Commentary

Standardization: teach it correctly (and leave it behind)

I still remember the lecture on standardization at the beginning of my MPH studies, several weeks into the first course in epidemiology. The word sounded about right – standards have always been praised – and the computation was simple, but I had no clear understanding why some kind of bias was removed. The instructor offered a hand-waving explanation that revolved around “making two groups comparable”. Who would have dared to question both “standardized” and “comparable”?

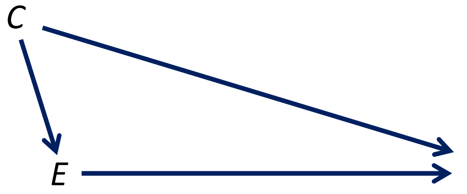
As far as I can tell from epidemiology textbooks on my bookshelf, matters have not improved much in the past two decades. Although many authors teach how to compute standardized rates (or probabilities), they don’t solidly explain why the difference between, and the ratio of, two standardized measures of frequency are unconfounded measures of association. Nor do they explain the arbitrary choice of a standard population, and how to reconcile different estimated effects from an infinite number of possible “standards”.

Here is an attempt to teach the true logic of the method – and its faulty component.

Confounding and deconfounding

Figure 1 shows the causal structure of confounding bias. The marginal association between *E* and *D* has two sources: the causal path *E*→*D* that we wish to estimate, and the confounding path *E*←*C*→*D*, a source of bias.

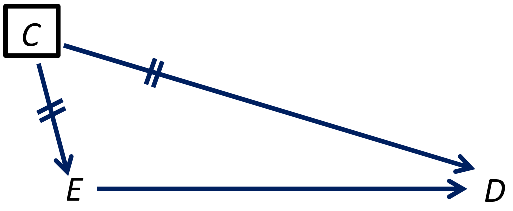
Figure 1. Confounding bias



Next, recall a simple principle: Strict conditioning on *C* (restricting *C* to one value) will eliminate the confounding path, because upon restriction *C* will not be associated with any variable. In particular, *C* will be disconnected from both *E* and *D* (Figure 2). Therefore, the association between *E* and *D*, conditional on *C*=c, is an unbiased estimator of the effect *E*→*D*.

Lastly, we may estimate the effect of *E* on *D* for each value of *C* and then replace the multiple estimates with a single weighted average (assuming negligible effect modification). Whichever weights are used, any such average is free of confounding bias by *C*.

Figure 2. Deconfounding by conditioning



Standardization

Standardization seems to be doing something quite different (but the truth will be discovered shortly).

In the first step we choose a “standard population”, which means a population with some distribution of *C*. In the simplest case of a binary *C*, the standard population contains N_1+N_2 people or person-time, where the subscripts 1 and 2 denote *C*=1 and *C*=2, respectively.

In the second step we compute a standardized rate (or a standardized probability) for exposed (R^e) and unexposed (R^u) as follows:

$$R^e = \frac{R_1^e N_1 + R_2^e N_2}{N_1 + N_2} = R_1^e \frac{N_1}{N_1 + N_2} + R_2^e \frac{N_2}{N_1 + N_2}$$

$$R^u = \frac{R_1^u N_1 + R_2^u N_2}{N_1 + N_2} = R_1^u \frac{N_1}{N_1 + N_2} + R_2^u \frac{N_2}{N_1 + N_2}$$

where

R_1^e and R_2^e denote the rate of *D*=1 in exposed people who belong to the stratum *C*=1 and the stratum *C*=2, respectively.

R_1^u and R_2^u denote the rate of *D*=1 in unexposed people who belong to the stratum *C*=1 and the stratum *C*=2, respectively.

To simplify, let

$$w_1 = \frac{N_1}{N_1 + N_2}$$

$$w_2 = \frac{N_2}{N_1 + N_2}$$

Then,

$$R^e = R_1^e w_1 + R_2^e w_2$$

$$R^u = R_1^u w_1 + R_2^u w_2$$

$$(w_1 + w_2 = 1)$$

Both standardized rates are, therefore, a weighted average of *C*-specific rates, using a single set of weights: the distribution of *C* in the standard population.

Finally, we compute the standardized rate difference (RD_s=R^e–R^u) and the standardized rate ratio (RR_s=R^e/R^u), and call them "adjusted" measures. As far as *C* is concerned, both are unbiased.

