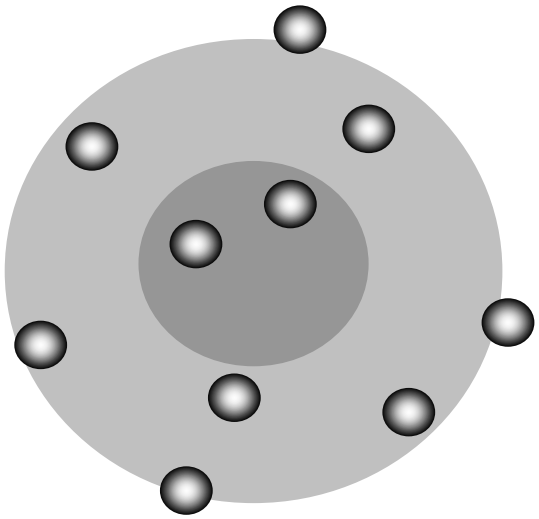
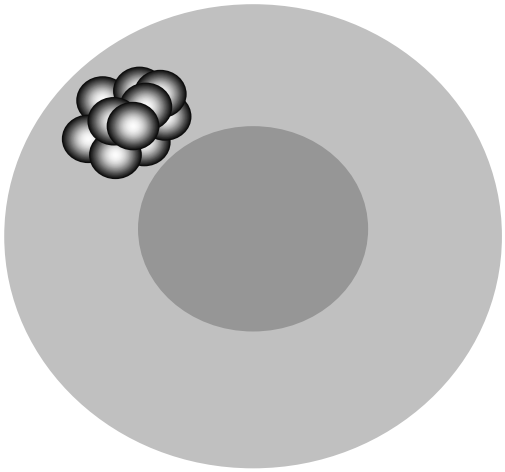
Stratification (open cohort study)

Accurate or Valid

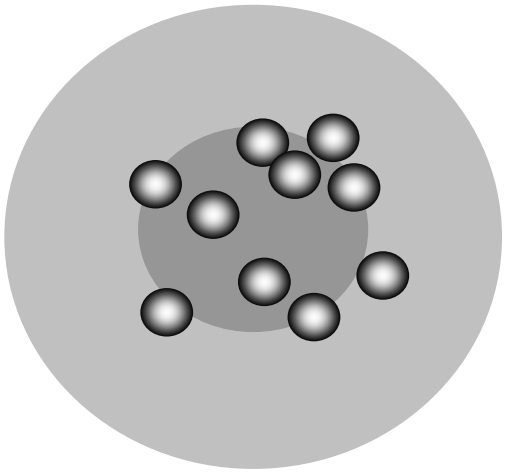
一般的にバイアス(bias)と効率(efficiency)の間には trade-off の関係があります。仮に多くのリスクファクターを心配するあまり多くのマスを作ってしまうと、１つのマスの中に存在する標本数が減少してしまうため、統計学的有意差を出しにくくなってしまいます。Confounder による bias は減るのですが、効率は悪くなります。ここでいう bias とは真の値と得られた値の差のことです。真の値は effect modification, confounders, selection bias, information bias が存在しない状態ではじめて得られます。逆にいうと bias のない状態を valid と呼びます。これは efficiency (=accuracy, precise) とは性格を異にします。もし細分化したマスを統合すれば 1 マス当りの標本数が増えるので efficiency は改善されますが、confounder の評価が下がり validity も下がります。



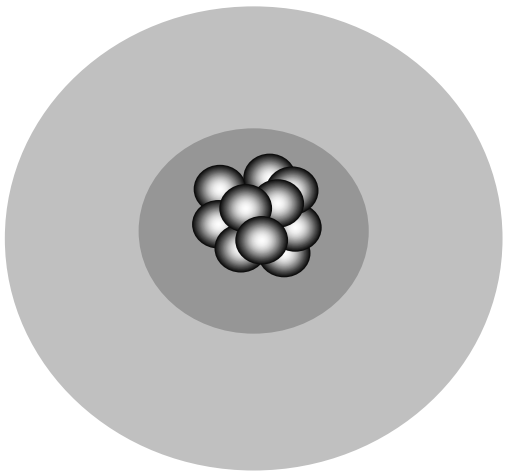
Not accurate not valid



Accurate but not valid



Valid but not accurate



Valid and Accurate

Open Cohort

通常の open cohort study は下の表であらわされます。

	exposed	unexposed	total
cases	a	b	M1
person-time	N_1	N_0	T

性別や年齢など confounder variable によって層化すると(stratify)、いくつかの似たような表ができあがります。i = 1, , , , , I 番目の strata としますと stratum i は以下のように表されます。

