30.2 處理偏誤的方法

總之，後設分析中的研究可能高估真正的效果值，因為這些研究是目標研究母群體的有偏樣本。但是，如何處理這個問題呢？唯一能真正檢驗出版偏誤的是比較正式發表研究的效應和未發表研究的效應。前提是必須有未發表的研究，如果可以得到，就沒有什麼擔心的了。但是，最好的方法是去進行一個全面的文獻檢索，以期減小偏誤。事實上，沒有證據表明這種方法一定程度是有效的。考科藍文獻回顧包含更多的研究，相較於發表在醫學期刊上的類似系統敘述性綜論呈現了更小的效果值。努力去查找未發表的及難以得到的文獻，因此，經典的考科藍文獻回顧可能降低出版偏誤的影響。

儘管從學位論文、專題文章、會議論文、政府報告、技術報告等類似的資源檢索資料將增加後設分析的資料來源，無條件地排除這些研究報告的綜合分析通常是不能被接受的，必須平衡得到灰色文獻潛在的效益。關於文獻檢索，想要得到更多指南的讀者可以參看相關文獻（Hopewell et al., 2005; Reed & Baxter, 2009; Rothstein & Hopewell, 2009; Wade et al., 2006）。

因為無法確定是否避免了偏誤，研究者發展了幾個針對評價偏誤對任何特定後設分析的潛在影響的方法。這些方法主要說明以下問題：

- 有偏誤存在的證據嗎？
- 全部效應都可能是偏誤的結果嗎？
- 偏誤的影響可能有多大？

下一節以被動吸菸與肺癌關係的後設分析來說明這些方法。

30.3 引例

Hackshaw 等人（1997）發表了所謂吸二手（被動）（second-hand, passive）菸與肺癌關係的後設分析，包含 37 個研究，文章並包含暴露於二手菸使不吸菸的配偶得到肺癌的風險增加了約 20%。問題是有著較大效應的研究更可能被發表，以至於更可能被納入分析，因此結論是值得懷疑的。

30.4 基本假設

為了衡量出版偏誤的影響，需要一個模型來說明哪些研究可能被漏掉了。常用的模型（包括我們用的模型）假設：1. 大型研究更可能被發表，不管是否

有統計顯著性，因為這些研究被投入大量的時間和資源；2. 中度大小的研究樣本有被漏掉的風險，但是中等的樣本大小下，即使中度效應也可能得到有統計顯著性的結果，所以也可能只有一部分研究被漏掉；3. 小樣本研究最有可能被漏掉，因為小的樣本大小，只有最大的效應才有可能得到有統計顯著性的結果，因此小到中等的效應可能不被發表。

綜合以上三點，我們假設偏誤隨樣本大小增加而降低。下面描述的方法都將基於這樣的假設。此外，還有更為複雜的方法用於估計漏掉的研究數和（或）考慮偏誤調整效應。但是因為操作上的困難，很少用於實際的研究，另一原因也因為使用者需要做出相對複雜的假設和選擇。

在繼續討論實例之前，需要提出一個重要的附加說明提請讀者注意。這裡描述的方法是尋找樣本大小和效果值的關係，如果有這樣的關係，說明有漏掉的研究。這是一個可能的原因，即小樣本研究中效果值越大，但是，也有一種可能是在小樣本研究中效果值確實越大。這裡提請讀者注意是為了說明隨後的討論背景，並將會在本章最後「小研究效應」一節進行討論。

30.5 圖示數據

評價偏誤的潛在影響，最好要先理解資料，森林圖可用於這個目的。圖 30.1 為被動吸菸後設分析的森林圖。本例中，相對風險性大於 1 說明風險增加。絕大部分研究顯示吸二手菸增加了肺癌風險，最後一行為固定效果模型綜合效應估計結果。相對風險估計為 1.204，95% 信賴區間為 1.120 ～ 1.295。