How does standardization work?

Somehow, standardization has generated unconfounded measures of effect. But why? We did not condition the association between *E* and *D* on *C*, as we should have done (Figure 2). We did not estimate the effect of *E* on *D* for each value of *C* and then computed a weighted average. Or did we?

Yes, we did. A little math shows that standardized differences and ratios are nothing more than a weighted average of conditional associations. Standardization is just an odd way to deconfound in the classic manner.

The standardized rate difference (RD_s) can be written as a weighted average of the *C*-specific rate differences (RD_1 and RD_2).

Proposition:

$$RD_s = R^e - R^u = RD_1 w_1 + RD_2 w_2$$

for some weights w_1 and w_2 $(w_1 + w_2 = 1)$

Proof:

$$RD_s = R^e - R^u =$$

$$= (R_1^e w_1 + R_2^e w_2) - (R_1^u w_1 + R_2^u w_2)$$

$$= (R_1^e - R_1^u) w_1 + (R_2^e - R_2^u) w_2$$

$$= RD_1 w_1 + RD_2 w_2 \qquad \text{QED}$$

Likewise, the standardize rate ratio (RR_s) can be written as a weighted average of the *C*-specific rate ratios (RR_1 and RR_2).

Proposition:

$$RR_s = \frac{R^e}{R^u} = RR_1 u_1 + RR_2 u_2$$

for some weights u_1 and u_2 $(u_1 + u_2 = 1)$

Proof:

$$RR_s = \frac{R^e}{R^u} = \frac{R_1^e w_1 + R_2^e w_2}{R_1^u w_1 + R_2^u w_2}$$

$$= \frac{\frac{R_1^e}{R_1^u} R_1^u w_1 + \frac{R_2^e}{R_2^u} R_2^u w_2}{R_1^u w_1 + R_2^u w_2}$$

$$= \frac{RR_1 R_1^u w_1 + RR_2 R_2^u w_2}{R_1^u w_1 + R_2^u w_2}$$

$$= RR_1 \frac{R_1^u w_1}{R_1^u w_1 + R_2^u w_2} + RR_2 \frac{R_2^u w_2}{R_1^u w_1 + R_2^u w_2}$$

To simplify, let

$$u_1 = \frac{R_1^u w_1}{R_1^u w_1 + R_2^u w_2}$$

$$u_2 = \frac{R_2^u w_2}{R_1^u w_1 + R_2^u w_2}$$

$$(u_1 + u_2 = 1)$$

Then,

$$RR_s = RR_1 u_1 + RR_2 u_2 \qquad \text{QED}$$

And that's why standardization works. No need to bring up any argument about standardized–adjusted–expected–predicted–summary–weighted–fictional–comparable rates.

The story – turned upside down

You can find similar algebra in *Modern Epidemiology*[1] (a dense textbook that should be praised for its comprehensiveness more than for clarity and order).

On pages 266-7, we read:

"The following algebra shows that a standardized rate difference is the weighted average of the stratum-specific rate differences..."

"Note that both RR_w [standardized risk ratio] and IR_w [standardized rate ratio] are weighted averages of stratum-specific ratios..."

Peculiarly to my mind, these results are mentioned as a casual observation, rather than the essence of the causal matter. The authors don't link the logic of standardization to a weighted average of stratum-specific effects. On the contrary, in various places we find attempts to inject logic into the original weights that are used to compute standardized rates or standardized probabilities. For example, about 120 pages later (under the heading "regression standardization") we read the following claim, the seeds of which were planted long ago.[2]

"Ideally, the standardized weights w(**z**) will reflect the distribution of the covariates **Z** in a target population of subject-matter relevance." (Page 386)

What exactly is "a target population of subject-matter relevance", which contains that ideal distribution of *C*? No, it is not what you probably have in mind. "The target population" is a foggy idea that thrives on ambiguity on the part of the writer, combined with misinterpretation on the part of many readers. Yet it managed to penetrate the literature in various guises,[3] most recently in the form of marginal structural models.[4-6] The epistemological poverty of that idea – an idol of deterministic causal inference – is exposed elsewhere.[7,8]

On the weighting of estimated effects

As the algebra shows, standardization is not conceptually different from the Mantel-Haenszel procedure – another method to deconfound, which is typically taught later in introductory courses. In both methods we condition on the confounder and compute a weighted average of the estimated effect across its strata. The two methods differ only in the chosen weights. Two questions may therefore be asked.

