定は2群の2元配置分散分析と同じである（analysis of variance）．

今の実験の例で言えば，つぎのようにすれば良いだろう．

1．実験室の環境（光，温度，湿度）は一定にする（⇒差は生じない）．

2．実験は体調が同一コンディションの時に行う（⇒差は無視できる）．

3．各ラットにどの薬剤を投与するかは無作為割り付けを行う．体重の違いが実験結果に大きく影響を与える場合には，体重でいくつかのブロックに分類しその中で処理の無作為割り付けを行う（⇒個体差は偶然変動へ転化される）．

4．実験順序も無作為化を行う．（⇒実験順序の差は偶然変動へ転化される）．

4. 臨床研究

ヒトを対象とした臨床研究では次節で述べるように病院に来院する患者（標本）の母集団からの偏りの問題がつきまとうが，ここではまず，前節と同様の「比較上」の問題に絞って議論しよう．

図1は，ある医学雑誌から抜粋したもので，2種類の処理A，Bに対する反応の比較をある同一疾患の患者の検体を用いて行った結果である．実験の興味は反応の平均値であ

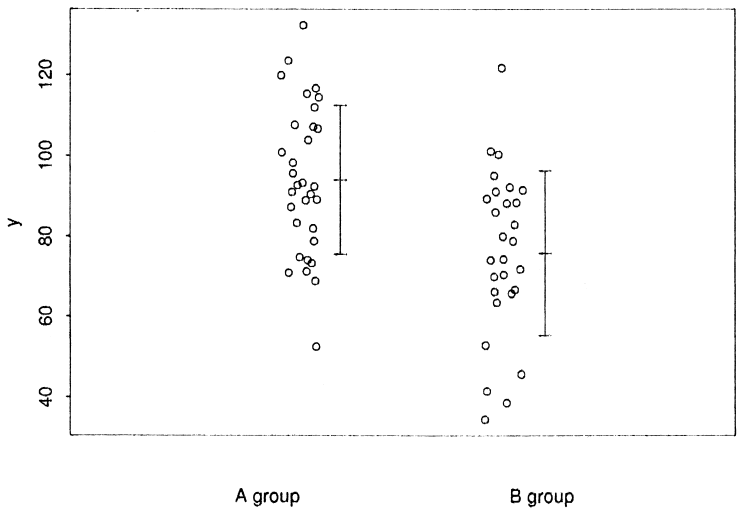

図1　ある医学論文に掲載された2群比較

る．一見すると2群間に差がありそうである．事実，二標本のWilcoxon rank sum testで計算した結果は有意差があった．しかし，その論文のMaterials and Methodsには「処理の割り付け」に関する記載が全くないのである．これでは信用できないではないか！なぜなら，「ある処理に対する生体反応Yはその生体の特性（X_1, $X2$, ...）によって大きく変化してしまう」ことが多いからである．しかも，患者を対象にする場合，同一疾患と言えどもhomogeneousな集団ではなく病気の状態が大きく異なるheterogeneousな集団である．たとえば，

・処理効果の差はなく，
・図2（a）に示すように反応Yと特性X_1との間に正の相関がある（つまり，特性X_1の値がもともと高い個体は処理の反応も高くなる傾向がある）

状況を考えてみよう．この場合，

1．特性X_1の値の高い患者がA群に多ければ図1に類似した図2（b）の結果がでる，
2．逆に，B群に多ければ，図2（c）の反対の結果でてしまう

という特性X_1に交絡（confound）した見かけの差（bias）が生じてしまうのである．X_1が観測不可能であれば，どちらが真実かは神様だけがご存じである．この見かけの差，つまり，交絡因子（confounding factors）によって生じたバイアスを交絡因子によるバイアス（confounding bias）と呼ぶ．さて，今の研究で，「無作為に割り付けを行えば，特性X_1の値の大きい個体と小さい個体の割合が2群間で（他の総ての特性値も！）確率的にバランスされ，一方の群に高い個体が多く集まるという可能性は小さくなり，真に処理差がなければ，正しい図2（d）の結果が期待され，図2（b），図2（c）に示すような見かけの結果が起こる確率は有意性検定の有意水準以下に制御できる」のである．もっとも，無作為化は各群の特性を均一にする「可能性が大」なのであって「必ず保証するものではない」．したがって，時にはいくつかの因子に関してバランスが保てないことも起こりえる．特に標本の大きさが少数の場合には偏りを生ずる確率も高くなる．したがって，重要な（観測結果に影響を与える）背景因子に偏りが見られた場合には解析で調整する必要がある．この方法として

1．反応が計量値であれば，共分散分析（analysis of covariance）．
2．反応が2値であれば，ロジスティック回帰分析（Logistic regression analysis），Mantel-Haenszel法

