Published in final edited form as:
Ann Emerg Med. 2010 June ; 55(6): 544–552.e3. doi:10.1016/j.annemergmed.2010.01.002.

The Importance of "Shrinkage" in Subgroup Analyses

Ari M. Lipsky, M.D., Ph.D.[1,2,3,4], **Marianne Gausche-Hill, M.D.**[1,2,3], **Muna Vienna, M.D.**[1], and **Roger J. Lewis, M.D., Ph.D.**[1,2,3]

[1]Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA

[2]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA

[3]Los Angeles Biomedical Research Institute, Torrance, CA

[4]Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel

Abstract

Background—Subgroup analyses examine associations (e.g., between treatment and outcome) within subsets of a larger study sample. The traditional approach evaluates the data in each of the subgroups independently. More accurate answers, however, may be expected when the rest of the data are considered in the analysis of each subgroup, provided there are three or more subgroups.

Methods—We present a conceptual introduction to subgroup analysis that makes use of all the available data, and then illustrate the technique by applying it to a previously published study of pediatric airway management. Using WinBUGS, freely available computer software, we perform an empirical Bayesian analysis of the treatment effect in each of the subgroups. This approach corrects the original subgroup treatment estimates toward a weighted average treatment effect across all subjects.

Results—The revised estimates of the subgroup treatment effects demonstrate markedly less variability than the original estimates. Further, using these estimates will reduce our total expected error in parameter estimation, as compared to using the original, independent subgroup estimates. While any particular estimate may be adjusted inappropriately, adopting this strategy will, on average, lead to results that are more accurate.

Conclusions—When considering multiple subgroups, it is often inadvisable to ignore the rest of the study data. Authors or readers who wish to examine associations within subgroups are encouraged to use techniques that reduce the total expected error.

Corresponding Author: Ari M. Lipsky, M.D., Ph.D., Department of Emergency Medicine, Box 21, Harbor-UCLA Medical Center, 1000 West Carson Street, Torrance, California 90509, Tel: (310) 222-3501, Fax: (310) 782-1763, aril@alum.mit.edu.

Previously Presented: Abdi M, Gausche M, Lewis RJ. Estimating the Efficacy of an Intervention in Distinct Patient Subgroups: A Large Study of Prehospital Pediatric Airway Management. Abstract presented at the (February) 1998 Meeting of the Ambulatory Pediatric Association Region IX and X, Carmel, California.

Previously presented: Abdi M, Gausche M, Lewis RJ. Estimating the Efficacy of an Intervention in Distinct Patient Subgroups: Interpretation of Data from a Large Study of the Prehospital Pediatric Airway Management. Abstract presented at the (February) 1998 American Federation for Medical Research Western Regional Meeting, Carmel, California. *J Investig Med* 1998;46:150A.

Previously presented: Lewis RJ, Gausche M, Abdi M. Subgroup Analysis of Data From a Prospective Randomized Study of Prehospital Airway Management in Children Using Classical and Bayesian Techniques. Abstract presented at the (May) 1998 Annual Meeting of the Society for Academic Emergency Medicine, Chicago, Illinois. *Acad Emerg Med* 1998;5:428.

Every Subgroup Is Part of a Larger Study

A colleague of yours hands you a paper[1] which describes a prospective, controlled trial in critically ill children in the prehospital setting, comparing the outcomes of children assigned to receive endotracheal intubation (ETI) with the outcomes of those assigned to receive bag-valve-mask (BVM) ventilation. Not surprisingly, the observed survival proportions demonstrated considerable variation depending on the illness or injury which necessitated the airway intervention. For instance, the proportion of subjects surviving sudden infant death syndrome (SIDS)—for both ETI and BVM—was essentially zero, whereas for children who required ventilatory support due to respiratory failure, the proportion surviving was much better. Anticipating the variability in survival proportion associated with different disease states, as well as the possibility that the effect of ETI relative to BVM would vary by illness category, the investigators presented both the overall survival proportions in the two arms of the trial, and the survival proportions within each illness category (subgroup). And within each subgroup, they focused on the performance of ETI relative to BVM. Considering the study population as a whole, the trial demonstrated no benefit of ETI over BVM and even a trend toward harm. Based on these results, the use of prehospital pediatric ETI was abandoned in some emergency medical services systems.

Your colleague then asks you an important question: While the overall study demonstrated a trend toward harm in choosing prehospital ETI over BVM, certain patient subgroups seemed to have fared better with ETI than with BVM. If we choose to never intubate children in the prehospital setting, are we not ignoring those subgroups in which ETI appeared to be advantageous? Moreover, given that the number of children in some of the subgroups was quite limited, what is the best estimate of the effect of ETI compared to BVM in any particular subgroup, and how much uncertainty is there in the estimates of the effect of ETI in these subgroups?

Before reading further, stop and ask yourself what your instinctive response to the last question would be. For instance, in the subgroup of children denoted “Other,” which was comprised of all subjects whose illness did not fall into the prospectively-defined subgroups, there were 10 survivors and 5 deaths among those who received BVM, and 10 survivors and 1 death among those who received ETI. You could calculate that the odds ratio for survival in the ETI arm versus the BVM arm is 5 [i.e., (10/1)/(10/5)], and that the 95% confidence interval is 0.5 – 50.8. Although the confidence interval includes the possibility of there being no treatment effect, this finding, you think, should generate considerable doubt among providers about withholding a potentially life-saving intervention from children who fall into the “Other” subgroup. Your initial response, then, might be to inform your colleague that it is not prudent to completely remove pediatric ETI from the prehospital armamentarium.

We argue below that, in most circumstances, examining data from one subgroup in isolation (as we did above) yields a less accurate estimate of the treatment effect in that subgroup than considering the subgroup data in conjunction with the data from the rest of the study. It turns out that a better estimate of the odds ratio for survival in the “Other” subgroup is 1.6, with a 95% probability interval of 0.5 – 5.2. We discuss how we obtained this adjusted estimate below.