	exposed	unexposed	total
cases	a_i	b_i	M_{1i}
person-time	N_{1i}	N_{0i}	T_i

１つの臨床研究で confounder はしばしば複数考えられます。例えば喫煙の冠動脈疾患による死亡リスクについて調べた場合、結果は当然年齢,性別、コレステロール値に影響されますし、アルコールの量にも影響されるかもしれません。例えば British Doctors Study revised につても表を下に示します。

年齢	35-44		45-54		55-64		65-74		75-84	
喫煙	s	ns	s	ns	s	ns	s	ns	s	ns
患者	32	2	104	12	206	28	186	28	102	31
PY	52,407	18,790	43,248	10,673	28,612	5,710	12,663	2,585	5,317	1,462

S: smoker, ns: non-smoker、PY: person-years

単純に上の表を合計すると、

	s	ns
患者	630	101
PY	142,247	39,220
IR	44.29	25.79
IRR	1.72	

全体でみると喫煙者は非喫煙者に比べ 1.72 倍の頻度で冠動脈疾患によって死亡しています。でも思った程ではありません。

もう一度最初の表に戻って、各年齢群での喫煙によるリスクを検討してみましょう。

年齢	35-44		45-54		55-64		65-74		75-84	
喫煙	s	ns	s	ns	s	ns	s	ns	s	ns
患者	32	2	104	12	206	28	186	28	102	31
PY	52,407	18,790	43,248	10,673	28,612	5,710	12,663	2,585	5,317	1,462
IR	6.11	1.06	24.1	11.2	72.0	49.0	146.9	108.3	191.8	212.0
IRR	5.76		2.15		1.47		1.36		0.90	

IR: incidence rate, IRR: incidence rate ratio

各年齢層の IRR に注目すると、年齢が若ければ若い程喫煙による冠動脈疾患死亡率が高くなっています。とてもはっきりしたトレンドとしてつかむことができます。よって年齢は冠動脈疾患による死亡の effect modifier といえます。

次に年齢は confounder となっているのでしょうか？もしそうだとすると、confounder を打ち消して正しい値を得るにはどうしたらよいでしょうか？

$chi^2 = (O - E)^2 / var$

O: observed, E: expected, var:variance

が基本です。

Open cohort study において confounding が存在しないと仮定します。

$chi^2 = [X - E(X/H_0)]^2 / var(X/H_0)$

X = 暴露された中での患者数 = a
$E(X/H_0)$ = confounder が存在しない状況下で暴露された中から発生することが期待される患者数 = $M_1(N_1/T)$
$Var(X/H_0) = M_1(N_1/T)(1- N_1/T) = M_1N_1N_0/T^2$

全体を統合した表(crude table)から計算しますと、

$= [630 - (142247\ 731/181467)]^2 / (731 \times 142247 \times 39220/181467^2) = 26.24$

となります。

さて confounder を打ち消すために層化(stratification)しましたが、それぞれの表を合算しなくてはなりません。基本公式は crude table のものと類似しています。

$chi^2 = [\Sigma X_i - \Sigma E_i(X_i/H_{0i})]^2 / \Sigma var_i(X_i/H_{0i})$
Σは最初から I 個の表を全部合算するという意味です。

X_i = 暴露された中での患者数 = a_i
$E_i(X_i/H_{i0})$ = confounder が存在しない状況下で暴露された中から発生することが期待される患者数 = $M_{1i}(N_{1i}/T_i)$

