

疫学・生物統計学 (2) (公衆衛生学第 3 回配布資料)

中澤 港

2017 年 10 月 22 日

1 因果関係

- 要因 (リスク因子) → 影響 (病気) の関係*1
- 完璧な証明は困難
- 統計学的な相関関係は何らかの関連を示唆する……が、
 - － 因果の向きは不明。
 - － 無関係なのに偶然相関がでてしまったかもしれない (第一種の過誤)
 - － 見かけの関連かもしれない
 - － バイアスや交絡 (後述) のせいかもしれない

もっとも強い因果関係は、要因が影響を起こすメカニズムが生物学的に明らかであり、それが常に成り立つ場合にいえる (生物学的因果関係)。疫学の究極の目的は、それを明らかにすることともいえる。

1.1 生物学的因果関係を導く指針 (不完全)

生物学的因果関係の存在を示す指針としては以下のものが知られている。しかし時間性を除きすべて例外がある。

- 科学的常識
- Henle-Koch の 4 原則 [(1) ~ (3) を 3 原則という]
 - (1) その病原体が当該感染症患者から分離される
 - (2) その病原体は他の疾病患者には見出されない
 - (3) 患者から分離培養された病原体が実験動物に同一疾患を発生させる
 - (4) 当該罹患動物から再び同一の病原体が分離される
- 動物実験・実験室の事実
- 病理学的事実
- 観察疫学的事実 (ただし、時間性以外はすべて例外があることに注意)*2
 - (1) Strength 関連の強さ。喫煙していない医師の死亡率が 0.07/1000/年、一日 1-14 本喫煙する医師の死亡率が 0.57/1000/年、一日 25 本以上喫煙する医師の死亡率が 2.27/1000/年というデータから、死亡率が 1-14 本の喫煙で 8 倍になり、25 本以上の喫煙で 32 倍になるといえる。喫煙と死亡の間には強い関連があると考えられる。
 - (2) Consistency 一貫性。その関連は異なる人、異なる場所、異なる状況、異なる時点でも一貫して見られるか?
 - (3) Specificity 特異性。特定の作業に従事する労働者に特定の部位の疾病が起こりやすいという関連性がある、その部位の疾病が他の労働者には稀であるならば、その作業への従事はその疾病の原因であるという因果関係を考える根拠となりうる。
 - (4) Temporality 時間性。原因は必ず時間的に結果より前に起こっているはずで、これは例外がない。
 - (5) Biological gradient 用量反応関係が成立しているか。それが真に病気の原因であれば、曝露の量が多かったり頻繁なほど、その病気にかかる確率は高まるはず、と考えられる。
 - (6) Plausibility 蓋然性。もっともらしさ。
 - (7) Coherence 整合性。他の知見と矛盾しないか。
 - (8) Experiment 実験あるいはそれに準ずる方法で証明できるか。
 - (9) Analogy 似た現象から類推できるか。
- 介入研究・実験疫学・臨床試験による
- メタアナリシスで複数の研究をまとめて解析→共通して同じ関係検出

1.2 個人レベルでの因果関係は立証可能か?

例えば、中澤は、ソロモン諸島の主なマラリア媒介蚊が踝から下を吸血するので、ソロモン諸島で調査するときは、ずっと靴下をはいている。いまのところマラリアに罹ったことはないが、それが本当に靴下をはいているから吸血予防ができていて罹らないのか、それとも、仮に靴下をはいていなくてもマラリアには罹らなかったのかは、靴下をはかなかった中澤が存在しない以上、わからない。

つまり、個人レベルでの因果関係は立証不可能といえる。こういう考え方を反事実 (counterfactual) モデルという。つまり、目の前の患者さんの病気について、因果関係を確定することはできない。できるのは、正しい可能性が高い因果推論 (causal inference) をすることだけ。

1.3 集団レベルでの因果推論のロジック

- 喫煙していて肺がんにかかった A さんが、もし喫煙しなかったら、という反事実は観察不能
- *しかし* 喫煙という曝露 (exposure) 条件をもつ集団 A に対して、喫煙以外の条件がほとんど同じ集団 B は設定可能
- ということは、集団 A と集団 B を追跡し、集団間で肺がんの発生率を比較することも可能

*1 その要因が原因となって、結果としての影響が表れる、という意味。

*2 <http://www.edwardtufte.com/tufte/hill> を参照。Sir Austin Bradford Hill が “The Environment and Disease: Association or Causation?,” *Proceedings of the Royal Society of Medicine*, 58: 295-300, 1965. で提示した、通称 Hill の基準