圖中從最大的精確度到最小的精確度來標注研究，所以樣本大的研究出現在前，樣本小的研究出現在後。這對於合併效果值沒有影響，但是可以藉此看出樣本大小和效果值之間的關係。從上到下，效應向右移動，正是上述模型預測的當偏誤存在時的情形。如果包含來自公開發行雜誌的研究和灰色文獻的研究，可以依來源分組或檢視灰色文獻（可能代表了任何漏掉的研究）是否傾向於呈現更小的效應。

漏斗圖

展示樣本大小和效果值關係的方法是漏斗圖（funnel plot）。

傳統的漏斗圖以效果值為 X 軸，樣本大小或變異數為 Y 軸。大樣本研究出現在圖的頂部，聚集在平均效果值周圍；小樣本研究出現在圖的底部，因為小

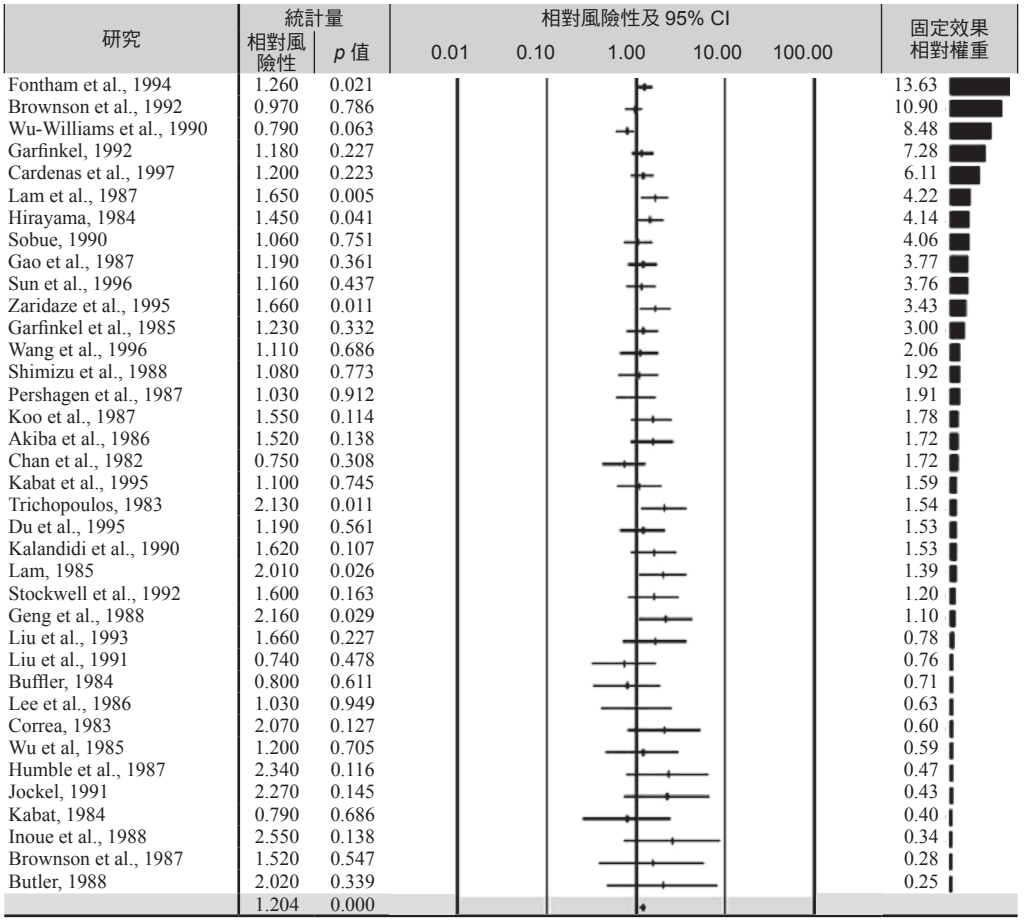

Figure 圖 30.1　吸二手菸與肺癌——森林圖

樣本研究有更大的抽樣誤差，分散在更寬的數值範圍。這種圖案像一個漏斗，因此被稱為漏斗圖（Light & Pillemer, 1984; Light et al., 1994）。

