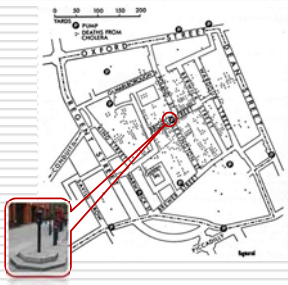
アウトカムの信頼性と妥当性

慶應義塾大学医学部クリニカルリサーチセンター
中川敦夫

2014.2.28

ブロード・ストリート事件　1854年

10日間で500人死亡



1854年 Broad Street周辺のコレラ死亡発生状況

当時、コレラは「空気感染」すると考えられており恐れられていた。

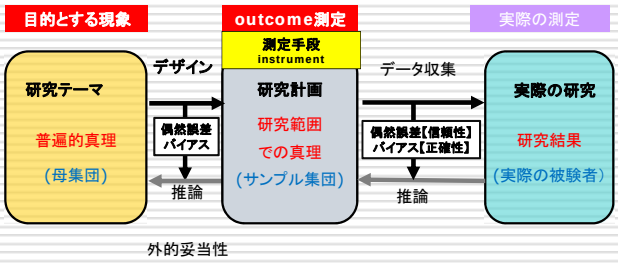
1854年 John Snowは患者発生状況を数え（測定）、ある井戸水を飲んでいる人がコレラに罹患すると結論

1883年 ロベルト・コッホがコレラ菌を発見

生物学的要因（病原体など）が不明であっても、社会的要因、状況の観察/測定から、感染症流行を止めることができた

測定measurement

目的とする現象を数値化するプロセス



目的とする現象は何か？ whose interest?

- 患者
- 介護者
- 医療者
- 研究者
- 政治家、行政
- 企業などのsponsor

目的とする現象は何か？ アウトカム　の　6D

アウトカム	
死亡　death	
疾患　disease	症状、臨床化学検査値異常など
不快　discomfort	痛み、吐き気、呼吸苦、掻痒、耳鳴
障害　disability	日常生活の機能、就労や余暇
不満　dissatisfaction	疾患やその介護に伴う感情(満足度)
貧困　destitution	個人や社会の疾患に伴うコスト

The six D's

抽象的な特性は複合的要素の組み合わせによって測定

うつ病治療における患者の求めるアウトカム

うつ病外来患者自身が「寛解」と判断するfactor

TABLE 1. Factors Identified by Depressed Outpatients as Very Important and Most Important in Determining Remission From Depression

Factor	Patients Identifying the Factor as Very Important (N=535)[a] N	%	Patients Identifying the Factor as Most Important (N=487)[b] N	%
Absence of symptoms of depression	375	70.6	51	10.5
Presence of positive mental health (e.g., optimism, vigor, self-confidence)	412	77.3	81	16.6
Not getting overwhelmed by stress	272	50.9	9	1.8
Coping well with stressful events	311	58.5	11	2.2
Able to cope with the normal stress of life	360	67.9	22	4.5
Functioning well	374	70.3	10	2.0
Return to usual level of functioning at work, home, or school	397	74.3	39	8.0
Able to fulfill usual responsibilities	359	67.2	5	1.0
Feeling happy most of the time	253	47.5	18	3.7
Satisfaction with life	308	57.7	34	7.0
Feeling in emotional control	384	71.9	42	8.6
General sense of well-being	372	69.7	52	10.7
Positive outlook on life	350	65.5	41	8.4
Participating in and enjoying usual activities	362	67.7	11	2.3
Participating in and enjoying relationships with family and friends	378	70.7	20	4.1
Feeling like your usual, normal self	397	75.6	70	14.4

[a] Because of missing responses, the total number of responses for each item ranged from 533 to 535.
[b] Thirty-three of the 535 patients did not indicate which factor was most important in determining remission or more factors, leaving 487 patients with usable responses.

- 健康的なメンタルヘルス（自信、楽観）
- 通常機能への回復
- 本来の自分という感覚

Zimmerman M, McGlinchey JB, Posternak MA, Friedman M, Attiullah N, Boerescu D. How should remission from depression be defined? The depressed patient's perspective. Am J Psychiatry. 2006 Jan;163(1):148-50.