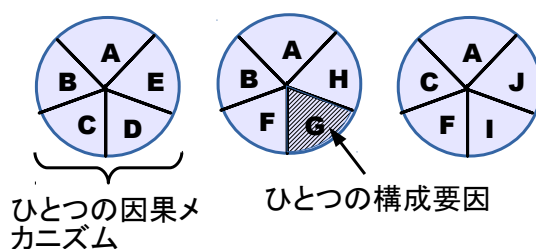
- ⇒集団レベルの因果関係がいれば、個人でもその可能性が高いだろうと推論できる
- 実は、この集団 A と B が、「喫煙という要因が肺がんという疾病に影響する関係」における、曝露群と非曝露群（対照群）
- 「喫煙という曝露条件以外の条件がほとんど同じ」になるように非曝露群を選ぶ操作をマッチング^{*3}と呼ぶ。このようなコホート研究では、マッチングは背景因子を揃える（曝露の有無以外の条件がほとんど同じと仮定できる）ことになるので有益。ただし、症例対照研究では、患者群に対して対照群（非患者群）をマッチングすると、母集団を正しく代表しなくなる危険があるので、必ずしも勧められない。

1.4 いろいろな因果関係

- 一要因⇒一症状：(例) 交通事故⇒外傷による死亡。
- 一要因⇒多症状：(例) 鉛の吸収過剰：腎臓、肝臓、中枢神経系や骨に蓄積し、食欲不振、筋肉の痛み、腹痛、不妊、脳疾患（鉛エンセファロパシー）、慢性腎炎などを引き起こす
- 多要因⇒一疾病：(例) 肝細胞がんが発生するには、B型肝炎やC型肝炎への罹患、多量飲酒や喫煙など、多くの要因が積み重なることが寄与する。
- 多要因⇒多疾病：(例) 糖尿病や冠動脈疾患など、多くの慢性疾患は、高血圧、メタボリックシンドローム、不摂生、ストレスなど多くの要因が複雑に絡み合って発症に至っている。

1.5 因果関係の整理～因果パイモデル

- 複数の因子を含む因果関係では、その因子の組み合わせは一通りとは限らない。
- それらの因子が揃えば必ず疾病が起こるという条件の組を十分要因群 (sufficient causes) という。十分要因群を構成する個々の要因を構成要因 (component causes) という。
- 十分要因群の組を円グラフの形で表したものを因果パイモデルと呼ぶ（下図は、ある疾病を引き起こす3種の十分要因群；出典は Rothman, 2002）



1.6 因果パイモデルの効用

ある疾病にかかわるすべての因果パイがわかれば、以下の利点がある。

- すべての因果パイを相互に比較する
- 各要因の相対的重要性を評価できる^{*4}

1つの因果パイについてみると、

- 最初の要因が作用してから、最後の要因が作用して発症に至るまでの時間（誘導期間=induction period）を評価できる
- 感染症に比べ慢性疾患では一般に誘導期間が長い

2 因果関係を歪めるもの～測り間違い

誤差

交絡と誤差が因果関係を歪める。誤差には偶然誤差と系統誤差がある。偶然誤差には測定機器の精度の低さ、サンプルサイズの小ささからくる標本誤差等が含まれる。系統誤差には選択バイアスと情報バイアスが含まれる。偶然誤差が大きいと、真値の推定値としての測定値の信頼区間の幅が広くなり、系統誤差が大きいと、測定値の真値とのズレが大きくなる。

2.1 何が因果関係を歪めるのか？

因果関係はさまざまな要因によって歪められ、見えにくくなっていることが多い。多くの場合、注目している要因以外の要因の作用によって関連が歪められている（交絡が存在する）だけでなく、測定や評価が正しくないこと、つまり誤差の影響を受ける。

誤差にはランダムな誤差（偶然誤差）と系統誤差（バイアス）がある。この2つを区別し、かつ、どういうメカニズムで誤差が起こるのかを整理しておく必要がある。

ランダムな誤差が大きい測定とは、精度の低い測定ともいえるが、統計学的には、測定により得られた推定値の信頼区間の幅が広い測定である。検査機器や技術の改善によって減らせるし、サンプルサイズを大きくすることでも減らせる。

^{*3} 細かくいうと、個人単位のマッチングと集団レベルのマッチングがあり、個人単位の場合も曝露群1人に対して非曝露群が1人だったり複数人だったりするなど、いろいろなやり方がある。