以標準誤而不是樣本大小或變異數作為Y軸的好處是所有點分布在圖的「下半部」，即小樣本研究分布在這裡。這個原則可使得判斷對稱性更加容易，且只影響到圖形展示，不影響統計量。

30.6 有偏誤存在的證據嗎？

如果沒有出版偏誤，研究將會對稱分布在平均效應的周圍，因為抽樣誤差是隨機的。存在出版偏誤時，研究如上述模型假定所述，在頂部對稱，較小樣

本研究缺失，底部有更多的研究缺失。如果效應的方向向右（如本例），接近圖的底部左側將會有一個缺口，如果能夠得到，將會是無統計顯著性的研究所在位置。

本例中（圖 30.2），X 軸為相對風險（對數單位），Y 軸為標準誤（因為是反方向的，故低值在頂部）。直觀上看，圖是不對稱的。往下看，大多研究出現在右側（表示更大的風險），和左側有些研究可能缺失是一致的。

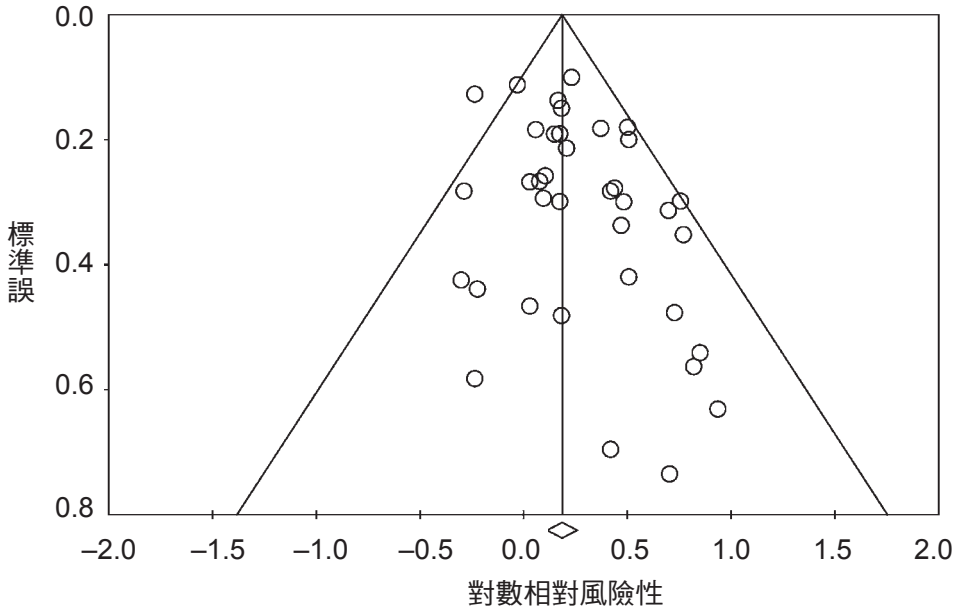

Figure 圖 30.2　吸二手菸與肺癌——漏斗圖

首先，一個問題是，是否有偏誤存在的證據。因為漏斗圖的解釋很大程度上是主觀的，因此，有幾種量化或檢驗樣本大小和效果值關係的檢驗方法被提出。兩個早期的方法被廣泛使用（Begg & Mazumdar, 1994; Egger et al., 1997）；相關的綜述參見 Higgins 與 Green （2008）著作的第 10 章。

儘管這些檢驗提供了有用的資訊，但值得注意的是：首先，正如漏斗圖，這些檢驗所用指標不同（風險差或風險比），可能產生非常不同的圖；第二，只有在樣本大小有著合理的離散度，研究數大小合理時才是有意義的；第三，即使這些條件滿足，研究的檢定力也比較低。這樣沒有統計顯著性的相關或迴歸也不能作為對稱性的證據。任何情況下，即使我們可以某種方法解決這些問題，研究所能回答的問題（即是否有偏誤存在的證據）的重要性也是有限的。