- Should we prefer a weighted average across the values of *C* over *C*-specific estimates?

- How do we choose the weights?

The first question alludes to the possibility of effect modification by *C* ($RD_1 \neq RD_2$, $RR_1 \neq RR_2$, or both). If effect modification is strong – and we compute a single estimate rather than two *C*-specific estimates – the estimator is plagued with effect modification bias.[9] For many weights, neither the effect of *E* when *C*=1, nor the effect of *E* when *C*=2, is captured unbiasedly by their weighted average. (And for *any* set of weights, the estimator is severely biased for at least one effect.)

Now to the second question: Assuming no effect modification, how do we choose the weights? Since the weighted average is unbiased – for any set of weights – the answer comes from a different domain: the weights determine the variance. We can therefore trace the following trail:

Choosing a standard population is equivalent to choosing weights for the estimated effect in the strata of the confounder, which in turn, will determine the variance of the unconfounded measure of effect.

Needless to say, we always prefer an unbiased estimator with the smallest variance.

Minimizing the variance

Continuing with a binary confounder, let's examine the variance of an unconfounded, weighted average of two *C*-specific measures of effect: rate ratio, rate difference, probability ratio, probability difference, log rate ratio, log probability ratio, and so on. To simplify, let *M* denotes the measure, and let *W* denotes the weights (formerly, *w* or *u*). As before, the subscripts 1 and 2 denote *C*=1 and *C*=2, respectively.

$$M = M_1W_1 + M_2W_2$$

$$Var(M) = Var(M_1W_1 + M_2W_2)$$

$$= W_1^2Var(M_1) + W_2^2Var(M_2)$$

Recalling $W_2 = 1 - W_1$

$$Var(M) = W_1^2Var(M_1) + (1 - W_1)^2Var(M_2)$$

A little algebra takes us to the following quadratic function of W_1:

$$Var(M) =$$

$$\big(Var(M_1) + Var(M_2)\big)W_1^2 - 2Var(M_2)W_1 + Var(M_2)$$

Setting the first derivative to zero, we find the value of W_1 which minimizes $Var(M)$:

$$W_1 = \frac{Var(M_2)}{Var(M_1) + Var(M_2)}$$

which can also be written as

$$W_1 = \frac{\frac{1}{Var(M_1)}}{\frac{1}{Var(M_1)} + \frac{1}{Var(M_2)}}$$

$$W_2 = 1 - W_1 = \frac{\frac{1}{Var(M_2)}}{\frac{1}{Var(M_1)} + \frac{1}{Var(M_2)}}$$

Using these weights, we find

$$Var(M)_{min} = \frac{1}{\frac{1}{Var(M_1)} + \frac{1}{Var(M_2)}}$$

These results accord with intuition. "Good" weights are typically related to the variance, and the formula allocates the weights by partitioning the sum of the inverse of the variance in the two strata of *C*. (The proof for any number of strata is shown in Appendix A.)

$Var(M_1)$ and $Var(M_2)$ can be estimated from the data, which means that the optimal weights, W_1 and W_2, can be estimated too. Recalling that the weights are functions of the standard population (N_1, N_2), we can trace the steps back and estimate the proportions of N_1 and N_2 that minimize the variance of *M*. This is the distribution of the confounder in the preferred standard population! Any other choice will result in a different set of weights and a larger variance of *M*. It cannot be justified on statistical grounds.

But why bother? Why use tortuous standardization in the first place, rather than estimate the optimal weights directly and compute a weighted average of *C*-specific effects? Indeed, that's exactly what other methods do. Trying to weigh by the inverse of stratum-specific variances is the essence of the Mantel-Haenszel estimator and the Woolf estimator. Weighting in meta-analysis is another example (though the issue is not confounding).