In order to provide an intuitive understanding of how we arrive at these corrected estimates, we will be using three different examples that involve proportions (though these methods can be used for estimating any type of parameter): 1. the proportion of heads in a series of coin flips; 2. the proportion of successful hits in a series of at-bats for baseball players; and 3. the proportion of children surviving in the pediatric trial of airway management. Because, however, for the pediatric trial we are interested in relative survival (ETI versus BVM) and odds are more convenient mathematically, we will be looking at the odds ratio of survival with

ETI versus BVM. And while we will begin the discussion with the example provided in a seminal paper by James and Stein that challenges our intuitions, this paper (including the worked example provided in the Technical Appendix) will use a different mathematical approach (empirical Bayes estimation) because we believe this approach will lead to a better understanding of the material.

Your Best Estimate

Imagine that you flip a coin ten times. This is not a normal coin—in fact, before you started this experiment you had no idea what the probability was of a flip turning up heads. If your ten flips yielded seven heads, what would be your best guess about the probability that an additional flip of this rather strange coin will turn up heads? It should be 0.7 (70%). Since we have assumed that you have no additional information about this coin, you have the best chance of "being correct" if you simply use the observed proportion as your best guess. You probably used this same intuition when you answered the question above about the best estimate of the effect of ETI on survival in the "Other" subgroup: the best estimate should just be the results obtained in that particular subgroup; after all, those are the most relevant data for answering the question.

The phrase, "being correct," which we wrote above, can be defined in several ways. One way of defining the most "correct" estimate in particular—used frequently in regression analyses—requires that we minimize the square of the distance between the estimate (in this case, the value derived from the data) and the unknown true value. This is called a "least squares" estimate: using the square ensures that both over- and under-estimates contribute positive values to the measurement of the error, and that the error grows fairly quickly the farther we are from the true value. For the coin flips, using any number other than 0.7—the observed proportion that turned up heads—as your estimate will probably increase the distance from your estimate to the truth, thus increasing the expected error. Applying this to the pediatric airway study above, your intuition likely led you to conclude that for each subgroup you should simply use each subgroup's data to generate the best estimate of the effect on survival of ETI relative to BVM in that particular subgroup.

For decades, however, it has been known that in the setting of multiple subgroups this intuition is misleading.[2] When the goal is to choose estimates for all the subgroups that yield the lowest overall estimation error, there are alternative approaches that are always likely to be better than (i.e., have a lower expected error than) the individual subgroup estimates. How could something always be better than that which seems so intuitively obvious?

Stein's Paradox & Baseball

Over half-a-century ago, Willard James and Charles Stein published an amazing finding.[2] They stated that as long as three or more groups are being compared, we can always expect to have a lower overall estimation error if we consider the data from all the groups when trying to estimate the parameters of the individual groups. While any particular estimate might be made worse, we are guaranteed to have lower expected error across all the estimates using this approach.

In 1977, Efron and Morris published in *Scientific American* an engaging review of what came to be known as Stein's paradox, so-called because the finding goes against the grain of what makes intuitive sense.3 (A more rigorous treatment of the subject by the same authors can be found in reference 4.) The authors describe the following, now oft-repeated example:

The baseball season is just underway, and the statistics on a group of players' batting averages are beginning to accumulate. Specifically, eighteen players' performances at their first forty-

five at-bats during the 1970 season were recorded. Just as your colleague asked you earlier about the subgroup analysis, Efron and Morris asked, given what we now know after these at-bats, what would best predict each player's final batting average to be once all the season's data are available? (To make the problem tractable, Efron and Morris define the true batting average as the player's average at the end of the 1970 season.) They go on to demonstrate that the predictions for individual players are much better if they take into consideration how the other players performed, essentially revising each player's batting average toward the overall group's average.

That should bother you. Why should the batting prowess of other, unrelated players in a league change how well you think a particular player will perform during the rest of the season? The answer lies in a concept known as "shrinkage."

More Coinage

Before we get to shrinkage, however, let's return to the coin example. This time you have in your hand a fair coin which, by definition, will turn up heads half the time as the number of flips becomes increasingly large. We know that by chance, however, shorter series of flips will vary from coming up 50% heads in a predictable way. After ten flips, for instance, we are most likely to get five heads, but it is also pretty likely that we will get three, four, six, or seven heads. And it is less likely, though not impossible, to get zero, one, two, eight, nine, or ten heads. (The probability of obtaining any one of these outcomes follows the binomial distribution which, after many flips, can be reasonably represented by a normal distribution.)

The fair coin is like a baseball player for whom we somehow know the true batting average (for "hitting heads") is fifty percent. If we only look at a short series of a coin's "at-bats" during the season, the observed average may be substantially higher or lower than fifty percent, but when the end of the coin's "season" is reached, the final average will be fifty percent. A real baseball player at the beginning of the season, on the other hand, is more like the strange coin mentioned earlier—they both have an underlying probability of a successful hit (or heads), but that probability remains unknown to us until after a much larger set of data have been collected.

Now imagine that we have ten strange coins, each with its own distinct but unknown probability of landing heads-up. Importantly, we have no reason to suspect (based, perhaps, on inspection) that any one coin has a greater probability of turning up heads than any other coin, an assumption known as exchangeability.[5] We then proceed to flip each coin ten times and record the percentage of heads in each series. If we consider each coin independently, then these percentages would represent our beliefs about the most likely value for the probability that each particular coin will turn up heads.

However, we know that these observed percentages will fluctuate around the true underlying likelihood of turning up heads solely due to the random chance described above. We would expect that a few of the series of coin tosses will overestimate their coin's true probability of turning up heads, a few will underestimate it, and the rest will likely cluster around the correct number. And because we had no reason to believe that any particular coin was more or less likely to turn up heads than any other coin, it is fair to examine more closely the coins with a relatively high or low percentage of flips that turned up heads: It is both more likely that they are true outliers (i.e., that these coins do indeed have higher or lower underlying probabilities of turning up heads) compared to the other coins, and also more likely that some of their exceptionalness is due to chance (i.e., that these coins are on streaks which do not reflect the true underlying probabilities of heads). The same reasoning applies to the coins closer to the middle of the pack, though to a lesser extent, in that their averages are likely less far from their true underlying probabilities: They are both less likely to be true outliers, and less likely to be far from their true underlying probabilities as a result of chance.