などを適用する．しかし，解析で事後的に調整することには限界があるので，事前にいくつかの因子が結果に影響を与えることがわかっている場合には，その因子を2－3つのカテゴリーに分けて，それぞれのカテゴリーの中で割り付けを無作為化する「層別無作為化（stratified randomization）を実施する」を行う．また，比較的小さい規模の試験であって，重大な影響を与える可能性がある予後因子を事前に明確に特定できる場合には層別無作為化に代わって予後因子の分布の偏りを強制的に最小化する割り付け法として「最小化法（Minimization）を実施する」ことが多い．これは患者が試験に登録される毎に交絡因子の分布の偏り状況を判断して行う逐次操作が必要でありコンピュータの利用が必須である．

5.　臨床試験－無作為割り付けは必須？

だれしも，乱数で自分の運命が左右されたのではたまったものではないと感じるであろう．その患者に有効なはずの（担当医師が経験的にそう思っているだけにすぎない）治療を受ける機会が奪われるといって無作為化臨床比較試験は倫理上問題があり実施できないと主張する臨床医が多い．一方

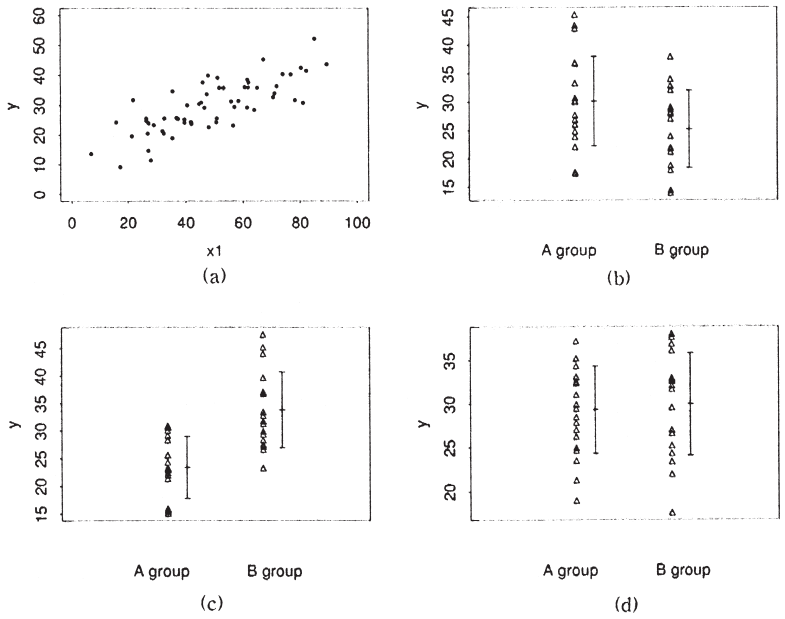

図2　交絡因子の影響

```
> b<-runif(500)
> round(b)
  [1] 1 0 0 0 0 1 1 0 1 1 1 1 0 1 0 0 1 0 1 0 1 1 0 0 0 0 0 0 1 1 0 0 0 1 0 1 1
 [38] 1 1 0 1 1 0 0 1 1 1 1 0 1 0 1 1 0 0 0 0 0 0 1 0 1 1 0 1 0 1 1 0 0 1 1 0 1
 [75] 0 0 1 1 0 1 1 1 0 0 1 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0
[112] 1 0 0 1 1 1 0 0 0 0 0 1 0 1 1 0 0 1 0 1 1 1 0 1 0 0 0 0 1 0 1 1 0 0 1 0 1
[149] 0 1 1 1 0 0 1 0 0 0 1 0 0 1 0 0 1 1 0 0 1 1 0 1 1 0 1 0 1 1 1 1 1 0 1 1 0
[186] 0 1 1 1 0 1 1 1 1 1 0 0 1 0 1 1 0 1 1 1 0 1 1 0 0 0 0 1 1 1 1 1 1 0 0 1 1 0
[223] 0 1 0 0 1 1 1 0 1 0 1 1 1 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1
[260] 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 0 0 1
[297] 0 1 0 1 0 1 0 1 1 1 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0
[334] 1 0 0 0 1 0 1 0 1 0 0 1 1 1 0 1 0 0 1 1 1 0 0 0 1 1 0 0 0 1 0 0 1 0 1 1 1 0
[371] 1 0 1 0 1 0 1 1 0 0 0 0 1 0 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 0 1 0 0 0 0 0 1
[408] 1 1 1 0 0 0 0 0 1 1 1 1 0 1 1 0 1 1 1 1 0 0 1 1 1 0 1 0 0 0 1 0 0 1 1 1 0 0
[445] 0 1 0 0 0 0 1 0 0 1 1 0 1 1 1 0 1 0 1 0 0 1 1 1 0 0 1 0 1 0 0 1 0 0 1 1 1 0
[482] 1 0 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 0 0 1
> a<-hist(b,breaks=c(0,0.5,1),plot=F)
> a$count
[1] 251 249
>
```