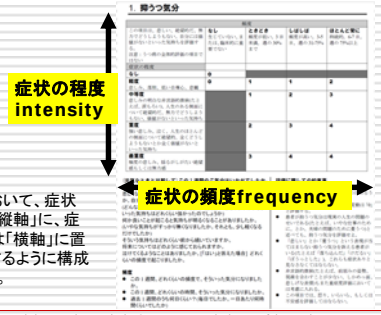
アウトカムの設定

- たくさんのアウトカムを測定したくなるが、primary outcomeは絞られているか？
- そのアウトカムは測定可能か？
  - 発生頻度は低すぎないか？差が検出できるか？
- 測定手段に妥当性はあるか？
- 測定手段に再現性はあるか？
- 測定手段の使用料や版権の問題は大丈夫か？

評価尺度instrumentは症状を定量化したものである

- 本質的には患者の主観的体験（現象）を評価尺度を用いて定量化したものである
  - 症状評価のinstrumentとしての問題
  - 臨床経験にもとづく臨床判断と必ずしも一致しない
- 評価尺度による重症性の評価は、治療選択を決定しない
  - 個別症例ごとの治療反応特異性

GRID版 ハミルトンうつ病評価尺度 (GRID-HAMD)



各項目において、症状の程度は「縦軸」に、症状の頻度は「横軸」に置いて評価するように構成されている。

Williams JB, Kobak KA, Bech P, et al. The GRID-HAMD: standardization of the Hamilton Depression Rating Scale. *Int Clin Psychopharmacol. 2008;23(3):120-129.*

GRID-HAMD：日本臨床精神薬理学会T. Furukawa, T. Akechi, N. Ozaki, N. Iwata, H. Naitoh, T. Higuchi, A. Kalali

測定手段によって得られたデータ

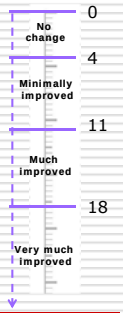
測定されたデータ （測定の尺度）

種類	特徴	例	統計量	情報量
定性的データ（カテゴリカル・データ）				power
名義変数 （2値data） **Nominal data**	**意味のない整理番号を割り当てたもの**	男女（男-1、女-2） 血液型（A-1,B-2,O-3,AB-4） 寛解の有無（Y-1、N-2）	割合	少ない
順序変数 **Ordinal data**	**順序には意味があるが間隔には意味がない**	オリンピックのメダル （金-1、銀-2、銅-3） 痛みの程度 （Wong Baker Face Scale）	割合 中央値	中程度
定量的データ				
離散変数 discrete	**範囲が大きいと連続変数に近い**	喫煙本数	割合 中央値 平均値 標準偏差	多い
連続変数continuous **間隔尺度** interval scale	**目盛が等間隔になっている（等間隔と仮定されている）**	気温（℃）、血圧		多い
比例尺度 ratio scale	**原点の決め方が定まっており，間隔や比率に意味が有**	身長，体重，絶対温度		多い

HAMDは 本来は順序尺度ordinal scaleである

CGI-C scale	Absolute reduction HAMD score
Very much improved	≥18
Much improved	11-17
Minimally improved	4-10
No change	0-3
Worsening	<-1

Furukawa TA, Akechi T, Azuma H, Okuyama T & Higuchi T (2007) Evidence-based guidelines for interpretation of the Hamilton Rating Scale for Depression. Journal of Clinical Psychopharmacology, 27, 531-534.