Of Shrinkage and Baseball

If we consider how the individual baseball players were doing in the beginning of the season, some will have performed exceptionally well and others exceptionally poorly, similar to our distribution of ten coins. This exceptionalness is exactly the issue. Though these outliers' batting ability may truly be exceptional, it is also reasonable to believe that they may be on lucky or unlucky batting streaks. These streaks, even among the intrinsically best and worst batters, are unlikely to continue for an entire season. We correct for this tendency of the exceptional players to be more extreme than truth by nudging each player's estimate toward the players' overall mean (i.e., the mean of all the players' averages thus far). (As an extreme example, consider a player who has hit successfully at his first two at-bats: Should we assume that he will maintain a perfect batting average for the rest of the season, or might it be better to nudge his current batting average toward the overall mean?) How much we correct (or nudge) each player's estimate depends in part on how far he is from the overall mean—how much of an outlier he is. Because we are correcting all the points toward a specific central point, this process is known as shrinkage: we are shrinking the distance between the individual players' estimates.

There is always the risk, of course, that the exceptionally good or bad player is truly exceptional and that we are either applying a correction where one is not needed, or even worse, nudging the estimate in the wrong direction. Though we cannot know which (if any) players' averages we have revised inappropriately, we can always expect that adopting a strategy that applies shrinkage to the group will outperform a strategy that uses the naïve, individual averages, and we therefore choose to use shrinkage. Put another way, the expected overall error in prediction is guaranteed to be less than it would have been had the individual batting averages been used to predict the players' performances during the remainder of the season.[4] In the Efron and Morris example, the shrinkage strategy paid off: The total squared error of the naïve averages turned out to be 0.077, while the total squared error of the James-Stein estimators was 0.022, representing a reduction in error by a factor of 3.5.

In addition to the distance from the overall mean, the amount of correction is also related to how much information is available. In the baseball example in which all the players had had the same number of at-bats, they all had roughly the same amount of information available. However, in our ETI/BVM subgroup analysis in which some subgroups may have had many subjects and others many fewer, we would, in general, need to apply a stronger correction to those subgroups with fewer subjects (i.e., less information). The more information available in any particular subgroup, the less likely it is that the original estimate has deviated substantially from the true value due to chance. Conversely, groups with little information may be expected to exhibit wider fluctuations around their true underlying values due to chance.

The concept of shrinkage is also known as "borrowing strength" because information is "borrowed" from all the other individuals or subgroups to help form the estimate for any particular individual or subgroup.

Borrowing Strength

We may be able to better understand the what is happening in borrowing strength by adopting a Bayesian,[6] multilevel modeling perspective. (Though there are important differences between the original James-Stein method and the Bayesian one, we chose this approach because it helps us develop a more intuitive understanding of the material; other approaches exist as well.) We know that at the end of the season each baseball player will have a particular batting average which we will assume (as Efron and Morris did) is the true value of interest. However, while we are still collecting data, the batting averages will continue to fluctuate: If a particular player is on a hot streak, his average may shoot higher than his end-of-season batting average.

On the other hand, if he is facing a series of top-notch pitchers, his average thus far may underestimate his eventual season performance. This is the first level of the model—the individual players.

The lower half of Figure 1 shows the first level of this model for three players: one player (the center distribution) is fairly average in his batting ability, one (on the right) is above average, and one (on the left) is below average. We have indicated the players' true averages with the lettered ticks (a–c). The data collected from each player at the beginning of the season give us one observed point (i.e., one observed batting average); these are labeled in the graphs with the filled shapes. Note that the distribution itself shows us the probability of observing a particular batting average early in the season given the underlying true batting average. Remember, though, that the distributions are constructed around points (a–c) that we cannot observe or have not yet observed. This should seem very similar to flipping three strange coins, where a–c would represent the unknown (true) probability of heads for a particular coin, and a filled shape would indicate the percentage of heads in the first ten tosses. The widths of the distributions are the same indicating that we have a roughly equal amount of information available for each player.

The upper half of Figure 1 shows the second level of the model. This distribution describes, collectively, all eighteen batters of interest to us. It is centered around the average ability of the group, and its width corresponds to how variable the batting ability is within the group. This distribution is thus made up of a bunch of (relatively) fair individual batters, as well as some (relatively) excellent and poor ones.

A link exists between the first and second levels because the observed batting averages for players in the first level are the points that we have available to us for defining the group distribution in the second level. While there exists a theoretical distribution which appropriately captures the mean and variance of the players' abilities (which is drawn in the second level of Figure 1), we cannot directly observe the parameters that define that distribution. Instead, we have only the observed batting averages available to us. The question then becomes, how do we best determine the parameters (i.e., mean and variance) of the group (level 2) distribution?

We may, for instance, be willing to make a rough assumption about the shape of the group performance distribution (e.g., that it is bell shaped), but we may not feel that we can make an educated guess about where that distribution should be centered (its mean) or how wide it should be (its variance). In order to learn about our group distribution, we find the values for the mean and variance that best explain the entire collection of observed batting averages. At the beginning of the baseball season, though we have only some of the information that will eventually be available to us, we can still find the best parameters for the group distribution. The specific approach used to pick the group mean and variance depends on the statistical tack being used. When conducting an empirical Bayesian analysis,[7] the observed group mean and variance are used to define the theoretical level 2 distributions while, in a fully Bayesian analysis, prior information and Bayes theorem are used. With the James-Stein approach (which is not Bayesian), we do not explicitly consider the group distribution; rather we derive a shrinkage factor based on the individual values, the overall average, and the overall variance.

Now that we have constructed our first and second levels and shown how they are linked, we can describe conceptually how shrinkage works. To correct a particular player's batting average for early season streaks, we first take into account the data generated thus far by that player. We then look at where that player appears to be in terms of the group distribution. Since it is unlikely that we would find a player many standard deviations away from the mean, we suspect that a player whose early data are quite extraordinary is on a streak that is not entirely representative of where he will end up, especially if he has had relatively few at-bats so far.