更有趣的問題是**偏誤有多大？對結論的影響是什麼？**

30.1 整個效應僅是偏誤的假象？

下面是關於觀察到的綜合效應是否穩健的問題，換句話說，我們是否有信心說效應不僅是偏誤造成的假象？

Rosenthal 的失敗安全數

一個早期的處理出版偏誤的方法是 Rosenthal 的失敗安全數（Fail-Safe *N*）。假定後設分析基於 *k* 個研究得到了有統計顯著性的 *p* 值。我們關心的是有著較小效果值的研究是否被漏掉了，如果檢索到漏掉的研究，並包含在分析中，合併效果值的 *p* 值就不再有統計顯著性了。Rosenthal （1979）提出計算我們需要加入多少個漏掉的研究使得分析結果變為沒有統計顯著性。為了練習我們假設漏掉的研究中平均效應為 0。如果我們僅需要很少的研究（如 5 或 10 個）就可以使效應逆轉，我們可以懷疑真正的效應就是 0，如果我們需要非常多的研究（如 20,000），那麼就沒有理由懷疑真正的效應為 0。

Rosenthal 稱這種方法為「文獻抽屜」（File drawer）分析（文獻抽屜是指假定的缺失研究的位置），Harris Cooper 則建議稱需要改變檢驗結果的缺失研究數為失敗安全數（Rosenthal, 1979; Begg & Mazumdar, 1994）。

儘管 Rosenthal 的作法對於處理出版偏誤問題是重要的，但為有條件的使用，主要有幾個原因：首先，該法是針對統計顯著性而不是實際意義。也就是說，它是問多少隱藏的研究可以使得效應沒有統計顯著性，而不是使得效應沒有實際意義；第二，公式假定漏掉研究的平均效應為 0，而事實上，可能是負值（只要求更少的研究就可以逆轉效應）或者比較小的正值；最後，失敗安全數是基於綜合研究的 *p* 值的顯著性檢定，這種綜合方法在 Rosenthal 提出失敗安全數法的當時是很普遍的。現在普遍的作法是計算合併效果值，然後計算這個效應的 *p* 值。而不同方法計算的 *p* 值實際上檢驗的是不同的虛無假設。也因為這些原因，這個方法並不適合針對效果值的分析方法。我們花相對長的篇幅來說明這一點是因為其重要的歷史地位。

也就是說，對於吸二手菸的系統評價失敗安全數為 398，說明我們需要 400 個平均相對風險性為 1 的研究加入分析中，才能使得累積效應變為無統計顯著性的結果。

Orwin 的失敗安全數

如前所述，Rosenthal 的失敗安全數的兩個問題分別為針對統計顯著性，而不是實際意義和假設漏掉研究的平均效應為 0。Orwin（1983）針對這兩個問題，提出了對 Rosenthal 方法的改進。首先，Orwin 的方法允許研究者決定多少漏掉的研究將會使得綜合效應達到一個特別的水準而不是 0。因此研究者可以選擇一個代表了有著實際重要性的最小的效應值，然後，問多少研究可以使得合併效果值低於這個值。第二，該法可以讓研究者確定缺失研究的平均效應為一特定值而非 0。因此，研究者可以模擬一系列其他分布的缺失研究 （Becker, 2005; Beggan & Mazumdar, 1994）。

上述吸二手菸的後設分析中，Orwin 失敗安全數為 103，說明需要 100 個平均相對風險性為 1 的研究加入分析才能使得綜合效應變得不重要（定義相對風險比為 1.05）。

30.8 偏誤的影響多大

上述方法是關於是否偏誤對觀察效應有影響（基於漏斗圖），或者是否完全決定了觀察效應（失敗安全數）。和這些極端的情形相比，第三種方法是努力估計偏誤的效應多大，估計當偏誤不存在時，效果值會是什麼。目的是將每個後設分析大概分為三種情形：