Indirect standardization

Standardization has been classified into two types, "direct" and "indirect", but the distinction is artificial. It can be shown that the SMR – the hallmark of indirect standardization – can be computed according to the generic procedure that was outlined earlier: choosing a standard population and dividing one weighted average of *C*-specific rates by another. Hence, the SMR may also be written as a weighted average of *C*-specific rate ratios: $SMR = RR_1u_1 + RR_2u_2$ for a binary *C*, and $SMR = \sum RR_iu_i$ in general. An example of that alternative computation is shown in Appendix B.

The "indirect" method is distinguished from the "direct" method only in the choice of the standard population. In the former, one of the groups serves as the standard, so we encounter what may be called "self-standardization". The rate in one group is standardized to the confounder distribution in that group. But self-standardization is a redundant exercise. Simple math can show that a self-standardized rate (or probability) is equal to the crude rate (or probability).

Many authors write that indirect standardization is preferred when one of the two groups is small, and that the smaller group should be chosen as the standard. Under some assumptions, that choice might approximate better the optimal weights than many arbitrarily chosen standards, but it is still not better than the optimal weights – those weights that approximate the inverse of the variance of stratum-specific estimated effects. Since the variances in question can be estimated from the data, indirect standardization is not needed either.

Some authors argue that the computation of the SMR is justified when *C*-specific rates are missing in one group. Not so. In that case, standardization also willfully opens the door to effect modification bias:[9,10] The missing rates preclude the estimation of stratum-specific effects (RR_i), and a single weighted average – the SMR – might conceal dissimilar effects in the strata of *C*. (No, it is not analogous to unintentional omission of unknown effect modifiers: the fact that bias might unknowingly creep into science is no excuse for knowingly letting a possible bias creep in.)

In retrospect

Why did so many minds fail to recognize the faulty feature of standardization as described here, namely, deconfounding with arbitrary weights and penalized variance? At least three explanations may be offered, none of which have to do with "a target population of subject-matter relevance".

First, many authors don't present the algebra which shows that a standardized measure of effect is a weighted average of stratum-specific effects (and those who do, note it in passing). Some authors don't even end the story with the computation of a measure of effect. They just tell the reader that the standardized rates "are comparable", "may be compared", "account for different distributions", and the like.

Second, a weighted average across the strata of a confounder is always constrained between the smallest and the largest stratum-specific estimates. If these two estimates are not too far apart (i.e., modest effect modification is estimated), a weighted average will not vary much, no matter which weights are used. Similarly, small departures from the optimal weights will result in only a slight increase in the variance of a weighted average.[11] Therefore, in many examples both the estimate and the variance are not very sensitive to the arbitrary choice of a standard population. The faulty component of the method doesn't leave its mark.

Third, many authors teach that the inverse of the variance is the appropriate weight for a weighted average of stratum-specific effects, but they don't offer a full explanation. They don't state that the ultimate goal is to minimize the variance of an unconfounded measure of effect, and they don't show the math that satisfies that goal. Read, for example, a typical explanation in a well-written book, *Statistics for Epidemiology*, of why we should use inverse-of-variance weights to compute the unconfounded odds ratio:[12]

"The weights account for the fact that the stratum Odds Ratios are estimated with different precision. In averaging quantities that are subject to varying levels of random uncertainty, it is best to use weights that are proportional to the *reciprocal* of the variance of the underlying estimator, so that imprecise components are given low weight." (Page 129)

Not a word on minimizing the variance of the unconfounded odds ratio. Not a word on the connection between that variance and the weights. It is best because imprecise components are given low weight and precise components are given high weight. True, but not the whole truth.

Epilogue

Standardization is founded on two valid ideas: conditional associations and their weighted average. Nonetheless, the choice of the weights does not follow the expected logic, namely, minimizing the variance of a weighted average.

We may teach standardization as a historical method to deconfound that contains a kernel of wisdom and its statistical flaw is now understood. There is no justification, however, for continued use of this method in epidemiology. Standardization should be abandoned, along with the empty term "the target population of subject-matter relevance". As for the latter: either there is no such thing (which is obvious in many examples of standardization), or such a thing is irrelevant to causal knowledge.[7,8]

Acknowledgement: Doron Shahar – for writing Appendix A and for comments on a draft manuscript.