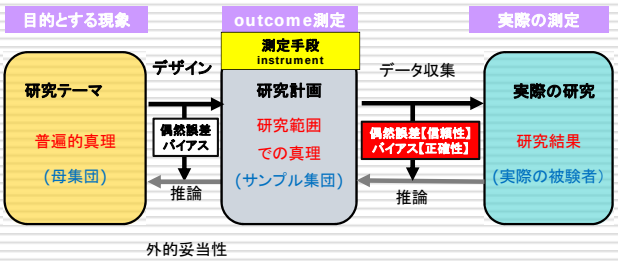


25点から 20点への減少と、7点から2点への減少は同じか？

いかに的確に目的の現象を測定するか

測定measurement

目的とする現象を数値化するプロセス



いかに的確に目的の現象を測定するか

- **定度(精度)precision**

＝再現性の高さ

  - ✓ 同義語:再現性reproducibility、信頼性reliability、一致度consistency

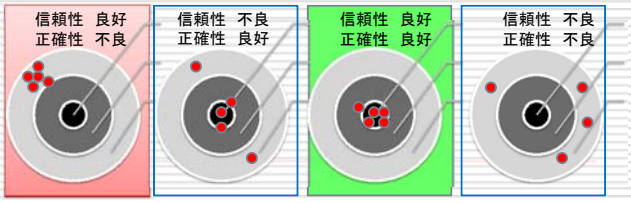
- **真度(正確性)accuracy**

＝目的とする現象(真の値)にどれほど近いか

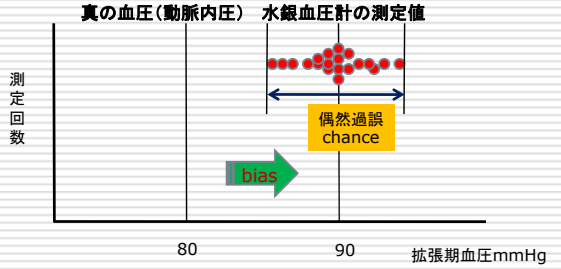
  - ✓ 妥当性validityに大きな影響を与える

定度precision(信頼性)と真度accuracy(正確性:妥当性に大きな影響)

	定度(信頼性)	真度(正確性)
定義	繰り返し測定の値が安定である度合	測定値が目的とする真の値と一致する度合
最も良い評価方法	反復測定し値の変動を調べる	Gold standardとの比較
研究にとっての意義	効果判定の検出力を高める	結論の妥当性を高める
影響する要因	偶然誤差(偶然変動)	系統誤差(bias)



バイアスと偶然過誤



定度(精度、信頼性)precision

偶然誤差(バラツキ)の影響を受ける

① **測定者による誤差 observer variability**

- 測定者自身の言葉づかい、測定機器を用いる時の操作のクセ

② **測定手段誤差 instrument variability**

- 測定環境や機器の老朽化の度合い

③ **対象者誤差 subject variability**

- 対象自身に起因するバラツキ (気分の変動や最終服薬からの時間など)

定度(精度、信頼性)precisionを高める対策

① 測定方法の標準化
- 実施マニュアル(手順書)の作成

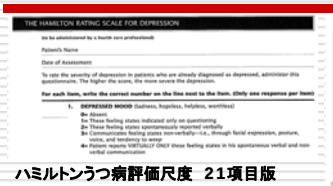
② 測定者のトレーニングと技能チェックcertification

③ 測定手段の改善
- 質問票の言い回し(clarity, simplicity, neutrality, time frame)

④ 測定手段の自動化・自記化
- 自動血圧計など

⑤ 測定の反復(困難な場合が多い)　Feasibilityとのバランス
- 2回以上測定してその平均値を用いる

ハミルトンうつ病評価尺度



ハミルトンうつ病評価尺度　21項目版

Hedlund JL, Vieweg BW (1979) The Hamilton rating scale for depression: a comprehensive review. J Operat Psychiatry 10: 149–165

アンカーポイントや標準化された質問法

GRID版 ハミルトンうつ病評価尺度

Williams JB, Kobak KA, Bech P, et al. The GRID-HAMD: standardization of the Hamilton Depression Rating Scale. *Int Clin Psychopharmacol.* 2008;23(3):120-129.

信頼性、再現性の検討法

- 再テスト信頼性
  - 同一の被験者に期間を空けて同一のテストをさせ、一回目と二回目のテスト結果を比較する
- 評価者間信頼性
  - 同じ対象を複数の評価者がテストしその結果を比較する
- 内的整合性　Cronbachのalpha係数(クロンバックのα係数)
  - 項目間の一貫性(ある質問項目でYesと回答した被験者は，同じ尺度内の別項目でもYesと回答する)

一致率の検定

- Kappa係数 :

  名義変数、名義変数

  (カテゴリカル・データ)

Kappa coefficient
0.93-1.00 excellent agreement
0.81-0.92 very good agreement
0.61-0.80 good agreement
0.41-0.60 fair agreement
0.21-0.40 slight agreement
0.01-0.20 poor agreement
<=0.01　no agreement

- Intraclass correlation (ICC)級内相関:

  連続変数

検者内信頼性 ICC(1,1) (1,K)　一元配置変量
検者間信頼性 ICC(2,1) (2,K)　二元配置変量　ランダム効果
検者間信頼性 ICC(3,1) (3,K)　二元配置混合　固定効果
K:測定回数

Cronbach's alpha

DSM-5 診断の一致率 (Kappa 係数)