図３　統計ソフトS-PLUSにより生成された０－１乱数例500個

で，ある治療法を２－３人の患者に実施して成績が続けて良かったりするとその治療法が良いと思い込んでしまう主観的判断が問題である．そこで，図３を見てみよう．統計ソフトSPLUSを利用して500個の0,1の乱数列を表示したものである．それぞれの生起確率は等確率（=1/2）である．確かに，500個の中で０は251個，１は249個，とそれぞれ約半数出現している．ところが，Xで示した10個の数列では０が８回現れている，またYで示したところは逆に１が８回連続している．つまり，２回に１回の出現が期待される事象であっても，一方が何度も連続して出現することがよくあることを示している．つまり臨床医の経験がこの乱数列のどの局面にいたかで治療法に対する「思い」が大きく変化してしまうのである．

また，治療法には，すべて，それを支持する人，批判的な人，無関心な人がおり，中立的な立場の人は少ないものである．したがって，その治療法が有効であると主張する客観的な証拠を提示しない限り，その治療法に熱心な集団を除いては，誰も評価はしてくれない！

１．対照も置かず，無作為割り付けもせずに実施された研究（オープン試験）では当該治療法に都合よい方向に偏った結論を導いたが，

２．後にきちんと対照群を置いて比較試験を実施した結果，対照群に比較して有意に劣ってしまった

いたという事例は，公表バイアス(publication bias)を考慮するとかなりの頻度にのぼるものと推測される．Glantzは「治療法に対する執着度と試験デザイン」との関連を，1950年代に肝硬変治療として実施されていた門洞静脈吻合術を評価した51の論文で調査した．結果は表１に示すように熱心な研究者ほど対照群すら置かずに，また対照群を設置していても無作為割り付けを実施していないことが分かる． 対照群を置かない研究でこれほどまでこの手術に支持が偏った理由はまさに観察者側の偏向と患者側のプラセボ効果（効果の如何にかかわらず手術を受けたというだけで回復する効果）の何者でもない．事実，この手術は現在行われていない．したがって，次のように宣言できる．「治療法Ａと治療法Ｂのどちらが有効かが誰も分からない無作為化比較臨床試験には倫理上の制約はない」むしろ，比較可能性が乏しいデータに「正しい統計手法」を適用して誤った結果を導くことのほうがはるかに倫理上の問題があるように思われる．将来，その結果に基づいて発生するであろう不必要な研究に費やされる不幸な研究者と研究協力者，費用，時間の地球規模の損失，不必要でかつ不適切な治療を受けることになる最も不幸な患者群を考えてみてほしい．なにが正しいか理解できるだろう．

表１　門洞静脈吻合術を評価した51の論文の評価 (Glantz, 1992)

試験デザイン	手術に対する執心度			計
	高い	中位	なし	
対照群なし	24	7	1	32
対照群あり（非無作為化）	10	3	2	15
対照群あり（無作為化）	0	1	3	4

6. random allocationができない疫学研究

ヒトの健康に悪い影響を与えるリスク因子を研究する疫学研究（epidemiology）では動物実験・戦争時代の軍部による人体実験を除くと，リスク因子を無作為にヒトに割り付けることは倫理的に許されない．したがって，喫煙に関する研究では「喫煙者 vs. 非喫煙者」，大気汚染の健康影響に関する研究では「主要幹線道路沿いの住民 vs. 緑の多い住宅街の住民」などを比較するというように，現在住んでいる一人一人の嗜好形態，行動様式，生活習慣，社会環境，環境汚染状況の違いを上手に利用して観察する研究

(observational study) にもとめなければならない．したがって，実験のところで強調した様々な潜在的交絡因子（性，年齢，職業など）が存在し，かつその一部しか実際には観測できないため，比較したい群どうしの比較可能性が保証されない．そのため，少数の交絡因子でマッチングをとったマッチドケースコントロール研究も行われるが，多くのまた未知の因子でのマッチングは不可能である．したがって疫学研究では「調査時点で除去できない交絡は統計解析で調整（adjust）する」ことが必須条件となるが，完全に調整することはできない点が疫学研究の方法論上の最大の問題点である．