$Var_i(Xi/H_{0i}) = M_{1i}(N_{1i}/Ti)(1- N_{1i}/T_i)=M_{1i}N_{1i}N_{0i}/T_i^2$
自由度は１です。
H_0: 喫煙と冠動脈疾患死亡率に関連がない。
H_A: 喫煙と冠動脈疾患死亡率に関連がある。

X_i 630
$E_i(X_i/H_{i0})$ と$Var_i(Xi/H_{0i})$ を表から計算します。

年齢	35-44		45-54		55-64		65-74		75-84	
喫煙	s	ns	s	ns	s	ns	s	ns	s	ns
患者	32	2	104	12	206	28	186	28	102	31
PY	52,407	18,790	43,248	10,673	28,612	5,710	12,663	2,585	5,317	1,462
IR	6.11	1.06	24.1	11.2	72.0	49.0	146.9	108.3	191.8	212.0
IRR	5.76		2.15		1.47		1.36		0.90	
$E_i(X_i/H_{i0})$	25.03		93.04		195.07		177.72		104.32	
$Var_i(Xi/H_{0i})$	6.61		18.41		32.45		30.06		22.50	

$\Sigma E_i(X_i/H_{i0}) = 595.2$
$\Sigma Var_i(Xi/H_{0i}) = 110.1$

よって
$chi^2= [\Sigma X_i - \Sigma E_i(X_i/H_{0i})]^2 / \Sigma var_i(X_i/H_{0i})=(630 - 595.2)^2 / 110.1 = 11.02$

$Pr(chi^2 > 11.02) = 0.001$

H_0は棄却され、喫煙と冠動脈疾患死亡率の間に関連があると結論できます。もちろん年齢についてはcontrol していますが(より正確にcontrol するには 10 年でなく、1 年単位で区切る方が正確になりますがブランクのマスができたりして効率的ではありません。よって適当に区切るとconfounder を完全には除去しきれないことになります:residual confounder)、さらに年齢層を 10 年でなく更にこまかくすることもできるのですから、confounder が残存していないとも限りません(residual confounder)。また他のconfounder (飲酒、性別など)、selection bias, information bias が存在しないことが前提となります。

年齢で層化していない場合のχ^2は 26.23 であったのに対して、年齢(confounder)で層化した後のχ^2は 11.02 と大分小さくなっています。しかし年齢というconfounder の存在を度外視しても、喫煙は冠動脈疾患の発生を有意に増加させるといえます。そして年齢はconfounder としてcrude data を過大評価させていたといえます。ですからadjust すると値は小さくなります。

さてそれでは何倍に増加させるのでしょうか？
論文にする際、それぞれの表として報告することも重要です。Effect modification 自体重要な所見ですので、多くはそのように報告されます。また明らかに差のある若年層だけとる研究者もいるかもしれません。しかしその場合多くのデータをゴミ箱に捨てることになり賢明とはいえません。やはり confounder（年齢）を除外した際の喫煙が冠

動脈疾患死亡率を全体で何倍に押し上げるのかを知りたいところです。

これを計算するためにはそれぞれの年齢層の表を合算しなくてはなりません。単純に合算してもよいものでしょうか？例えば高齢者のPYは小さくなっていますが、均等に合算すると高齢者のIRRの比率(weight)が過大評価されてしまいます。それではPYの比率(weight)に従って分配したらよいでしょうか。そうすると若い世代は心筋梗塞患者数が少ないのにこれを反映しないことになってしまいます。Inverse variance weights {$w_i = 1/\mathrm{var}[\ln(RR_i)] = 1/(1/a_i + 1/b_i)$}でもよいのですが、a or b が0だとweight も0になってしまうためそのstrata のデータは捨てなくてはなりません。よってMantel-Haenszel weights がそれぞれの表を合算する際、どの程度の重みをもって評価するかに用いられます。

$w_i = b_i N_{1i}/T_i$

しかしながら effect modification が存在しなければどの weight を選択してもかまいません。

$$RR_{MH} = \Sigma w_i \, RR_i \,/\, \Sigma w_i = \Sigma b_i N_{1i}/T_i (a_i/N_{1i})/(b_i/N_0) \,/(\Sigma b_i N_{1i}/T_i)$$
$$= \underline{(\Sigma a_i N_{1i}/T_i)/(\Sigma b_i N_{1i}/T_i)}$$