The closer the data are to the center of our group distribution, the less suspicious we are. The correction comes from a weighted average of the player's individual data and the group mean, with the weights determined by how narrow we feel the group distribution appears to be: The narrower the group distribution appears to be, the more we will pull the data points toward the center, because players have to be less far from the mean to be considered exceptional.

Considered from a slightly different perspective, the shrinkage correction occurs because we believe that all the players' means come from a single distribution (the one at level 2). The information we gain about 17 players gives us some insight into how the 18th will perform—it is unlikely, though possible, that his performance will be considerably different from theirs. If, for instance, the pack of 17 players seems to be performing particularly well, then we are more willing to believe (unless we see sufficient contrary data) that the 18th player is also performing well. And how well the group is performing is what determines the location (mean) of the level 2 distribution.

As mentioned earlier, we add an additional level of complexity when we recognize that the individuals' distributions also have associated variances or uncertainties. This is easier to comprehend in terms of information—the more information available about a single individual, the smaller the uncertainty or variance and the narrower the distribution. How far we shrink each player's individual average toward the group mean should also take into account these information (or inverse variance) weights. If we have a lot of information available for a particular individual, we have less reason to worry that his exceptional performance (good or bad) represents primarily a random fluctuation; we would therefore not want to push his estimated batting average as much toward the center of the group distribution. In other words, the batting average in a series of many at-bats (as opposed to few at-bats) would more likely approach the true batting average for that player, reducing the need for correction toward the group mean. Similarly, when we calculate the group mean, we should weight more heavily the averages of players with more at-bats.

The technique we described above requires inferring the parameters for the group distribution using only information from the individuals who comprise it, an approach known as empirical Bayes estimation.[7] In a fully Bayesian approach, we would make a prior educated guess about the group (level 2) distribution *before* considering the accumulating players' data, and describe our prior beliefs by introducing a third level to Figure 1 from which the parameters for the level 2 distribution would be sampled. Whether empirical or fully Bayes, the players' individual data can be thought of as being pulled toward the center; the only difference is whether that center is determined using the data alone (empirical Bayes) or using both the data and prior beliefs (fully Bayes).[8] Further, the closer the center is to the true center (whether derived empirically or with prior information), the more efficient these methods will be in improving the estimates.

With James-Stein estimation, if we are considering three or more independent level 1 estimates, the expected total mean squared error (across all the estimates) is guaranteed to be lower than if we were to use the naïve estimates. With four or more level 1 estimates (and the assumption of exchangeability in place of the stronger independence assumption), this expected error is reduced further using empirical Bayes estimation. (Although counterintuitive, the James-Stein method is valid with any independent groups, so that estimating the cost of various teas in China may enjoy reduced error if the batting averages are considered in the same problem. Of course, the more different these groups are, the less efficient the reduction in expected error.)

All of the techniques mentioned thus far, whether James-Stein or Bayesian, demonstrate the principle of bias-variance tradeoff.[8] When we shift the original estimates toward a specific point to reduce the squared error, we are introducing bias in the sense that we can no longer

expect that our estimates, on average (i.e., across multiple repetitions of the experiment), are equally likely to fall on either side of the true value. While this may seem statistically disconcerting, from a clinical perspective the tradeoff is generally worth it to achieve more accurate estimates. To reiterate an earlier point, we choose to use shrinkage not because it guarantees better results for estimating any particular baseball player's true average (it doesn't), but rather because as a strategy it is likely to yield more correct estimates overall than if we were to use the naïve (non-shrunk) estimates.

Our Approach to the Pediatric Airway Study

[For further discussion, see the Technical Appendix.]

We used empirical Bayes estimation, as described above, to determine improved estimates for the effect of ETI versus BVM on survival in each of the patient subgroups in the airway management study. The batters' averages discussed above have been replaced with the odds ratios for survival among the children in a particular subgroup, comparing the ETI arm to the BVM arm. The better ETI is relative to BVM, the higher the odds ratio will be, allowing us to represent the relative effect of ETI versus BVM with a single variable. And just as we had group and individual batting averages in the previous example, here we have odds ratios of survival for both the group as a whole and for the individual subgroups. (Note that each individual baseball player has an average computed from multiple at-bats, just as each subgroup's odds ratio is computed from multiple patient outcomes.)

Following the empirical Bayes approach, we first calculated the mean and variance of the group (level 2) distribution using the subgroup (level 1) data. Using this group distribution, we corrected the local estimates (i.e., those estimates derived using only the data available in a particular subgroup) using the approach mentioned earlier. By specifying the group distribution using all of the subgroup data, we allow the individual subgroup estimates to be shrunk toward the center of the group distribution by an amount justified by the consistency in the subgroup data. The degree to which a particular subgroup's estimated odds ratio is shrunk toward the overall mean depends on how far that odds ratio is from the mean, and how much information is available from that subgroup's data relative to how much information is available when considering all the subgroups' data.

Results

Using the standard (non-shrinkage) approach, the overall odds ratio for a successful outcome in the ETI- versus BVM-treated subjects is 0.82 (95% confidence interval, 0.61 – 1.11). The point estimate of the odds ratio and the confidence interval suggest that survival is likely better in the BVM arm of the trial, though it is not statistically significant at the 95% level. At this confidence level, the data are also consistent with no effect and even a harmful effect of BVI vs ETI. The single odds ratio toward which the individual odds ratios are shrunk in the empirical Bayes analysis is analogous to a weighted average of the subgroup odds ratios; this odds ratio is 0.77.

In Table 1 we have provided the raw data, the point estimates of the odds ratios and associated confidence intervals based on using each subgroup's data individually, and the same estimates and credible intervals after incorporating shrinkage. A credible interval is the Bayesian analog of the frequentist confidence interval. Details of the calculations leading to these results are provided in the technical appendix.