References

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology* (third edition), Lippincott Williams & Wilkins, 2008
2. Greenland S, Maldonado G. The interpretation of multiplicative-model parameters as standardized parameters. *Statistics in Medicine* 1994;13:989-99
3. Maldonado G, Greenland S. Estimating causal effects. *International Journal of Epidemiology* 2002;31:422-9
4. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550-560
5. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology* 2003;14:680–686
6. Weuve J, Tchetgen Tchetgen EJ, Glymour MM, et al. Accounting for bias due to selective attrition: the example of smoking and cognitive decline. *Epidemiology* 2012;23:119-28
7. Shahar E. Estimating causal parameters without target populations. *Journal of Evaluation in Clinical Practice* 2007;13:814-816
8. Shahar E, Shahar DJ. Marginal structural models: much ado about (almost) nothing. *Journal of Evaluation in Clinical Practice* 2013;19:214-22 [Epub Aug 23, 2011]. Part 2 is posted on this website.
9. Shahar E, Shahar DJ. Causal diagrams and three pairs of biases. In: *Epidemiology – Current Perspectives on Research and Practice* (Lunet N, Editor).http://www.intechopen.com/books/epidemiology-current-perspectives-on-research-and-practice, 2012:pp. 31-62
10. Choi BC, de Guia NA, Walsh P. Look before you leap: stratify before you standardize. *American Journal of Epidemiology* 1999;149:1087-96
11. Kalton G. Standardization: a technique to control for extraneous variables. *Journal of the Royal Statistical Society*, *Series C* (*Applied Statistics*) 1968;17:118-136
12. Jewell NP. *Statistics for Epidemiology*, Chapman & Hall/CRC, 2004

Appendix A

Let C be a confounder for the effect $E \to D$ that is not an effect modifier. Let $1, ..., n$ be the values of C, and let M_i be the estimated effect of E on D in the stratum $C = i$. The observations are assumed to be independent, and therefore, the M_i are independent. (M_i may be any measure of effect.) Then, any weighted average $M = \sum_{i=1}^n W_i M_i$ ($\sum_{i=1}^n W_i = 1$ and $W_i \geq 0, \forall i$) is an estimate from an unbiased estimator of the effect $E \to D$ (as far as C is concerned). Since the estimator is unbiased regardless of the weights, we want to choose weights that minimize the variance of M

Proposition: *For independent* $\{M_i\}_{i=1}^n$ *where* $n \geq 2$, $\mathrm{Var}(\sum_{i=1}^n W_i M_i)$ *is minimized when* $W_i = \frac{\frac{1}{\mathrm{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}}$, *and its minimum value is* $\frac{1}{\sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}}$.

The first proof uses the method of Lagrange multipliers.

Proof 1: $\mathrm{Var}(\sum_{i=1}^n W_i M_i) = \sum_{i=1}^n \mathrm{Var}(W_i M_i) = \sum_{i=1}^n W_i^2 \mathrm{Var}(M_i)$, because the M_i are independent. We wish to minimize the previous expression under the constraint that $\sum_{i=1}^n W_i = 1$ and $W_i \geq 0, \forall i$. The set T of all $(W_1, ..., W_n) \in \mathbb{R}^n$ for which $\sum_{i=1}^n W_i = 1$ and $W_i \geq 0, \forall i$ is closed and bounded. The extrema in the interior of T can be found by considering only the first constraint, which may be written as $\sum_{i=1}^n W_i - 1 = 0$. Later we shall find the extrema on the boundary of T.