年齢	35-44		45-54		55-64		65-74		75-84	
喫煙	s	ns	s	ns	s	ns	s	ns	s	ns
患者	32	2	104	12	206	28	186	28	102	31
PY	52,407	18,790	43,248	10,673	28,612	5,710	12,663	2,585	5,317	1,462
IR	6.11	1.06	24.1	11.2	72.0	49.0	146.9	108.3	191.8	212.0
IRR	5.76		2.15		1.47		1.36		0.90	
lnIRR	0.531		0.093		0.041		0.041		0.042	
$a_i N_{1i}/T_i$	8.445		20.59		34.27		31.53		22.00	
$b_i N_{1i}/T_i$	1.47		9.62		23.34		23.25		24.31	

$\Sigma a_i N_{1i}/T_i = 116.835$
$\Sigma b_i N_{1i}/T_i = 82.09$
$RR_{MH} = (\Sigma a_i N_{1i}/T_i)/(\Sigma b_i N_{1i}/T_i) = 1.42$

年齢を調整した結果、IRR は 1.72 より 1.42 に変化しました。他の confounding, bias が無いものとすると、喫煙により冠動脈疾患による死亡率が 42%増加することになります。

先にp value を出していますが、95% confidence interval (CI) を出してみましょう。

$\ln IRR_{MH} \pm 1.96\sqrt{\mathrm{var}(\ln IRR_{MH})}$

$\mathrm{var}(\ln IRR_{MH}) = A/BC$

$A = \Sigma M_{1i}N_{1i}N_{0I} / T_i^2$
$B = \Sigma a_iN_{0i} / T_i$
$C = \Sigma b_iN_{1i}/T_i$

$var(\ln IRR_{MH}) = A/BC = 0.01149$

$\ln(1.42) \pm 1.96\sqrt{0.01149} = (0.141, 0.560)$
$e^{(0.141, 0.560)} = (1.15, 1.76)$

結論
もうこれ以上 confounding、あるいは bias が無いとして、confounder や bias は known/unknown を含め無限大にあり、adjust したからといって confounder が完全に除去されるわけではありません。ですから統計学的結論を述べるときには confounder や bias が存在しないとして云々ということになります。喫煙によって 15%から 75%冠動脈疾患による死亡率の有意な増加をみると 95%の信頼をもっていうことができます。最初の IRR は 1.72 倍であり、年齢という positive confounding により過大評価されていたことになります。

次に rate difference (RD)について検討してみましょう。

$RD = \Sigma w_i' RD_i / \Sigma w_i'$
$Var(RD) = a/N_1^2 + b/N_0^2 = (aN_0^2 + bN_1^2)/N_1^2N_0^2$
$w'_i = 1/vari(RD_i) = N_{1i}^2N_{0i}^2 /(N_{0i}^2a_i + N_{1i}^2b_1)$

年齢	35-44		45-54		55-64		65-74		75-84	
喫煙	s	ns	s	ns	s	ns	s	ns	s	ns
患者	32	2	104	12	206	28	186	28	102	31
PY	52,407	18,790	43,248	10,673	28,612	5,710	12,663	2,585	5,317	1,462
IR	6.11	1.06	24.1	11.2	72.0	49.0	146.9	108.3	191.8	212.0
IRD	0.000504		0.001280		0.002296		0.003856		-0.002020	
Var(RD)x10^7	0.173		1.609		11.104		53.502		181.113	