The degree of shrinkage for each subgroup's estimate of treatment effect is illustrated in Figure 2. When the odds ratios derived from the raw subgroup data are far from 0.77, or when they are based on less underlying data (i.e., when there is greater uncertainty), then the resulting

shrinkage is larger. The Other category exhibits considerable shrinkage because its raw estimate of treatment effect is both far from the overall treatment effect and less precise (there are only 26 patients in this category). If we carefully compare the Head Injury (HI) and Multiple Trauma (MT) subgroups, we see that despite their starting at approximately the same point, HI is shrunk slightly more because of its relatively greater uncertainty—HI includes 39 patients whereas MT includes 76. In other words, because three is less available information, the HI estimate appropriately borrows more information from the other subgroups than the MT estimate. The SIDS category represents an anomaly in that it seems to change in the wrong direction. This resulted from our taking into account the fact that there were no survivors in either ETI or BVM, which would lead to a divide-by-zero error in the calculation of the initial point estimate for the odds ratio. To circumvent this problem, we used the common approach of adding 0.5 "subjects" to all four cells (i.e., ETI recipients who lived and died, and BVM recipients who lived and died).[9]

Discussion

When we first examined the observed subgroup odds ratios from the pediatric airway management study, we noted that patients with Other causes of respiratory compromise might have considerable benefit from prehospital ETI, though the estimate is rather imprecise. However, after we borrow strength from the rest of the data, we quickly see that this benefit is very likely exaggerated. Credible intervals computed similarly (i.e., which incorporate appropriate shrinkage) further help the reader determine whether he or she feels that there is sufficient information to declare intubation superior in particular subgroups.

In Figure 2, we see that there is a natural ordering (or ranking) of the odds ratios before shrinkage. As an example, we see that Head Injury has a higher associated odds ratio for survival than does Submersion Injury. While a ranking also exists after shrinkage, the order may change considerably from that observed before shrinkage due to the variable application of shrinkage to the individual subgroups. (As noted above, this arises because of differing distances from the center and information available at each subgroup.) After shrinkage, the odds ratio for survival in the Head Injury subgroup is in fact lower than that in the Submersion Injury subgroup. We can expect the ordering of the estimates after shrinkage to be more accurate. (See reference 10 for a real-world application.)

As mentioned earlier, the overall expected error is lower using an empirical Bayes technique, making shrinkage the preferred strategy even if we cannot know if any particular estimate was corrected appropriately. Thus, we can expect that the odds ratios derived from shrinkage will be better than simply taking the independently calculated ratios when four or more subgroups are available. The shrinkage methods also tend to reduce type I and II error rates, though the specifics depend on the setting.

For this empirical Bayes approach to be valid, it is important that we not believe *a priori* that the treatment effect (here, the odds ratio) in one group is likely to be larger than in any other group. More specifically, before the study begins we should believe (for instance) with equal probability the following two statements: 1. the treatment effect is higher in the Multiple Trauma subgroup than in the Seizure subgroup, and 2. the treatment effect is higher in the Seizure subgroup than in the Multiple Trauma subgroup. Technically, we say that those two subgroups are exchangeable in that we could exchange 'Seizure' for 'Multiple Trauma' without changing our *a priori* belief that the statement is true. With exchangeability and borrowing strength, we are able to learn something about the treatment effect in the Seizure group by observing the Multiple Trauma patients, and vice-versa, which does indeed make intuitive sense.

On the other hand, if we feel (for instance) that ETI would be better than BVM in seizure patients but less good or even worse than BVM in multiple trauma patients, then we cannot consider the subgroups exchangeable. (Note that here we are referring to the relative treatment effects—the odds ratios—being exchangeable from subgroup to subgroup, not the survival proportions. We would not expect the survival proportions to be similar across subgroups.) If this is the case, then the parameters for the subgroup (level 1) distributions cannot be assumed to have come from the same group (level 2) distribution, thus unlinking the subgroup distributions (which had been linked via the group distribution) and disrupting our borrowing. There are methods to deal with this, but they are beyond this discussion. Importantly, however, even moderate violations of this exchangeability assumption will not violate the general result.[8]

Shrinkage techniques, of course, should not be used indiscriminately or be seen as a panacea. Biases or other flaws in the design of a study will not be corrected with these techniques. Further, the additional complexity required to implement shrinkage may not be warranted if there are a lot of data available for each subgroup, minimizing the impact of shrinkage.

Conclusion

In any analysis that involves multiple subgroups, reporting the observed individual subgroup estimates of treatment effects is likely to lead to a predictable increase in the overall error of the estimates when compared to an approach which borrows strength across subgroups. Researchers who report subgroup performance are strongly encouraged to consider this concept in their analyses.

Acknowledgments

The authors thank Howard A. Bessen, MD, for his valuable feedback after reviewing an early manuscript, and the anonymous reviewers for their insightful comments.

Support: This publication was made possible by Grant Number 1F32RR022167 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. This work was previously supported by a Research Fellowship Grant from the Emergency Medicine Foundation to AML. The conduct and data analysis of the Pediatric Airway Management Project was supported by grant HS-09065-01 from the Agency for Healthcare Research and Quality (AHRQ), grant EMS-3036 from the State of California Emergency Medical Services Authority, and grant MCH064004-01-0 from the Bureau for Maternal and Child Health of the Public Health Service.