- 偏誤的影響可能是不重要的。如果包含所有相關的研究，效果值將保持不變。
- 偏誤的影響可能是中等程度的。如果所有相關的研究都被包含，效果值可能有所偏移，但是關鍵的結論（即效應是否有實際意義）可能保持不變。
- 偏誤的影響可能是重要的。如果所有相關的研究都被包含，關鍵的結論（即效應是否有實際意義）可能改變。

Duval 和 Tweedie 的修剪填補法

如上所述，漏斗圖的基本思想是出版偏誤將表現在資料點分布的不對稱。如果有更多的小樣本研究分布在右側，我們考慮在左側有些研究缺失（在引例中，我們考慮左側的研究缺失，但是在其他情形下可能考慮右側的研究缺失。該法要求研究者確定期望的方向）。

修剪填補法（Trim and Fill）採用迭代法（iterative procedure）從漏斗圖正的一邊移去最極端的研究，重新計算效果值，直到得到對稱的漏斗圖（關於新的綜合效果值）。理論上，這將會得到效果值的不偏估計量（unbiased estimator）。這種修剪產生了調整的效果值同時也降低了效應的變異數，得到一個過窄的信賴區間。因此，該法隨後將除去的原始研究再加入到分析中，為每個研究填補一個鏡像。這種填充對點估計沒有影響，但可用於校正變異數（Duval & Tweedie, 2000a, 2000b）。

該法的一個重要優點為它回答了一個重要問題：**什麼是最好效果值的不偏誤估計量**？該法的另一個優良特徵是有助於直觀的圖示。包含修剪填補法的電腦程式可以生成包含觀察到的研究和填補研究的漏斗圖，所以研究者可以看到當填補了相應的研究時效果值是如何變化的。如果這種變化是不重要的，研究者有更充分的信心認為報告效應是有效的。這個方法的問題是過分依賴關於為什麼研究有缺失的模型假設，發現不對稱的演算法可能被一兩個異常的研究所影響。

我們可能重新構造漏斗圖，考慮對修剪填補法做調整。圖 30.3 中，實際觀察到的研究用空心圈來表示，基於觀察點得到的對數單位下的效應估計用空心菱形表示為 0.185（0.113, 0.258），相應的相對風險性為 1.204 （1.120, 1.295）。7 個填補的研究用實心圓點來表示，加入填補點後估計的對數單位下的效應估計用實心菱形來表示，為 0.156（0.085, 0.227），相應的相對風險性為 1.169 （1.089, 1.254）。調整的點估計得到比原來的分析更小的風險。

儘管如此，關鍵是調整估計與原來的估計非常接近，這種情況下，1.17 的相對風險性和 1.20 的風險比實際意義是相同的。

僅分析大規模研究

如果出版偏誤主要作用於小樣本研究，那麼，因為一般來說無論結果如何大型研究都會被發表，那麼僅對大型研究進行分析，理論上可以克服任何問題。問題是如何定義何謂大型研究。我們不可能提供一個一般的指南，而一個可能有用的方法是透過累積後設分析（cumulative meta-analyses）說明所有可能的臨界值（threshold）。

累積後設分析是只從一個研究開始的後設分析，然後每加入一個研究，就重複進行一次後設分析，直到包含所有的後設分析。同樣的，累積森林圖中第一行展示了基於一個研究的效應估計，第二行為基於兩個研究得到的效應估計，