^{*4} 一つの考え方としては、多くの因果パイに含まれる要因ほど重要と考えられる。

系統誤差（バイアス）は、個々の測定値が真値から同じ向きに偏ることである。ゼロ点の合っていない機械による測定値は、いくら精度が高くてもバイアスが大きくなるし、バイアスがあったら、いくらサンプルサイズを大きくしても真値からの推定値のズレは小さくならない。バイアスには、大別すると、選択バイアス（selection bias; 観察対象が母集団を正しく代表しないこと）と、情報バイアス（information bias; 観察対象から得る情報が偏ってしまうこと）がある。

2.2 交絡 (confounding)

交絡とは、注目している因果関係を歪める第3の因子。交絡要因は、以下の3条件（例：肥満⇒高血圧に対する年齢）を満たす。

- 注目している要因（肥満）ではない
- 注目している要因（肥満）と結果である健康影響（高血圧）の両方と因果的に関連
- 要因曝露の結果ではない（肥満が高齢をもたらすのではない）

2.3 交絡の制御

交絡を制御するには、デザインによる制御と分析段階での制御がある。例えば、性別や年齢による交絡がありそうな因果関係についての研究を行う場合に、性別と年齢で層別した層別無作為抽出をすることや、コホート研究であれば曝露群と非曝露群について性別と年齢のマッチングを行うことや、特定の性別・年齢に限定して研究を実施することは、デザインによる制御である。

分析段階の制御には、下記のように、層別解析、標準化、プール化、多変量解析がある。

層別解析 交絡因子によって層別し、別々に解析すること。年齢による交絡があるなら、年齢階級別に死亡率を比べてみるなど

標準化 交絡因子について共通の重みを設定し、標準化した指標を求めて比較すること。例えば、年齢調整死亡率など。

プール化 交絡因子によって層別した各層に共通の関連性がみられると想定し、関連性を要約すること。マンテルハントウの要約オッズ比など。

多変量解析 統計学的に、説明変数群の中に、交絡が疑われる変数を共変量として投入することによって、交絡の影響を制御することができる。ロジスティック回帰分析などがよく行われる。

3 疫学研究の歴史と研究倫理

3.1 古典的疫学研究

- Panum PL (1820-1885) の麻疹の研究：デンマーク領ファロー島はそれまで 65 年間なかった麻疹が 1846 年に大流行したので、Panum は数千人の患者を診察して、大流行の原因、感受性と終生免疫、接触から発症までの期間、感染可能期間などを明らかにした。
- Gregg NM：1941 年にシドニーなどで多発した先天性白内障の観察から、その前年に起こった風疹流行と母親の当該児妊娠初期が一致することを発見した。
- 高木兼寛 (1849-1915) の脚気の研究：1884 年に脚気の原因が食事の欠陥にあると見当をつけ、脚気が多発していた海軍の軍艦乗組員の遠洋航海の際の食事を調べて、大麦、大豆、牛肉を増やすことによって、炭素に対する窒素割合を増やしたら、脚気が減った。真の原因がビタミン B1 不足であることまではわからなかったが、この介入研究によって、食事に原因があることは明らかになった。
- Snow J (1813-1858) のコレラの研究：当時コレラ菌は未知だったが、ロンドンではコレラが日常的に流行していた。1854 年の大流行時に、John Snow は、2つの研究によって、飲み水によってコレラが伝播している可能性が高いことを示した。1つは、コレラ死亡者の発生地図 (spot map) と発症日別のコレラ死者数の度数分布を作成した記述疫学研究により、コレラ流行の原因がブロードストリート地区の1つの共同井戸利用にあったことを突き止めたこと。もう1つは、水会社間の供給人口当たりのコレラ死亡数の比 (L社：461/173748 に対して S&V 社：4093/266516 が約 5.8 倍) を計算し、テムズ川下流域から取水している S&V 社から水供給を受けている人がコレラに罹りやすいことを示した自然実験研究。

3.2 Frammingham Heart Study (フラミンガム研究)

- 米国東部マサチューセッツ州ミドルセックス郡フラミンガム町
- 人口：65598 人 (2005 年)
- 1948 年 (米国民の死因の 8 割が循環器疾患だった頃) 開始
- 疫学的手法を循環器疾患に適用した最初の研究
- オリジナルコホート (1948 年) 5209 人、子供世代コホート (1971 年) 5124 人、第三世代コホート (2005 年) 4095 人。それぞれはクローズド・コホート。町全体としてみればオープン・コホート。
- 検査項目：
(オリジナルコホート 1998 年時) 身体測定、血圧、尿検査、生活習慣、病歴聴取、MD 検査、生化学検査、心電図、X 線撮影、肺機能検査、身体機能、認知機能、心理社会的状態
(子供世代コホート) X 線撮影を除く上記項目+心エコー、頸動脈エコー、ホルター心電図
- <http://www.framinghamheartstudy.org/>