$RD = \Sigma w_i' RD_i / \Sigma w_i' = 0.00061$

Confounder, bias がないものと仮定して、10,000PY あたり 6.1 人の冠動脈疾患死亡は喫煙によると結論できます。

$RD \pm 1.96\sqrt{var(RD)}$
$Var(RD) = 1 / \Sigma w_i$

$0.00061 \pm 1.96\sqrt{0.0000000154} = (3.7/10^4\ PY, 8.5/10^4\ PY)$

$RD_{crude} = 1.85/1{,}000PY\ (12.4/10^4\ PY, 24.6/10^4\ PY)$

RD_{crude}は$RD_{adjusted}$に比較して随分大きいことが判ります。このことは年齢がRDに関し

てconfounder として強く働いたことを示しています。

コンピュータプログラム

それでは上で用いた喫煙と冠動脈疾患に関する British Doctors Study について STATA6 を用いて解析してみましょう。

まずはコマンドにデータをこれから入れることを指示します。time は person-time の意味で open cohort 研究であることを知らせています。con は condition のことであり、ここでは 35―44 歳の strata を１、45―54 歳を２、55―64 歳を３、65―74 歳を４、75―84 歳を５とします。exposed は禁煙をしていたかしていなかったかで、前者を１、後者を０で表しています。cases は冠動脈疾患の数です。単純に 32 スペース 1 スペース 1 スペース 5247 リターンと次々に入力します。そして最後の end を入力し、データ入力が終了したことをコンピュータに知らせます。

```
. input cases exposed con time

        cases    exposed        con       time
  1. 32 1 1 52407
  2. 2 0 1 18790
  3. 104 1 2 43248
  4. 12 0 2 10673
  5. 206 1 3 28612
  6. 28 0 3 5710
  7. 186 1 4 12663
  8. 28 0 4 2585
  9. 102 1 5 5317
 10. 31 0 5 1462
 11. end
```

次に incidence rate ratio, incidence rate difference とそれぞれの 95%CI を計算するようコンピュータに指示します。コマンドに ir cases exposed time と入力するだけで下記のような表が現れます。

```
. ir cases exposed time

              | exposed             |
              |   Exposed   Unexposed  |    Total
--------------+------------------------+----------
         cases |      630        101  |      731
          time |   142247      39220  |   181467
--------------+------------------------+----------
              |                        |
  Incidence Rate |  .0044289    .0025752  |  .0040283
              |                        |
              |     Point estimate    |  [95% Conf. Interval]
              |------------------------+----------------------
 Inc. rate diff. |        .0018537      |  .0012439    .0024635
 Inc. rate ratio |       1.719823       |   1.39197    2.143531  (exact)
 Attr. frac. ex. |        .4185447      |  .2815935     .53348  (exact)
 Attr. frac. pop |        .3607157      |
              +-------------------------------------------------
                (midp)  Pr(k>=630) =                0.0000  (exact)
                (midp) 2*Pr(k>=630) =                0.0000  (exact)
```

書いてある通りで、喫煙者 142,247 person years に対して 630 人の冠動脈疾患の発生をみています。一方非喫煙者の場合 39,220 person years に対しては 101 人の発生しかみていません。

その下は incidence rate で 1,000person-years 当り 44 人の冠動脈疾患の発症をみています。

Incidence rate difference は喫煙者、非喫煙者間がどれくらい違うかで、incidence rate ratio は何倍違うかです。Attributable risk および attributable risk population も算出されます。意味は前章を参照してください。さらに左にはそれぞれの 95%信頼区間が示してあります。

上の表は全体をまとめた crude data でした。以下各年齢層に分けた表について検討してみましょう。

```
. ir cases exposed time, by(con)

           con |     IRR     [95% Conf. Interval]   M-H Weight
---------------+--------------------------------------------------
             1 |  5.736638    1.463519   49.39901     1.472169  (exact)
             2 |  2.138812    1.173666   4.272307     9.624747  (exact)
             3 |   1.46824    .9863626   2.264174     23.34176  (exact)
             4 |   1.35606    .9082155    2.09649     23.25315  (exact)
             5 |  .9047304    .6000946   1.399699     24.31435  (exact)
---------------+--------------------------------------------------
         Crude |  1.719823     1.39197   2.143531             (exact)
   M-H combined |  1.424682    1.154703   1.757784
---------------+--------------------------------------------------
 Test of homogeneity (M-H)   chi2(4) =    10.41  Pr>chi2 = 0.0340
```