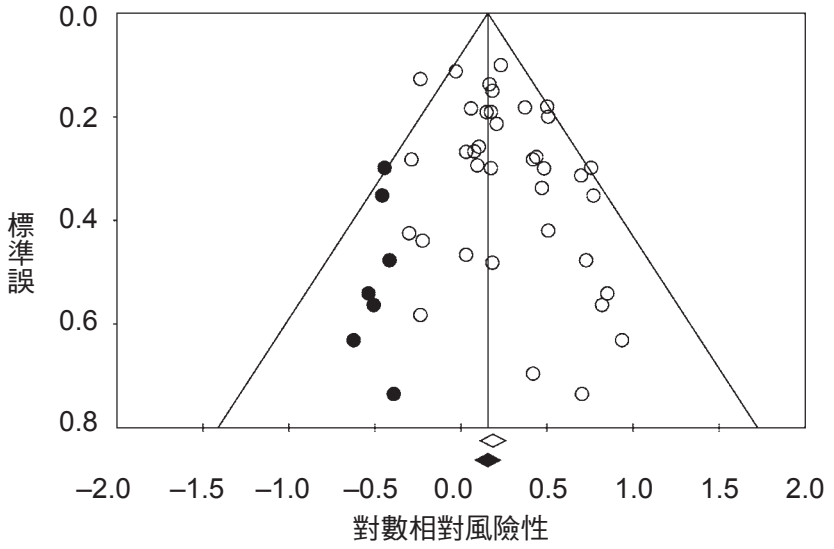

Figure 圖 30.3　吸二手菸與肺癌——包含填補研究的漏斗圖

以此類推。累積後設分析將會在第 42 章詳細討論。

為了檢驗大規模研究的不同定義界值，研究從大到小（或者從最大的精確度到最小的精確度）依次排序，進行依次加入每個研究的累積後設分析。如果在加入更大的研究時點估計是穩定的，加入小樣本研究時不改變，那麼就沒有理由認為小樣本研究引入了偏誤（因為小樣本研究，選擇偏誤可能是最大的）。另一方面，如果在加入小樣本研究時，點估計改變，那麼，至少表面上這是偏誤存在的確鑿證據，研究會希望找到改變的原因。

這個方法還是僅基於大的研究得到效果值的估計，且相較於修剪填補法，這種方法是完全透明的：我們基於大的研究計算效果值，然後加入小樣本研究，觀察效果值是否有變化或如何變化（大型研究和小型研究的明確區分通常是不存在的，且也不需要）。

圖 30.4 為資料的累積森林圖。注意，先前已經展示了累積森林圖和標準森林圖的不同。這裡，第一行為僅基於 Fontham 等人的研究之後設分析，第二行為基於兩個研究（Fontham 等和 Brownson 等）的後設分析，以此類推。最後一個加入的研究是 Butler（1988），所以，Butler 列的點估計和信賴區間估計（用線表示）與標「固定」（fixed）的合併效果值估計是相同的。而且這個圖上尺度從 0.5 ～ 2.0。

研究從最大的精確度到最小的精確度排序（大體上對應最大到最小的）。

研究	統計量 估計值	*p* 值	累積相對風險性和 95% CI 0.50 1.00 2.00	固定效果 相對權重
Fontham et al., 1994	1.260	0.021		13.63
Brownson et al., 1992	1.122	0.124		24.53
Wu-Williams et al., 1990	1.025	0.699		33.01
Garfinkel, 1992	1.052	0.388		40.29
Cardenas et al., 1997	1.070	0.213		46.40
Lam et al., 1987	1.109	0.046		50.62
Hirayama, 1984	1.132	0.013		54.75
Sobue, 1990	1.127	0.013		58.81
Gao et al., 1987	1.131	0.009		62.58
Sun et al., 1996	1.132	0.006		66.34
Zaridaze et al., 1995	1.154	0.001		69.76
Garfinkel et al., 1985	1.157	0.001		72.77
Wang et al., 1996	1.155	0.001		74.83
Shimizu et al., 1988	1.154	0.001		76.75
Pershagen et al., 1987	1.150	0.001		78.66
Koo et al., 1987	1.158	0.000		80.44
Akiba et al., 1986	1.165	0.000		82.15
Chan et al., 1982	0.154	0.000		83.87
Kabat et al., 1995	1.153	0.000		85.46
Trichopoulos, 1983	1.166	0.000		87.00
Du et al., 1995	1.166	0.000		88.53
Kalandidi et al., 1990	1.173	0.000		90.05
Lam, 1985	1.182	0.000		91.45
Stockwell et al., 1992	1.187	0.000		92.65
Geng et al., 1988	1.195	0.000		93.75
Liu et al., 1993	1.199	0.000		94.53
Liu et al., 1991	0.194	0.000		95.29
Buffler, 1984	1.190	0.000		96.00
Lee et al., 1986	1.189	0.000		96.63
Correa, 1983	1.193	0.000		97.23
Wu et al, 1985	1.193	0.000		97.82
Humble et al., 1987	1.197	0.000		98.29
Jockel, 1991	1.201	0.000		98.72
Kabat, 1984	1.199	0.000		99.12
Inoue et al., 1988	1.202	0.000		99.46
Brownson et al., 1987	1.202	0.000		99.75
Butler, 1988	1.204	0.000		100.00
	1.204	0.000		