3.3 久山町研究 (http://www.epi-c.jp/e001_1.0001.html)

- 福岡県糟屋郡久山町 (人口 8075 人: 2007 年 1 月 1 日)
- 1961 年～、40 歳以上の男女を対象としたコホート研究
- 登録数：第 1 集団 1621 人 (1961 年～)、第 2 集団 2038 人 (1974 年～)、第 3 集団 2637 人 (1988 年～)、第 4 集団 3500 人 (2002 年～)
- 定期健診 (通常毎年+一斉 5 年毎)：アンケート調査、食事調査、身体測定 (身長、体重、皮下脂肪厚、握力)、尿検査 (pH、蛋白、糖、潜血、ケトン体、ウロビリノーゲン)、血圧測定 (坐位、臥位)、血液検査 (肝機能等 31 項目)、貧血検査、糖負荷試験、心電図、眼底検査、問診
- 特徴：全国平均とほぼ同じ年齢・職業分布～平均的な日本人。剖検率が高い (1964 年 100%、その後も 80%)、追跡率 99%、一斉健診 78% 受診
- 1961 年当時、日本は脳卒中が死因の 1 位。欧米と違って脳出血が脳梗塞の 12.4 倍。欧米の研究者から死因への疑いが呈され、剖検に

より正しい死因を確定する目的で開始。後に心血管病、ゲノム疫学へ展開

- 主な結果：脳出血による死亡は脳梗塞による死亡の1.1倍（第1集団）。脳卒中の最大の危険因子が高血圧であることを解明、保健指導（全ての健診結果が出た時点で医師が個別指導）＋降圧剤服用で脳卒中死亡を大幅に減らした介入成果、等

3.4 INTERSALT 研究

INTERSALT 研究は、個人ベースで多くの国と地域について実施された横断的研究だが、地域集団ごとに集計した値を単位とした地域相関分析によってもさまざまな関連性が示されている。

- 32 国 52 集団（うち日本 3 集団）、20-59 歳の男女、1 集団 200 例ずつ、1985 年に横断的に実施
- 登録数 10079 例（男性 5045 例、女性 5034 例）
- 調査項目 血圧（5 分間の安静ののち、ランダムゼロ血圧計を用いて座位で 2 回測定）、24 時間蓄尿（尿量、ナトリウム、カリウム、マグネシウム、カルシウム）、心拍数、運動状況、飲酒、喫煙、服薬状況、年齢、身長、体重、人種、教育、経産歴（女性のみ）、閉経状況（女性のみ）。
- 主な結果：地域相関で、BMI \propto 血圧、Na/K 比 \propto 血圧、塩分摂取量 \propto 血圧、ヤノマミなど低塩分摂取集団では加齢に伴う血圧上昇なし、等

3.5 代表的な症例対照研究

- SMON についての椿忠雄の研究⇒キノホルム服用が原因
- 四肢の全部あるいは一部が短い奇形をもつ新生児の多数発生⇒サリドマイド服用が原因

3.6 人を対象とする医学系研究に関する倫理指針

- 2002 年に文部科学省と厚生労働省が合同で、「疫学研究倫理指針」を発表し、2007 年全面改訂され、2008 年 12 月にさらに一部改訂された。その後、2014 年 12 月 22 日に臨床研究に関する倫理指針と統合され、「人を対象とする医学系研究に関する倫理指針」となり、2017 年 2 月 28 日に一部改正された*5。医学研究倫理についてのヘルシンキ宣言や患者の権利についてのリスボン憲章に依拠している。最近の改正は個人情報保護法との関連が大きい*6。
- 疫学研究は人間を対象とするので、倫理面での配慮が不可欠
- とくに介入研究では曝露条件をセッティングするので、十分に統制された実験をする必要がある
- 観察的研究や記述疫学研究であっても、プライバシーへの配慮が必要
- 文書によるインフォームドコンセントが多くの場合に必要（侵襲を伴わない場合は不要とされるが、得ておくほうが無難）
- 倫理審査委員会による審査を通らないと研究できない（成果を学術雑誌に投稿しても事務的に受け付けを拒否されてしまう）

4 実験計画における Fisher の 3 原則

反復 1 つの処理に対して少なくとも 2 回以上の繰り返しが必要

無作為化 実験の順序や空間的にどの場所にどの実験群を割り当てるのかを無作為に決める必要がある。

局所管理 実験が大規模で、実験全体を無作為化するのが妥当でないとき、実験をある程度細分化してブロックを構成し、ブロック内で処理条件を無作為化し、ブロック内のバックグラウンドが均一になるよう管理すると、系統誤差の一部がブロック間変動として除去できる。