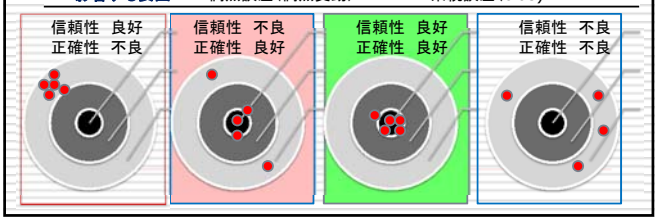
(field trial data 1593 Adults, 665 Children)
Kappa=1.0 perfect, >.8 excellent, .8-.6 good, .4-.6 moderate, .2-.4 fair (acceptable), <.2 poor

Disorder	DSM-5	DSM-IV	ICD-10	DSM-III
Schizophrenia	.46	.76	.79	.81
Schizoaffective	.50	.54	.51	.54
Autism spectrum disorder	.69	.59-.85	.77	- .01
PTSD	.67	.59	.76	.55
Child ADHD	.61	.59	.85	.50
Complex somatic disorder	.60	-	.45	.42
Major depressive disorder	**.32**	**.59**	**.53**	**.80**
Bipolar Disorder	.54	-	.69	-
Oppositional defiant disorder	.41	.55	-	.66
GAD	.20	.65	.30	.72

APA 165th annual meeting, 2012

定度precision(信頼性)と 真度accuracy(正確性:妥当性に大きな影響)

	定度(信頼性)	真度(正確性)
定義	繰り返し測定の値が安定である度合	測定値が目的とする真の値と一致する度合
最も良い評価方法	反復測定し値の変動を調べる	Gold standardとの比較
研究にとっての意義	効果判定の検出力を高める	結論の妥当性を高める
影響する要因	偶然誤差(偶然変動)	系統誤差(bias)



真度（正確性）accuracy
バイアス（系統誤差）の影響を受ける

① **測定者バイアス observer bias**
- 重めに評価する傾向、誘導質問をする

② **測定手段バイアス instrument bias**
- ある機器や質問票で常に高めに記録がなされる

③ **対象者バイアス subject bias**
- recall bias

うつ病-症例Sの日米評価者のMADRSの総得点分布
日本語MADRS面接ビデオを評価し、同じ面接ビデオに英語字幕をつけ、その評価をアメリカで実施

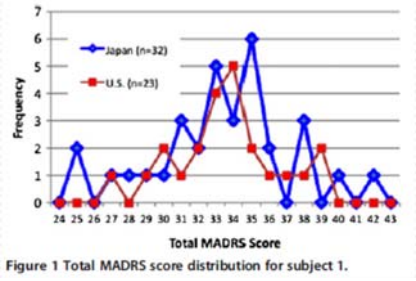
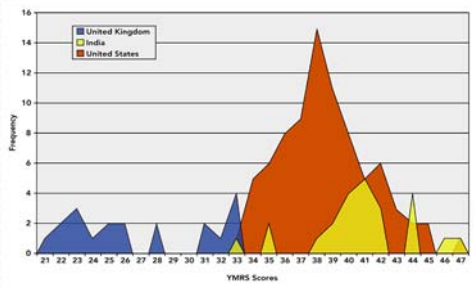


Figure 1 Total MADRS score distribution for subject 1.

Japan=33.3 (SD3.9), US=33.6(SD 3.0) p=ns

Targum SD, Nakagawa A, Sato Y.A cross-cultural comparison study of depression assessments conducted in Japan. Ann Gen Psychiatry. 2013 Apr 3;12(1):9.

躁病症例Bに関するアメリカ、イギリス、インド評価者の Young Mania Rating Scale総得点分布



The most profound differences were noted for **mood elevation** (item 1; *P*<0.001), **irritability** (item 5; *P*<0.001), **thought content** (item 8; *P*<0.001), and **disruptive–aggressive behaviour** (item 9; *P*<0.001).

cultural biases

Mackin P, Targum SD, Kalali A, Rom D, Young AH. Culture and assessment of manic symptoms. Br J Psychiatry. 2006 Oct;189:379-80.