Figure 圖 30.4　吸二手菸與肺癌——累積森林圖

從最上面開始，基於最大的 18 個研究（包含（Chan & Fung, 1982））估計的累計相關風險為 1.15。隨著另外 19 個（小樣本的）研究，點估計向右移動，相關風險為 1.2。本身來看相關風險增加了。但是，關鍵的問題是即使僅基於大的研究進行後設分析，相對風險性也就是 1.15（95% 信賴區間為 1.07 ～ 1.25），兩者的臨床意義會是相同的。

注意，包含 37 個研究的後設分析中分配 83% 的權重給 18 個大的研究（右側列的條圖所示）。換言之，如果小樣本研究會引入偏誤，那麼小樣本研究給予更小的權重一定程度上可以減小這種傾向。但是，與固定效果模型後設分析相比，隨機效果後設分析給予小型研究相對大的權重，因此，隨機效果的累積後設分析不可能顯示這種傾向。

這個方法的主要優點是可以得到效應不偏估計量（在強的模型假設下），而且有助於直觀的圖形展示。與修剪填補法相比，該法不會被一兩個異常研究影響。

30.9 引例結果摘要

不同的統計方法從不同的角度探討了偏誤的問題。不必期望不同的方法相互一致，因為每種方法回答不同的問題。而目標應該是綜合不同方法所得到的不同的資訊。

資料概述

引例分析中有相當多的研究，儘管大部分顯示了增加的風險，但是只有少數得到有統計顯著性的結果。說明這種情況下，基於統計顯著性考察出版偏誤的方法不會有檢定力。

有偏誤存在的證據嗎？

漏斗圖明顯不對稱，大量小型研究集中在均值的右側。Egger 的檢驗得到有統計顯著性的 p 值，證實了漏斗圖的視覺印象。秩相關（rank correlation）檢驗沒有得到有統計顯著性的 p 值，但這可能是由於檢定力低的問題。總之，小樣本研究傾向於呈現吸二手菸與肺癌間更高的關聯程度。

觀察到的關聯可能僅是偏誤的結果嗎？

Rosenthal 失敗安全數為 398，說明需要近 400 個平均風險比為 1.0 的研究加入到後設分析才能使得累計效果值沒有統計顯著性的結果。Orwin 失敗安全數為 103，說明需要超過 100 個平均風險比為 1.0 的研究加入到後設分析才能使得累計效果值變得不重要（定義風險比為 1.05）。假定後設分析的作者僅能檢索到 37 個關於吸二手菸與肺癌關係的研究，不可能有接近 400 或 100 個研究被漏掉，且即使說我們高估了吸二手菸導致的風險，其實際風險不可能是 0。