4.1 実験計画におけるサンプルサイズの設計原則

- 臨床試験の主要なエンドポイント（評価項目）と統計解析の方法が決まっていなくてはならない。
- 割合を比較する場合なら (1) 有意水準、(2) 検出力、(3) コントロール治療での臨床イベント発生割合、(4) 試験治療のイベント発生割合がコントロール治療よりどれくらい小さければ臨床的に意義があると考えられるか、その最小の値
- 例えば有意水準が片側 5%、検出力が 80%、コントロール治療でのイベント発生が 30%、試験治療により期待できるイベント発生が 15% の場合なら、各群 95 人となる*7。

4.2 毒性試験

- 化学物質などについて生物に好ましくない作用の有無またはその強さの程度を調べるための試験
- 試験は、評価する毒性の項目（一般毒性、特殊毒性）、使う生物の種類（哺乳動物、魚など）と形態（全体、組織、細胞など）、曝露経路（経口、吸入、経皮など）、曝露期間（長期、短期など）によって様々な種類がある。目的によって、適切な試験方法を決定する必要がある
- とくに、用量反応関係（量-反応関係ともいう。dose-response relationship）については多くの方法が開発されてきた

4.3 臨床試験

- ヒトに対する実験（侵襲あり）を臨床試験という

*5 <http://www.mhlw.go.jp/file/06-Seisakujouhou-10600000-Daijinkanboukouseikagakuka/0000153339.pdf>

*6 厚生労働省の説明資料：<http://www.mhlw.go.jp/file/06-Seisakujouhou-10600000-Daijinkanboukouseikagakuka/0000170955.pdf>

*7 R では、

```
power.prop.test(p1=0.15, p2=0.3, sig.level=0.05, power=0.8, alternative="one.sided")
```

で OK。

- 新薬とか新しい治療法は、モデル動物で効果があるだけではダメで、どうしてもヒトに効くか、ヒトに有害作用がないかどうかを確認する必要があり、臨床試験は必須
- 無駄になってはいけないので、科学的かつ倫理的に考え抜かれた計画に従って行われねばならない。様々なガイドラインがあって、それに沿って計画する必要がある。
- きちんと計画しない臨床試験は許されない（後になってから、サンプルサイズが不十分だったので有意差が見られませんでした、ではダメ）

4.4 臨床試験の4段階

- 前段階（非臨床試験）：細胞，組織，動物を使った実験（主として上述の毒性試験）＝安全性を確かめる
- 第I相試験：健康な成人のボランティアを対象として，薬物動態や最大許容量を調べる
- 第II相試験：比較的少数の患者を対象として，有効性，安全性，用量反応関係を調べる
- 第III相試験：数百から数千の患者を対象として，「薬の候補」の有効性を，科学的に検証する目的で行う。通常は **RCT (Randomized Controlled Trial)** になる。
- 第IV相試験：市販後に大勢の患者が実際に服用した結果，新薬がどういう特徴をもっているか，副作用はでないかを調べる（市販後臨床試験）

4.5 臨床試験における倫理的要求

- ヘルシンキ宣言や，それに基づいて各国が出している倫理指針が大原則。しかしこれだけでは具体的にどうすればいいのか曖昧。
- 具体的には，日米欧による International Conference on Harmonization (ICH) という会議により，いくつかのガイドラインが公表されている。
- 医薬品の臨床試験の実施の基準に関する省令 (GCP)
- 臨床試験のための統計的原則について（統計ガイドライン）
- 臨床試験における対照群の選択とそれに関連する諸問題（対照群に関するガイドライン）

4.6 臨床試験の手順

試験実施計画書の作成 実験なので当然。

臨床研究倫理審査 IRB (Institutional Review Board) による倫理審査により研究実施について承認を受ける必要がある

試験の実施 試験実施計画書に沿って実施するのは当然。ただし，有害作用に苦しむ患者が発生した場合，同じ治療を続けることは倫理的に正しくないので，計画書からの逸脱が正当化される（省令 GCP 第 46 条）

データ解析 薬の候補を使う，使わないはランダムに割り付けられるが，倫理的問題を含めて割り付けが守られない（服薬コンプライアンスが悪い）場合があるのが問題。その場合の扱いとしては，実際の服薬にかかわらず割り付け通りに解析する ITT (intention to treat / intent to treat) にするのが原則*8。