References

1. Gausche M, Lewis RJ, Stratton SJ, et al. Effect of out-of-hospital pediatric endotracheal intubation-effect on survival and neurologic outcome: a controlled clinical trial. JAMA 2000;283:783–790. [PubMed: 10683058]
2. James, W.; Stein, C. Estimation with quadratic loss. In: Neyman, J., editor. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability; Berkeley: University of California Press; 1961. p. 311-319.
3. Efron B, Morris C. Stein's paradox in statistics. Sci Amer 1977;236:119–127.
4. Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 1975;70:311–319.
5. De Finetti, B. The Theory of Probability. Vol. Vols. 1 and 2. New York: Wiley; 1974.
6. Spiegelhalter DJ, Myles JP, Jones DR, et al. An introduction to Bayesian methods in health technology assessment. BMJ 1999;319:508–512. [PubMed: 10454409]
7. Casella G. An introduction to empirical Bayes data analysis. Am Stat 1985;39:83–87.
8. Greenland S. Principles of multilevel modelling. Int J Epidemiol 2000;29:158–167. [PubMed: 10750618]

9. Agresti A. On logit confidence intervals for the odds ratio with small samples. Biometrics 1999;55:597–602. [PubMed: 11318220]

10. Shahian DM, Torchiana DF, Shemin RJ, et al. Massachusetts cardiac surgery report card: implications of statistical methodology. Ann Thorac Surg 2005;80:2106–2113. [PubMed: 16305853]

11. Lunn DJ, Thomas A, Best N, et al. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000;10:325–337.

Appendix

Technical Appendix

The treatment effect for the pediatric airway study is the log odds ratio of survival with ETI as compared to BVM ventilation. Because we are comparing two effects in each group (i.e., the control and the treatment), we actually need to consider distributions of both the control effect and the treatment effect. We call *pc* the probability of survival in the control arm, and *pt* the probability of survival in the treatment arm. The log odds ratio, θ, is defined as

$$\theta = \ln\left[\frac{pc\,(1-pt)}{pt\,(1-pc)}\right].$$

Because of this relationship, the two effects are completely described by knowing *pc* and θ.

We chose to use normal distributions for the logit of *pc* and for θ The means and variances that define these normal distributions are known as hyperparameters because they determine the distributions from which the parameters are drawn that define the local distributions.

In a fully Bayesian hierarchal model, we would define prior probability distributions on all of the hyperparameters. These four hyperparameters define: 1. an “average” rate of survival proportion with the control treatment (BVM) across subgroups (*pc*); 2. the variability in that survival proportion; 3. the “average” treatment effect associated with the use of ETI as compared to bag-valve-mask ventilation (θ); and 4. the variability in that treatment effect across subgroups.

As discussed earlier, empirical Bayes methods do not require that we specify the prior distributions for these hyperparameters. Instead, we replace them with fixed values which convey the weighted overall means and variances. Here, we have used a quasi-maximum-likelihood approach to provide those values.

All Bayesian calculations were conducted using WinBUGS, Version 1.4, a publicly-available program which can be used to determine posterior probability densities in Bayesian models.[11] Initially, a fully Bayesian model was defined but with diffuse, non-informative priors for the four hyperparameters. The resulting values from this model provided the quasi-maximum-likelihood estimates of the four hyperparameters, which were then used as the parameters for the group distributions in the empirical Bayes model. Our revised estimates of the treatment effects across the subgroups are taken from the empirical model.

The mean of the overall distribution for θ, calculated during the first WinBUGS run, was −0.2666. Taking the exponential of this number, we derive the overall odds ratio toward which our subgroups should be shrunk, 0.77.

WinBUGS Code

model EBSubgrp;

```
# We first define the number of subgroups so we can build variable
# vectors.

const Num;

# Next we define the variables. The ones with "[Num]" are vectors, so
# that, for instance, st[4] refers to the number of successes in the
# fourth treatment subgroup, in this case Head Injury.

# Note that in WinBUGS the inverse of the variance is usually used to
# specify distributions; this variable is commonly referred to as tau.

# Though all the different subgroups have their own log odds for
# survival (lopc[]), the means and inverse variances are all defined by
# shared distributions (mu_pc, tau_pc). This is what links together the
# individual subgroups (see Figure 1). Similarly for the log odds
# ratios (theta, mu_theta, tau_theta).

var
  st[Num], nt[Num], # Successes and total subjects in treatment arms
  sc[Num], nc[Num], # Successes and total subjects in control arms
  pc[Num], pt[Num], # Probabilities of survival for control and
                    # treatment

  lopc[Num],        # Log odds of survival in control groups
  mu_pc,            # A distribution of means from which is sampled the mean
                    # for the distribution that represents the log odds of
                    # survival in each control group
  tau_pc,           # A distribution of inverse variances from which is
                    # sampled the inverse variance for the distribution that
                    # represents the log odds of survival in each control
                    # group

  theta[Num],       # Log odds ratio of survival in treatment vs
                    # control groups
  odds[Num],        # Exp(theta[]) (i.e., the odds ratios)
  mu_theta,         # A distribution of means from which is sampled the mean
                    # for the distribution that represents the log odds ratio
  tau_theta;        # A distribution of inverse variances from which is
                    # sampled the inverse variance for the distribution that
                    # represents the log odds ratio

# We now code our likelihoods and some parameter definitions.
# We have a binomial distribution defining the likelihood of seeing sc
# survivors of nc subjects given a probability of pc. Similarly with
# the treatment group.
# We then define the log odds of pc as lopc, and the log odds of pt as
# (lopc + theta), thus making theta the log odds ratio.
# We next have another likelihood that relates the probability of
# seeing a particular lopc given a normal distribution with mu_pc and
```