偏誤對風險比可能有什麼影響？

完整的後設分析顯示吸二手菸可以導致肺癌風險增加 20%。相反的，僅基於大型研究的後設分析則呈現肺癌風險增加 16%，就算因修剪填補法建議而刪

除不對稱的研究，肺癌風險也是增加 15%。前面提到分析出版偏誤的目的應該是將結果分為三類：1. 偏誤的影響不重要；2. 偏誤的影響是存在的，但是不會改變主要的結論；3. 主要的結論可能是有問題的。這個後設分析看起來屬於第二種情形。小型研究有著較大的效應的證據，和我們對於出版偏誤的假設是一致的。但是沒有理由懷疑主要結論的有效性，即吸二手菸可以引起肺癌風險的增加，這在臨床上是有意義的。

30.10 一些重要的警告

本章討論的大部分方法是尋找小型研究中的效果值更大的證據，然後將這解釋為出版偏誤存在的證據。效果值和樣本大小的關係假設及相關和迴歸檢驗是森林圖的核心，同時，修剪填充法和僅基於大型研究進行後設分析的基本邏輯也是基於兩者的關係假設。

因此，警覺這些方法的過程是否合乎以下幾個警告。不同的分析指標（如風險差或相對風險性）下這些方法可能得到不同的圖形，這些方法可能輕易漏掉真正的離散度，只有當樣本大小的離散度不大或研究數不是很小時，這些方法可能是有效的。即使這些條件滿足，檢驗（相關和迴歸檢驗）方法常常檢定力較低，因此，沒有找到不對稱的證據並不能得到肯定的結論。

30.11 小型研究效應

同樣重要的是，當有不對稱的明確證據時，也不能假定這反映了出版偏誤的存在。小樣本研究中效果值可能是大的，因為我們得到了一個小型研究的有偏樣本；也有可能是因為完全不相關的原因導致的小型研究中的效果值確實比較大。

例如，小型研究中納入病人病情較重，因此，更可能表現出好的藥效（如新藥的早期試驗）；或者，小型研究有著比大型研究更好的（或更差的）品質控制。Sterne 等人（2001） 用小型研究效應來描述小型研究中高的效果值的情況，強調這種效應的機制不清楚的事實。

調整出版偏誤是應當考慮這種情形，例如，我們應該呈現「如果不對稱是由於偏誤，我們的分析建議調整效果值為……」，而不是斷言「不對稱是由於偏誤，因此，真正的效應為……」。

30.12 結語

後設分析中包含出版偏誤的評價很重要，這既可以告訴評價者結果的穩健性，同時又警告他們結果是可疑的。這對於保證單個後設分析的真實性是重要的；同時對於保證這個領域的真實性也是重要的。如果後設分析忽視了可能的偏誤，事後又被發現結果是不正確的，人們就會形成一種認識：後設分析不值得相信。

Summary Points 要點概覽

- 當包含在後設分析中的研究和所有應該被包含的研究有著系統的差別時，導致出版偏誤，一般來說，效應高於平均效應的研究更有可能被出版，這將導致合併效果值的高估。
- 有一些評價出版偏誤可能影響的方法。應用這些方法，我們呈現如果消除偏誤得到調整估計，1. 結果可能沒有根本性改變；2. 效果值可能改變，但是基本結論，即處理效應有效或無效不會改變；3. 基本結論改變。
- 處理出版偏誤的方法要求做出許多基本假設，包括偏誤導致結果的方式，偏誤服從一定的假設等。
- 出版偏誤是後設分析中的問題，也是敘述性綜論或任何涉及文獻檢索工作中的問題。

Further Reading 延伸閱讀

Chalmers, T.C., Frank, C.S., & Reitman, D. (1990). Minimizing the three states of publication bias. *JAMA* 263: 1392–1395.

Dickersin, K., Chan, S., Cha.mers,T.C., Sacks, H.S., & Smith, H. (1987) Publication bias in clinical trials. *Controlled Clinical Trials* 8: 348–353.

Dickersin, K,. Min, Y.L., & Meinert, C.L. (1992). Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *JAMA* 267: 374–378.