4.7 ITT の理屈

以下，ITT の理屈を佐藤俊哉，松岡淨 (2001)*9 に基づいて説明する。臨床試験で A か B の 2 つの処理をランダムに割り付けるとして，実際には割り付けが 100% 守られるとは限らない。結果として以下 4 群が生じる。

- Aa 割り付けが A，実際に使ったのが A
- Ba 割り付けが B，実際に使ったのが A
- Ab 割り付けが A，実際に使ったのが B
- Bb 割り付けが B，実際に使ったのが B

全員が Aa または Bb なら問題ないが，実際は Ba や Ab の人が生じる。誰と誰を比べればいい？

- 計画書を守った人たちだけを比べる Aa を A 群，Bb を B 群とする。これは，治療効果が出たり副作用が出たりした人が減るという選択バイアスがかかるのでダメ。
- 実際に使われた処理で比べる Aa + Ba を A 群，Ab + Bb を B 群とする。計画書を守っていないのでダメ。Ab を Bb と一緒に B 群に入れると，A が効かなかった人が B 群に入る傾向がでて，やはり選択バイアスがかかる。
- 実際の使用は無視して，割り付け通りに比べる Aa + Ab を A 群，Ba + Bb を B 群とする。これが正しい。なんで？

なぜ割り付け通りに比べる ITT が正しいのか？

- 割り付けを守らなかった Ab と Ba には理由がある⇒中には飲み忘れや交通事故にあって服薬を続けられなくなったなど，治療効果と関係ない場合もあるが，多くの場合ランダムでない（治療が効かないから止めた，治療が効いて良くなったから止めた，副作用が出た，等）。
- A が新薬，B がプラセボだとして，
 - － 対立仮説「A が有効」の下では，Aa+Ba を A 群とすると治療効果を良く見せる。Aa+Ab を A 群とすると治療効果を薄める。
 - － 帰無仮説「A は効果なし」の下では，Aa+Ba を A 群とすると治療効果を良く見せるが，Aa+Ab を A 群とすると，正しく「効果なし」と判定できる。
- 計画書を守った人だけ解析したり実際に受けた治療で群分けして解析すると，必ずバイアスがかかるが，割り付け通りなら，帰無仮説の下では正しい解析ができるので，第一種の過誤を守る。

言い換えると，ITT をすると差が薄まる場合があるけれども，差が無い場合は正しく解析できる。本当は効かない薬が効いたと判定してしまうよりは，この試験では統計学的に有意な差は検出できなかったという方がマシ。両方やって，一致した結果が得られれば OK。違うときは ITT の方がよさそうというのも一つの判断。

*8 ただし，ランダム割り付け後に実験参加不適格であることが判明したとか，1度も薬を飲まなかったとか，ランダム割り付け後のデータが一切ない人については，解析から除外してもいい場合があり，その場合は，残りの「最大の解析対象集団」について，割り通りに解析する。

*9 <http://www.kbs.med.kyoto-u.ac.jp/01Sep15.pdf>

4.8 欠損値はどうか

- ITT の考え方から、最大の解析対象集団を解析するとして、欠損値はどうしたらいいのか？
- 欠損の種類により異なる
 - － 完全にランダムな欠損なら検出力が落ちるだけでバイアスは問題にならない
 - － 欠損値のある変数についてランダムな欠損で、他の変数についてランダムでない場合は、多重代入法 (Multiple Imputation) の利用を考慮する*¹⁰。
 - － ランダムでない欠損はいかんともしがたい

4.9 治療効果の判定指標 (その意味)

- 相対リスク減少率 (1 - リスク比)
- 絶対リスク減少率 (リスク差 = 超過危険 = 寄与危険)
- 要治療数 (NNT = 絶対リスク減少率の逆数)

(例) コレステロール低下薬の効果

5年間プラバスタチンを服用した1000人のうち32人が冠動脈疾患で死亡、偽薬を飲んでた1000人のうち41人が死亡。新聞報道は、「プラバスタチンを飲むと死亡リスクが22%低下した」
一般市民の多くは、プラバスタチンを飲むと1000人の高コレステロール血症患者のうち220人が心臓発作を免れると誤解したが、それは正しくない。