To find the extrema in the interior of T, let $F(W_1, ..., W_n, \lambda) = \sum_{i=1}^n W_i^2 \mathrm{Var}(M_i) - \lambda(\sum_{i=1}^n W_i - 1)$. By the method of Lagrange multipliers, the values of $W_1, ..., W_n$ for which $\frac{\partial F}{\partial W_j} = 0$ are the critical points of $\mathrm{Var}(\sum_{i=1}^n W_i M_i)$. (These contain all the extrema of $\mathrm{Var}(\sum_{i=1}^n W_i M_i)$ in the interior of T.) $\frac{\partial F}{\partial W_j} = 2W_j \mathrm{Var}(M_j) - \lambda$. If we set $\frac{\partial F}{\partial W_j}$ equal to zero, we find that $W_j = \frac{\lambda/2}{\mathrm{Var}(M_j)}$. Since $1 = \sum_{j=1}^n W_j = \sum_{j=1}^n \frac{\lambda/2}{\mathrm{Var}(M_j)} = \frac{\lambda}{2} \sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}$, then $\lambda = 2 \frac{1}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}}$. Thus, $W_j = \frac{\lambda/2}{\mathrm{Var}(M_j)} = \frac{\frac{1}{\mathrm{Var}(M_j)}}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}} > 0$. $(W_1, ..., W_n)$ is indeed in the interior of T for those values of W_j. By changing our index, $W_i = \frac{\frac{1}{\mathrm{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}}$. For these values of W_i, $\mathrm{Var}(\sum_{i=1}^n W_i M_i) = \sum_{i=1}^n W_i^2 \mathrm{Var}(M_i) = \sum_{i=1}^n \frac{\frac{1}{\mathrm{Var}(M_i)^2}}{(\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)})^2} \mathrm{Var}(M_i) = \frac{\sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}}{(\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)})^2} = \frac{1}{\sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}}$.

The boundary of T is characterized by having some of the W_i equal zero. For any point on the boundary, let $S = \{i : W_i \neq 0\}$. At such a point, $\mathrm{Var}(\sum_{i=1}^n W_i M_i) = \mathrm{Var}(\sum_{i \in S} W_i M_i)$. Using the method of Lagrange multipliers again, the critical points of $\mathrm{Var}(\sum_{i=1}^n W_i M_i) = \mathrm{Var}(\sum_{i \in S} W_i M_i)$ are found to be $(W_1, ..., W_n)$ where $W_i = \begin{cases} \frac{\frac{1}{\mathrm{Var}(M_i)}}{\sum_{j \in S} \frac{1}{\mathrm{Var}(M_j)}} & \text{if } i \in S \\ 0 & \text{if } i \notin S \end{cases}$

(These contain all the extrema of $\mathrm{Var}(\sum_{i=1}^n W_i M_i)$ on the boundary of T.) For these values of W_i, $\mathrm{Var}(\sum_{i=1}^n W_i M_i) = \sum_{i=1}^n W_i^2 \mathrm{Var}(M_i) = \sum_{i \in S} \frac{\frac{1}{\mathrm{Var}(M_i)^2}}{(\sum_{j \in S} \frac{1}{\mathrm{Var}(M_j)})^2} \mathrm{Var}(M_i) = \frac{\sum_{i \in S} \frac{1}{\mathrm{Var}(M_i)}}{(\sum_{j \in S} \frac{1}{\mathrm{Var}(M_j)})^2} = \frac{1}{\sum_{i \in S} \frac{1}{\mathrm{Var}(M_i)}} \geq \frac{1}{\sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}}$, because $\sum_{i \in S} \frac{1}{\mathrm{Var}(M_i)} \leq \sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}$.

Thus, of all critical points, $\mathrm{Var}(\sum_{i=1}^n W_i M_i)$ is smallest when $W_i = \frac{\frac{1}{\mathrm{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}}$. Therefore,

$\mathrm{Var}(\sum_{i=1}^n W_i M_i)$ is minimized when $W_i = \frac{\frac{1}{\mathrm{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}}$, and its minimum value is $\frac{1}{\sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}}$. QED

The second proof is done by induction.

Proof 2: The case $n = 2$ will be our base case for the induction. (This was shown in the article.)

$$\begin{aligned}
\mathrm{Var}\Big(\sum_{i=1}^{2} W_i M_i\Big) &= \mathrm{Var}(W_1 M_1 + W_2 M_2) \\
&= W_1^2 \mathrm{Var}(M_1) + W_2^2 \mathrm{Var}(M_2) && \text{because } M_1 \text{ and } M_2 \text{ are independent} \\
&= W_1^2 \mathrm{Var}(M_1) + (1 - W_1)^2 \mathrm{Var}(M_2) \\
&= W_1^2 (\mathrm{Var}(M_1) + \mathrm{Var}(M_2)) - 2W_1 \mathrm{Var}(M_2) + \mathrm{Var}(M_2)
\end{aligned}$$