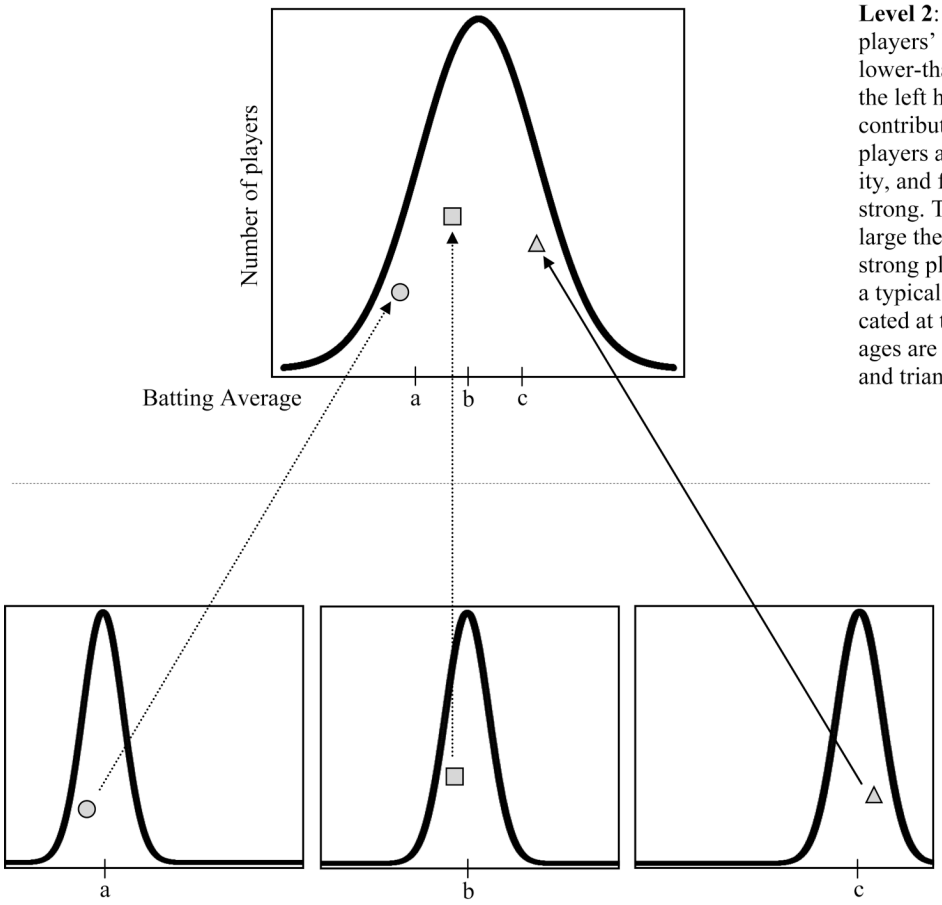
```
# tau_pc as its parameters, and similarly with theta.

{
  for (i in 1:Num)
       {
       sc[i]~dbin(pc[i], nc[i]);
       st[i]~dbin(pt[i], nt[i]);
       logit(pc[i])<- lopc[i];
       logit(pt[i])<- lopc[i] + theta[i];
       lopc[i]~dnorm(mu_pc, tau_pc);
       theta[i]~dnorm(mu_theta, tau_theta);
       odds[i]<-exp(theta[i]);
       }

# The following four lines define the prior information for the
# last two distributions mentioned above; these contain the
# so-called hyperparameters.

mu_pc~dnorm(0.0, 1.0E-6);
tau_pc~dgamma(0.001, 0.001);

mu_theta~dnorm(0.0, 1.0E-6);
tau_theta~dgamma(0.001, 0.001);
}

# Enumerate the observed data.

list(Num = 11,
sc = c(12, 9, 9, 4, 5, 5, 17, 34, 0, 18, 10),
nc = c(130, 22, 13, 17, 31, 10, 17, 36, 59, 54, 15),
st = c(14, 1, 5, 8, 12, 3, 14, 25, 0, 18, 10),
nt = c(124, 22, 13, 22, 45, 10, 19, 29, 80, 41, 11))

# Initialize the data to help the MCMC algorithm find a good place to
# start.

list(lopc = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
    theta = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
    mu_pc = 0, tau_pc = 0.1,
    mu_theta = 0, tau_theta = 0.1)
```

To generate the EB estimates of the log odds ratios, the last four lines of the distribution definitions and the last two lines of the initialization list were removed (i.e., those lines describing the hyperparameters), and the lopc [i] and theta[i] lines were replaced with (using the medians of the values computed in the first step):

```
        lopc[i]~dnorm(−0.4596, 0.1841);
        theta[i]~dnorm(−0.2666, 1.662);
```

The reader’s results should be fairly close to those shown here, with small

discrepancies expected due to, among other things, the specified sampling parameters such as the number of updates.



Level 2: This distribution represents a group of players' baseball ability. Weaker players, with lower-than-average batting averages, populate the left half of the curve, while stronger players contribute to the right half of the curve. Most players are close to the middle or average ability, and few players are very weak or very strong. The width of this curve depends on how large the spread is between very weak and very strong players. The true averages of a weak (a), a typical (b), and a strong (c) player are indicated at the bottom. The players' observed averages are indicated with a shaded circle, square, and triangle, respectively.

Level 1: Each of these three distributions represents one particular player. A sampled point from each distribution contributes to one point in the group distribution above. Weaker players have lower means, and stronger players have higher means. The width of the curve depends on how widely the player's ability fluctuates around his "true" average. The shapes indicate the observed values of the players' batting averages, and the ticks at the bottom indicate their true averages.

Figure 1.

Two-level, Hierarchical Model of Individual Baseball Players (Level 1) and the Players as a Group (Level 2). Note that in keeping with standard notation, we have labeled the individual level '1' and the group level '2.' If the parameters for the group 2 distribution were sampled from a distribution above it, that higher distribution would be labeled '3.'