真度（正確性）accuracyを高める対策

① **測定方法の標準化**
- 実施マニュアル（手順書）の作成

② **測定者のトレーニングと技能チェックcertification**

③ 測定手段の改善
- 機器の調整・キャリブレーション
- 質問票の言い回し(clarity, simplicity, neutrality, time frame)

④ 測定手段の自動化・自記化
- 自動血圧計

Feasibilityとのバランス

⑤ 気付かれない方法で測定を行う

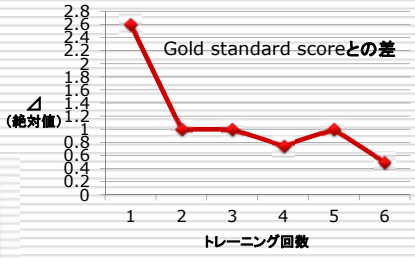
⑥ **盲検化blinding**

GRID-HAMD mock interview

NCNP認知行動療法センター GRID-HAMDビデオトレーニングと技能チェック

HAMDビデオを複数の評価者が視聴し、各自評価を行い比較した

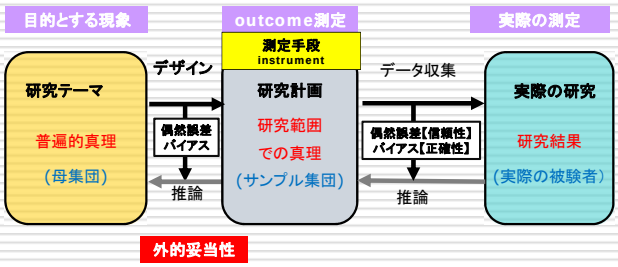
評価者間信頼性
ICC:0.94-0.98



Interpretation variance(観察した症状に対する解釈)とInformation variance (注目し観察している症状)のバラツキが小さくなる

測定measurement

目的とする現象を数値化するプロセス



妥当性 validity

① 内容妥当性 content validity
- 測定しようとする現象の全体を十分適切に捉えているか。
- **表現妥当性face validity**:適切さを主観的に判断される場合(専門家から見て無理なく妥当であるという判断)。

② 構成概念妥当性construct validity
- 測定しようとする現象の理論的概念を正確に表現しているか。
- たとえば、診断基準であれば、その疾患の生物学的・心理社会的要因や病理と一致するかを判断

③ 基準関連妥当性criterion-related validity
- 確実性の高い(gold standard)測定結果との相関
- 予測妥当性predictive validity:転帰や治療反応性が予測できるもの

"寛解=HAMD≦7" (Frank et al.1991) は妥当か？

医師の主観的評価 CGI	Furukawa 2007	Riedel 2010	Romera 2011	Zimmerman 2005,2012
Normal/not at all ill (remission)	**0-3**	**0-6**	**0-5**	**0-2**
Borderline ill	4-7			
Mildly ill	8-15			
Moderately ill	16-26			
Markedly or severely ill	≥27			

- Furukawa TA, Akechi T, Azuma H, Okuyama T & Higuchi T (2007) Evidence-based guidelines for interpretation of the Hamilton Rating Scale for Depression. Journal of Clinical Psychopharmacology, 27, 531-534.
- Riedel M, Möller HJ, Obermeier M, Schennach-Wolff R, Bauer M, Adli M, Kronmüller K, Nickel T, Brieger P, Laux G, Bender W, Heuser I, Zeiler J, Gaebel W, Seemüller F. Response and remission criteria in major depression--a validation of current practice. J Psychiatr Res. 2010 Nov;44(15):1063-8.
- Romera I, Pérez V, Menchón JM, Polavieja P, Gilaberte I. Optimal cutoff point of the Hamilton Rating Scale for Depression according to normal levels of social and occupational functioning. Psychiatry Res. 2011 Mar 30;186(1):133-7.
- Zimmerman M, Posternak MA, Chelminski I. Is the cutoff to define remission on the Hamilton Rating Scale for Depression too high? J Nerv Ment Dis. 2005 Mar;193(3):170-5.
- Zimmerman M, Martinez J, Attiullah N, Friedman M, Toba C, Boerescu DA, Rahgeb M. Further evidence that the cutoff to define remission on the 17-item Hamilton Depression Rating Scale should be lowered. Depress Anxiety. 2012 Feb;29(2):159-65.

まとめ:信頼性と妥当性の向上

- ☐ **測定値は正確で、再現性があるか**
- ☐ **測定が目的とする真の現象と近く、十分妥当か**

✓ **測定方法の標準化**
✓ **測定者のトレーニングと技能チェック**