The above expression has a minimum when $W_1 = \frac{\mathrm{Var}(M_2)}{\mathrm{Var}(M_1)+\mathrm{Var}(M_2)}$. Dividing the numerator and denominator by $\mathrm{Var}(M_1)\mathrm{Var}(M_2)$ we find that $W_1 = \frac{\frac{1}{\mathrm{Var}(M_1)}}{\frac{1}{\mathrm{Var}(M_1)} + \frac{1}{\mathrm{Var}(M_2)}}$, and $W_2 = 1 - W_1 = \frac{\frac{1}{\mathrm{Var}(M_2)}}{\frac{1}{\mathrm{Var}(M_1)} + \frac{1}{\mathrm{Var}(M_2)}}$. The minimum variance is then $\sum_{i=1}^2 \frac{\frac{1}{\mathrm{Var}(M_i)^2}}{(\sum_{j=1}^2 \frac{1}{\mathrm{Var}(M_j)})^2} \mathrm{Var}(M_i) = \frac{\sum_{i=1}^2 \frac{1}{\mathrm{Var}(M_i)}}{(\sum_{j=1}^2 \frac{1}{\mathrm{Var}(M_j)})^2} = \frac{1}{\sum_{i=1}^2 \frac{1}{\mathrm{Var}(M_i)}}$.

For the induction step, suppose that $\mathrm{Var}\Big(\sum_{i=1}^n W_i M_i\Big)$ is minimized when $W_i = \frac{\frac{1}{\mathrm{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\mathrm{Var}(M_j)}}$ for some $n \geq 2$, and its minimum value is $\frac{1}{\sum_{i=1}^n \frac{1}{\mathrm{Var}(M_i)}}$. Then,

$$\begin{aligned}
\mathrm{Var}\Big(\sum_{i=1}^{n+1} W_i M_i\Big) &= \mathrm{Var}\Big(\sum_{i=1}^{n} W_i M_i + W_{n+1} M_{n+1}\Big) \\
&= \mathrm{Var}\Big(\sum_{i=1}^{n} W_i M_i\Big) + \mathrm{Var}(W_{n+1} M_{n+1}) && \text{because the } M_i \text{ are independent} \\
&= \mathrm{Var}\Big(\Big(\sum_{j=1}^{n} W_j\Big)\Big(\sum_{i=1}^{n} \frac{W_i}{\sum_{j=1}^n W_j} M_i\Big)\Big) + \mathrm{Var}(W_{n+1} M_{n+1}) \\
&= \Big(\sum_{j=1}^{n} W_j\Big)^2 \mathrm{Var}\Big(\sum_{i=1}^{n} \frac{W_i}{\sum_{j=1}^n W_j} M_i\Big) + W_{n+1}^2 \mathrm{Var}(M_{n+1}) \\
&= (1 - W_{n+1})^2 \mathrm{Var}\Big(\sum_{i=1}^{n} \frac{W_i}{\sum_{j=1}^n W_j} M_i\Big) + W_{n+1}^2 \mathrm{Var}(M_{n+1}) \\
&= (1 - W_{n+1})^2 \mathrm{Var}\Big(\sum_{i=1}^{n} U_i M_i\Big) + W_{n+1}^2 \mathrm{Var}(M_{n+1}) && \text{where } U_i = \frac{W_i}{\sum_{j=1}^n W_j} \text{ are weights} \\
&= W_{n+1}^2 \Big(\mathrm{Var}(M_{n+1}) + \mathrm{Var}\Big(\sum_{i=1}^{n} U_i M_i\Big)\Big) - 2W_{n+1} \mathrm{Var}\Big(\sum_{i=1}^{n} U_i M_i\Big) + \mathrm{Var}\Big(\sum_{i=1}^{n} U_i M_i\Big)
\end{aligned}$$