(Quiz) 3つの判定指標を計算してみると？

1000人中32人の死亡と1000人中41人の死亡の比較という点は同じ。

相対リスク減少率 $1 - (32/1000)/(41/1000) = 1 - 32/41 = 9/41 = 0.2195...$

*新聞報道の「22%」はこれだった。

絶対リスク減少率 $(41/1000) - (32/1000) = 9/1000 = 0.009$

*この値を使うなら、「0.9%の死亡を防げる」あるいは、「1000人服用したとき9人の死亡を防げる」といえる⇒効果は劇的に見えないが誤解も少ない

NNT $1/(9/1000) = 1000/9 = 111.11...$

*この値を使うなら、「111人服用して初めて1人の死亡を防げる」といえる⇒より誤解されにくい

5 スクリーニング (Screening) とは？

集団を対象に、「迅速に実施可能な検査、手技を用いて、無自覚の疾病または障害を暫定的に識別すること」

スクリーンとは元々「篩に掛ける」ことを意味する。疫学や公衆衛生学でスクリーニングといえば、一般には健診や集団検診のこと。

5.1 スクリーニングの目的

究極の目的は、集団全体としての死亡率を下げる（普通は発生率は不変）。

主目的 疾病の早期発見、早期治療 = 二次予防

他の目的 ある疾病のハイリスク群の検出 = 一次予防、集団の特性の把握「集団診断」等

5.2 事例：米国乳がんスクリーニング

- 複数の機関による矛盾する勧奨/知見も食い違い
- 有病割合の低さと偽陽性の多さによる不利益
- 2009年11月のUS PSTF (preventive services task force) 勧告を巡る論争：勧告は、40代は定期的マンモグラフィ不要、50-74は2年おきのマンモグラフィ、75歳以上は不要、マンモグラフィと視触診の併用は無意味、自己触診は過剰診断をもたらす、となっていたため、反論多々。

5.3 スクリーニングにおけるバイアス

スクリーニングにおけるバイアス

自己選択バイアス、リードタイムバイアス、レングスバイアスがあることに注意

5.4 スクリーニング実施の原則 (Wilson JMG and Jungner G, 1968)

1. 目的とする疾患が重要な健康問題である
 - その疾患の頻度が高い (eg. がん検診)
 - 緊急な対処が必要 (eg. 新生児代謝異常検査)
2. 早期に発見を行なった場合に、適切な治療法がある (治療法がないと「負のラベリング効果」になることがあるため、スクリーニングはしない)
3. 陽性者の確定診断の手段、施設がある

*¹⁰ 理屈は難しいが、Rではmiceなどのパッケージで可能。伝統的に単純代入法として用いられてきた、最悪値の代入、直前値の代入、グループ間の差が縮まる値を代入、などは非推奨。

4. 目的とする疾病に潜伏期あるいは無症状期がある
5. 目的とする疾病に対する適切なスクリーニング検査法がある（「適切な」は、費用や判定に要する時間も含む）。
6. 検査方法が集団に対して適用可能で、受け入れやすい
7. 目的とする疾病の自然史がわかっている
8. 患者、要観察者に対する追跡 follow-up システムが確立している
9. スクリーニング事業の費用－便益が成立する
10. スクリーニングの意味、内容が受診者に周知されている（受診者が過度な期待をしないことにより、不要な争いを避けることができる）

5.5 現代における新しいスクリーニング基準

Andermann ら (2008) は、とくに遺伝的スクリーニングの問題を強調している。

遺伝情報の特殊性 (玉井, 2006 年) 不変性・予見性, 個人特異性, 家系内共有性 (eg. 「重い障害を持つ子どもがいる。次は健康な子がほしい。でも、上の子の存在を否定しているような気がして、出生前診断を受けることに罪悪感を覚えてしまう。似たような状態にある人はみんな受けてるじゃないか、と言いつつ考えている自分がいやだ。でも、もうひとり同じ障害の子が生まれたら、育てていけるだろうか」)

特有の倫理的問題 優生思想につながりやすい。健康増進と障害者福祉の極端な場合の衝突 (DALYs の問題点と絡む)

インフォームドチョイス 受診者が十分な情報に基づいて受診するかどうかを選択する (インフォームドチョイス) べき。信頼できる情報源としてプライマリケア医が重要

5.6 スクリーニング方法の評価

スクリーニングの性能評価指標

有効性 (感度, 特異度) と信頼性 (再現性) の他, 迅速性, 簡便で安価であることが求められる

	疾病	健康
検査陽性	a 人	b 人
検査陰性	c 人	d 人

感度 $\frac{a}{(a+c)}$

特異度 $\frac{d}{(b+d)}$

陽性尤度比 $\frac{(a/(a+c))}{(b/(b+d))}$ 感度を (1 - 特異度) で割った値。

陰性尤度比 $\frac{(d/(b+d))}{(c/(a+c))}$ 特異度を (1 - 感度) で割った値。

これらと紛らわしいが, 実際スクリーニングを行った場合に得られる判定的中度は, 対象となる集団における有病割合に依存するので, 混同してはならない。前述のものと同じ 2×2 の分割表を使って説明すると (ただし, この場合は実際のスクリーニングなので, 事前には疾病の有無はわからず, 事後的に精密検査等で疾病の有無を確定して得られる数値と考えて), 以下の通り。