若い世代の方が喫煙者により多くの冠動脈疾患の発生をみていることがわかります。さらに 95%CIも示されていますが、１（35―44 歳）のように 95%CIの幅が広い場合には統計学的に信憑性が低いことを示しています。若い世代では冠動脈疾患発生の頻度が少ない為このような結果となってしまったのです。また Mantel-Haenszel による weight も示されています。つまり３，４の世代により重点を置いて評価することを示しています。そしてCrude data (RR_{crude})とweight を用いて計算しなおしたIRR_{MH}を比較することができます。実際の計算と一致しました。

この調査では British Doctor を対象に行なわれたものですが、喫煙者は 55―64 をピークにしているのに対して、非喫煙者は 35―44 歳をピークとしています。冠動脈疾患は高齢者において発生頻度が増加しますから、年齢は confounder となりえます。Confounder を調整(adjust)するために direct standardization （後述）を用いて計算してみましょう。ここでは非喫煙者を標準としています。よって下の表で weight の数値は非喫煙者の PY となります。すなわち若い人により大きいウエイトを置く形となります。

```
. ir cases exposed time, by(con) es

           con |     IRR     [95% Conf. Interval]       Weight
---------------+-------------------------------------------------
             1 |   5.736638    1.463519   49.39901        18790  (exact)
             2 |   2.138812    1.173666   4.272307        10673  (exact)
             3 |    1.46824    .9863626   2.264174         5710  (exact)
             4 |    1.35606    .9082155    2.09649         2585  (exact)
             5 |   .9047304    .6000946   1.399699         1462  (exact)
---------------+-------------------------------------------------
         Crude |   1.719823     1.39197   2.143531             (exact)
 E. Standardized |   1.428377    1.157184   1.763125
```

年齢で標準化(standardized)した値は元の値(crude)と比較して小さくなっています。この値は Mantel-Haenszel 法で得た値とも一致します。何故年齢を調節すると値が小さくなったのでしょうか？年齢が上がると喫煙者の割合が増えて、しかも年齢があがると冠動脈疾患の頻度も増えるため、年齢を調節しないと結果を過大評価してしまうからです。

Incidence rate ratio についてだけではなく、incidence rate difference についても算出してみましょう。

```
. ir cases expose time, by(con) es ird

           con |     IRD     [95% Conf. Interval]       Weight
---------------+-------------------------------------------------
             1 |   .0005042    .0002463   .0007621        18790
             2 |   .0012804    .0004941   .0020667        10673
             3 |   .0022961    .0002308   .0043614         5710
             4 |   .0038567   -.0006767   .0083902         2585
             5 |  -.0020201   -.0103612    .006321         1462
---------------+-------------------------------------------------
         Crude |   .0018537    .0012439   .0024635
 E. Standardized |   .0011032    .0007922   .0014141
```

前と同様の理由で標準化した incidence rate difference も小さくなっています。

今度は間接法で incidence rate ratio を調整します。

```
. ir cases exposed time, by(con) is

           con |      IRR     [95% Conf. Interval]       Weight
---------------+-------------------------------------------------
             1 |   5.736638    1.463519   49.39901        52407  (exact)
             2 |   2.138812    1.173666   4.272307        43248  (exact)
             3 |    1.46824    .9863626   2.264174        28612  (exact)
             4 |    1.35606    .9082155    2.09649        12663  (exact)
             5 |   .9047304    .6000946   1.399699         5317  (exact)
---------------+-------------------------------------------------
         Crude |   1.719823     1.39197   2.143531             (exact)
 I. Standardized |   1.417609    1.146777   1.752403
```

間接法により incidence rate difference を計算します。

```
. ir cases exposed time, by(con) is ird

           con |      IRD     [95% Conf. Interval]       Weight
---------------+-------------------------------------------------
             1 |   .0005042    .0002463   .0007621        52407
             2 |   .0012804    .0004941   .0020667        43248
             3 |   .0022961    .0002308   .0043614        28612
             4 |   .0038567   -.0006767   .0083902        12663
             5 |  -.0020201   -.0103612    .006321         5317
---------------+-------------------------------------------------
         Crude |   .0018537    .0012439   .0024635
 I. Standardized |   .0013047    .0009929   .0016165
```

以上の結果より confounder および bias が存在しないと仮定して、「年齢により adjust したところ、喫煙により CHD のリスクは 42%上がる」と結論できます。