The U_i do not depend on W_{n+1}. So for any possible values of U_i, the above expression is minimized when $W_{n+1} = \frac{\text{Var}\Big(\sum_{i=1}^n U_iM_i\Big)}{\text{Var}(M_{n+1})+\text{Var}\Big(\sum_{i=1}^n U_iM_i\Big)}$. Furthermore, we assumed that $\text{Var}\Big(\sum_{i=1}^n U_iM_i\Big)$ is minimized when $U_i = \frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}$. Therefore, $\text{Var}\Big(\sum_{i=1}^{n+1} W_iM_i\Big)$ is minimized when

$$\begin{aligned}
W_{n+1} &= \frac{\text{Var}\Big(\sum_{i=1}^n U_iM_i\Big)}{\text{Var}(M_{n+1}) + \text{Var}\Big(\sum_{i=1}^n U_iM_i\Big)} \\
&= \frac{\sum_{i=1}^n U_i^2\text{Var}(M_i)}{\text{Var}(M_{n+1}) + \sum_{i=1}^n U_i^2\text{Var}(M_i)} \\
&= \frac{\sum_{i=1}^n \left(\frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}\right)^2\text{Var}(M_i)}{\text{Var}(M_{n+1}) + \sum_{i=1}^n \left(\frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}\right)^2\text{Var}(M_i)} \\
&= \frac{\frac{1}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}}{\text{Var}(M_{n+1}) + \frac{1}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}} \\
&= \frac{\frac{1}{\text{Var}(M_{n+1})}}{\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)}} \qquad \text{after mulitplying by } \frac{\frac{1}{\text{Var}(M_{n+1})}\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}{\frac{1}{\text{Var}(M_{n+1})}\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}
\end{aligned}$$

and when

$$\begin{aligned}
W_i &= \Big(\sum_{j=1}^n W_j\Big)U_i \\
&= (1 - W_{n+1})U_i \\
&= \Bigg(1 - \frac{\frac{1}{\text{Var}(M_{n+1})}}{\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)}}\Bigg)\frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}} \\
&= \Bigg(\frac{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}{\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)}}\Bigg)\Bigg(\frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}\Bigg) \\
&= \frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)}} \qquad \text{for } i \in \{1, ..., n\}
\end{aligned}$$

Therefore, $\text{Var}\Big(\sum_{i=1}^{n+1} W_iM_i\Big)$ is minimized when $W_i = \frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)}}$. The minimum variance is then $\sum_{i=1}^{n+1} \frac{\frac{1}{\text{Var}(M_i)^2}}{(\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)})^2}\text{Var}(M_i) = \frac{\sum_{i=1}^{n+1} \frac{1}{\text{Var}(M_i)}}{(\sum_{j=1}^{n+1} \frac{1}{\text{Var}(M_j)})^2} = \frac{1}{\sum_{i=1}^{n+1} \frac{1}{\text{Var}(M_i)}}$. This completes the induction step.

So for all $n \geq 2$, the weighted average $\text{Var}(\sum_{i=1}^n W_iM_i)$ is minimized when $W_i = \frac{\frac{1}{\text{Var}(M_i)}}{\sum_{j=1}^n \frac{1}{\text{Var}(M_j)}}$, and its minimum value is $\frac{1}{\sum_{i=1}^n \frac{1}{\text{Var}(M_i)}}$ QED

Appendix B

An example of indirect standardization, written as a weighted average of RRi*

	Miners (exposed)			General population (unexposed)					Standard population			
Age	N	Deaths (observed)	Rate (per 10^5)	Rate (per 10^5)	Deaths (expected)		**RRi**	R^u_i	w_i **	R^u_i x w_i	$\mathbf{u_i}$	RR_iu_i
20-24	74,598	10	13.41	12.26	9.15		1.09	12.26	0.14	1.71	0.05	0.055
25-29	85,077	20	23.51	16.12	13.71		1.46	16.12	0.16	2.57	0.08	0.110
30-34	80,845	22	27.21	21.54	17.41		1.26	21.54	0.15	3.26	0.10	0.121
35-44	148,870	98	65.83	33.96	50.56		1.94	33.96	0.28	9.46	0.28	0.541
45-54	102,649	174	169.51	56.82	58.33		2.98	56.82	0.19	10.91	0.32	0.961
55-59	42,494	112	263.57	75.23	31.97		3.50	75.23	0.08	5.98	0.18	0.618
Total	534,533	436			181	**2.41**			1.00	33.88	1.00	**2.41**
						SMR						ΣRR_iu_i

* The example (on the left) was taken from an online epidemiology course (http://ocw.jhsph.edu/courses/FundEpi/PDFs/Lecture7.pdf) where the classic computation of the SMR is shown. On the right, I computed the SMR as a weighted average of age-specific rate ratios, following the notation in the commentary

** The miners serve as the standard population