陽性反応的中度 (positive predictive value) $\frac{a}{(a+b)}$

陰性反応的中度 (negative predictive value) $\frac{d}{(c+d)}$

5.7 信頼性

スクリーニングの信頼性は, 検査再検査信頼性 (test-retest reliability), 施設間差 (inter-institute difference) が小さいこと, 測定者間差 (inter-rater difference) が小さいことなどを指す。

test-retest reliability は, 同じ対象者に同じ検査をしたときに結果が一致すること。κ 係数 = 実際の一致率と偶然の一致率の差を, 1 と偶然の一致率の差で割った値。完全一致のとき 1, 偶然の一致と同じとき 0, それ以下で負となる。

5.8 連続量の測定における ROC 分析

- 1 つのスクリーニング方法について陽性・陰性の基準値を最小値から最大値まで段階的に変えると, 偽陽性率 (= 1 - 特異度) も感度も変化
- 基準値 = 最小値で感度も偽陽性率も 1, 基準値 = 最大値で感度も偽陽性率も 0
- 偽陽性率を横軸, 感度を縦軸にとって線で結ぶと, 基準値の変化に対応する「ROC 曲線」になる。
- 左上 (0,1) 近くを通るほどスクリーニング有効性が高い (実際は費用など他の条件も考慮)
- 曲線上で一番左上に近い点が最適なカットオフ値といえる
- 最適カットオフ値の探索にも使えるし複数のスクリーニング方法の優劣の判定にも使える
- ROC 曲線が左上近くを通ることと AUC (Area under curve; 曲線下面積) が大きいことは同値

5.9 新生児マススクリーニング検査

2003 年遺伝医学関連 10 学会発表『遺伝学的検査に関するガイドライン』の指針には, 下記のように定められている。

1. 新生児の先天性疾患の早期診断・早期治療による発病率・死亡率の低下を目的
2. 医師は検査の意義を積極的に両親に説明し同意を得て検査実施に努め, 必要に応じて遺伝カウンセリングを考慮する

発達障害については, 1 歳 6 カ月児健診でも 3 歳児健診でも 84~85% の自治体で実施されている。ただし, そのうち 8 割は問診の中で聞き取っているだけ。13~14% の自治体は発達障害の疑いのある児に特別な検診を用意している。

先天性代謝異常等については, ハイリスクグループ検出と早期治療が目的で実施される。

5.10 特定健診：高齢者医療確保法による

- 40歳～74歳を対象
- メタボリック・シンドロームを検出する目的
- 内科系7学会合同提案によるスクリーニング基準値（必須項目として男性85cm以上、女性90cm以上の腹囲または内臓脂肪断面積100平方cm以上+以下のうち2項目以上：TG150mg/dL以上、HDLコレステロール40mg/dL未満、SBP/DBPが130/85mmHg以上、空腹時血糖110mg/dL以上）の信頼性・妥当性には反対意見多数あり（まだ変わっていない）
- 中年以降の男性の半数が予備軍になってしまうのでハイリスクグループ検出という意味は無い
- 特定保健指導の有効性にも疑問があるのでポピュレーション・アプローチとしても疑問がもたれている

5.11 がん検診：がん対策基本法による

- 部位別にさまざまな検査がある
- 2007年の受検割合は2～3割
- 国立がんセンター「がん検診ガイドライン」*¹¹
 - 子宮頸がん 細胞診のみ、相応な根拠あり（HPV検査は子宮頸がん死亡率減少効果の有無を判断する証拠が不十分）
 - 胃がん X線検査と内視鏡検査は、死亡減少効果を示す相応な根拠あり
 - 大腸がん 便潜血検査は十分な根拠あり。大腸内視鏡検査は死亡減少効果を示す根拠はあるが無視できない不利益があるため対策型検診は非推奨
 - 肺がん 胸部X線（高危険群に対しては喀痰細胞診併用）は相応な根拠あり
 - 乳がん 40～74歳でマンモグラフィ単独で死亡率減少効果を示す相応な根拠あり、偽陽性、過剰診断、放射線誘発乳がんの発症の可能性という不利益もあるが対策型検診推奨。40～64歳については視触診併用も同様の効果と不利益を示す。

*¹¹ <http://canscreen.ncc.go.jp/guideline/matome.html>