疫学研究は実験ではなく観察研究であるから，観察・調査に付随した問題点は避けられない．冒頭に述べた交絡の問題以外にも，

1．調査に回答しない回答拒否（non-response）健康状態と関連していることが少なくない（selection bias）
2．伝統的な測定手段であるアンケート調査の正確度・精密度がよくわからないことが多い．これは精度に格段の進歩が見られる臨床検査を測定手段とした研究に比べると極めて切れ味が悪い道具である（information bias）
3．面接調査の方が郵送調査より正確な情報が得られると言われるが，面接者は面接しようとする対象がケースかコントロールかについてブラインドがかかっていないことが多く，ケースの面接に熱心となる傾向がある（interviewer bias）
4．患者の記憶に多くを依存するケースコントロール研究ではケースの記憶のほうがコントロールの記憶より明確であることが多い（recall bias）
5．食習慣と各種がんに関する研究における食習慣の測定，電磁波と白血病を調査する研究における電磁波の暴露量の測定，のように過去のリスク因子への暴露量に関する測定の信頼性が低い（測定誤差，measurement errors）

などが疫学研究の「疫病」として立ちはだかっていて研究結果の再現性を極めて低いものにしている．1995年のScienceでは「疫学は限界に直面している」という表題で特集記事を掲載している．その主旨は「健康を脅かすリスク因子として食習慣（肉類，コーヒー，ヨーグルト，アルコール，など），環境因子（除草剤，殺虫剤，電磁波，ダイオキシンなど），薬剤（経口避妊薬など）に関する疫学研究の結果が雑誌に次々と発表されるが，有意に関連があったという発表が出るやいなや有意な関連がなかったという矛盾した発表が相次ぐため国民はなにを信じたらよいか分からない！国民の間には疫学研究がもたらした不安病が蔓延している（Anxiety epidemic）ではないか！疫学研究の結果は信用できるのか？」というかなりきついものである．この問題に対して欧米の著名な疫学者，医学統計学者が登場して，上述した「疫病」が矛盾した研究結果が相次ぐ主要な原因であるとのべているが，一方で，

1．報道の仕方にも大きな問題がある．疫学研究では一つの研究結果だけでリスクの大きさと因果関係を評価することはできない．数多くの一連の研究で類似の結果がでても，生物学的な因果関係が確認されるまでは，そのリスクが明確に評価されることは少ない．これに対し，報道関係者はたった一つの研究結果を，他の研究結果と分離して，しかも有意な関連が認められた部分だけを大げさに報道するから混乱が生じるのである，
2．更に，喫煙の数多くの有害性，肥満と多くの病気との関連，身体的運動の心疾患予防効果，多くの職業暴露のリスク（ベンゼン，アスベスト），日光と皮膚がん，薬害（サリドマイド），果物と野菜摂取のがん予防効果，などの，多くのがん，心疾患の予防に関する有用な知識の多くは疫学研究から得られたことに全く触れないのは明らかに偏った報道である（media bias）

と反論している．しかし，この問題は，基本的にはリスクの無作為割り付けができないため十分な交絡因子の調整が不可能で，結果として無視できないバイアスの大きさが研究によって異なることに起因している．したがって，単一の疫学研究の結果だけではリスク評価はできず，30－40もの類似の研究をまとめて評価することが必要になる．そのための統計技法として最近メタアナリシス（Meta analysis）が有効な方法として期待されているが，それを適用しようとすると，今度は別の問題，公表バイアス（publication bias）が立塞がる．つまり，今日の医学研究論文が採択される基準は「統計学的有意差が必要」となっていることが多く，有意でない結果の論文は採択されない，または，論文を投稿しない傾向がある．したがって，有意な効果を示した論文だけが雑誌に掲載され，その論文だけをまとめてメタアナリシスを行うと，明らかに有意な方向にバイアスを持った結論が導かれることになる．

臨床試験でも最近メタアナリシスにより薬剤の効果を世界レベルで再評価しようという動きが著しい．疫学研究ほどバイアスは大きくないが，公表バイアスの問題は同様である．真の意味のメタアナリシスを可能にするためには，地球上で行われている医学研究すべてをデザインから解析結果まで，登録できるシステムを構築しなければならない．

参考文献

[1] 丹後俊郎. 統計学のセンスーデザインする視点・データを見る目. 朝倉書店，1998.
[2] 丹後俊郎. 新版：医学への統計学，朝倉書店，1993.
[3] 宮原英夫・丹後俊郎編：医学統計学ハンドブック，朝倉書店，1995.
[4] Berkson, J. Limitations of the application of fourfold table analysis to hospital data. Biometrics Bulletin, 2, 47-53, 1946.
[5] 丹後俊郎，山岡和枝，高木晴良：ロジスティック回帰分析，朝倉書店，1996.
[6] Glantz. Primer of Biostatistics, 3rd edition, McGraw-Hill,1992.
[7] Taubes,G. Special News Report: Epidmiology faces its limits. Science, 269, 14 July 1995, 164-169 (1995).