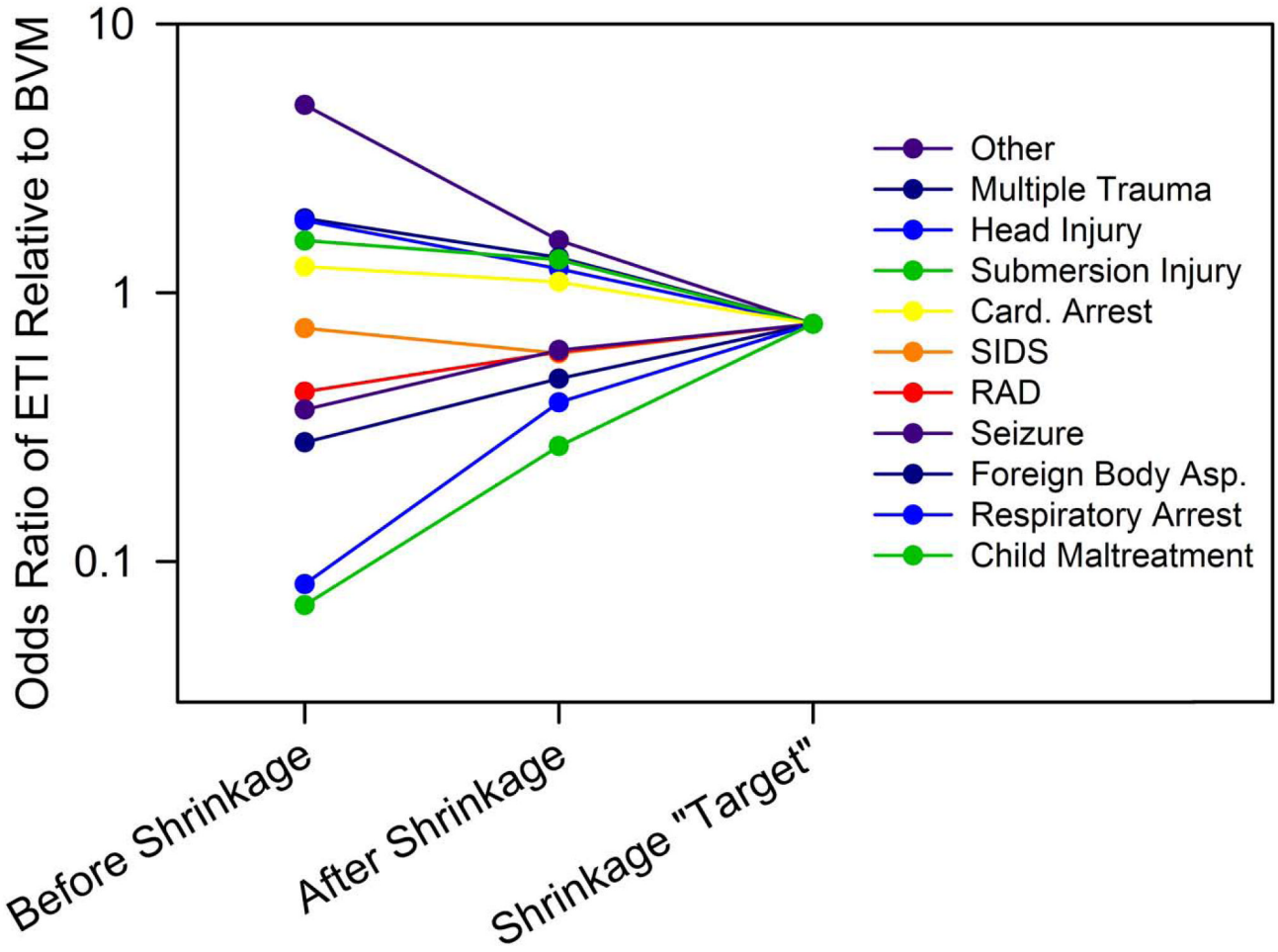


Figure 2.
Application of Shrinkage to Odds Ratios by Subgroup. The *y*-axis has a log scale. ETI – endotracheal intubation; BVM – bag-valve-mask; Card – cardiopulmonary; SIDS – sudden infant death syndrome; RAD – reactive airway disease; Asp – aspiration.

Table 1

Outcomes by Patient Subgroup, Before and After Shrinkage Applied

Treatment	ETI Survival	BVM Survival	Odds Ratios	
Subgroup[a]			Before Shrinkage (95% CI)	After Shrinkage (95% CrI)
Cardiopulmonary Arrest	14/124 (11%)	12/130 (9%)	1.25 (0.55, 2.82)	1.10 (0.53, 2.25)
Child Maltreatment	1/22 (5%)	9/22 (41%)	0.07 (0.01, 0.61)	0.27 (0.09, 0.78)
Foreign Body Aspiration	5/13 (38%)	9/13 (69%)	0.28 (0.05, 1.41)	0.48 (0.16, 1.42)
Head Injury	8/22 (36%)	4/17 (24%)	1.86 (0.45, 7.67)	1.23 (0.44, 3.47)
Multiple Trauma	12/45 (27%)	5/31 (16%)	1.89 (0.59, 6.05)	1.35 (0.55, 3.38)
Reactive Airway Disease	3/10 (30%)	5/10 (50%)	0.43 (0.07, 2.68)	0.60 (0.18, 1.93)
Respiratory Arrest[b]	14/19 (74%)	17/17 (100%)	0.08 (0.00, 1.48)	0.39 (0.12, 1.25)
Seizure	25/29 (86%)	34/36 (94%)	0.37 (0.06, 2.17)	0.61 (0.20, 1.87)
SIDS[b]	0/80 (0%)	0/59 (0%)	0.74 (0.01, 37.8)	0.60 (0.14, 2.45)
Submersion Injury	18/41 (44%)	18/54 (33%)	1.57 (0.68, 3.61)	1.33 (0.64, 2.77)
Other	10/11 (91%)	10/15 (67%)	5.00 (0.49, 50.8)	1.57 (0.48, 5.21)
Total[c]	110/416 (26%)	123/404 (30%)	0.82 (0.61, 1.11)	0.77 (0.30, 1.50)

ETI – endotracheal intubation; BVM – bag valve mask; CI – confidence interval; CrI – credible interval; SIDS –sudden infant death syndrome.

[a] In order to create mutually exclusive subgroups, we have made minor changes to the numbers presented here, as compared to the Table that appeared in the original publication.

[b] Subgroups that had treatments with no survivors or only survivors had 0.5 added to all of the underlying cells. For Respiratory Arrest, the odds ratio is thus (14.5×0.5)/(5.5×17.5) because there were no deaths in the BVM arm, and for SIDS it is (0.5×59.5)/(0.5×80.5) because there were no survivors in either arm.

[c] The “Before Shrinkage” odds ratio was calculated assuming that all the data were in one large group. In the calculation of the “After Shrinkage” odds ratio, the clustered nature of the data (i.e., that it is being analyzed by subgroup in our empirical Bayes approach) was preserved. The increased width of the “After Shrinkage” versus the “Before Shrinkage” interval is a direct consequence of using the subgroup model, as variability arises from differences among subjects both within and between their respective subgroups.