Borrowing Strength When Explicit Data Pooling Is Prohibited

Jerome P. Reiter[1]

When using regression models where units can be classified into distinct groups, similar parameters in each group can be estimated via explicit data pooling, such as in hierarchical models. Sometimes, however, external constraints prohibit explicit data pooling. In this article, I propose techniques that may be acceptable under such external constraints and yield more accurate estimates than those obtained by regressing separately in each group. These techniques utilize the information in multiple groups' parameter estimates to specify the model in each group, but ultimately estimate the parameters selected for each group's model using only that group's data. The techniques can be conceptualized as existing on a continuum ordered by how directly each relies on data pooling to make estimates; those techniques that look more like explicit data pooling are typically more accurate yet less likely to be acceptable. I present several methods for evaluating the procedure empirically.

Key words: Information pooling; legal constraints; hierarchical modeling; model selection.

1. Introduction

When units are classified as members of distinct groups, the goal of a statistical analysis is often to estimate distinct but similar parameters in each group. For example, given samples of patients from different hospitals, health statisticians may seek a model in each hospital that accurately predicts the probability of a patient's recovery given various risk factors for the patient. Or, given samples of households from different states, a national statistical agency may seek estimates of the population shares of many sub-populations (e.g., male black children, female white adults) in each state. In such tasks, one approach is to estimate the parameters in each group using just the data from that group: the health statisticians might build separate logistic regression models in each hospital, and the statistical agency might use standard survey estimators in each state. However, this segregated approach can lead to estimates with relatively large variances, particularly if sample sizes in some of the groups are not large.

Another approach is to estimate these parameters by using models that explicitly pool the data from the different groups. By *explicit data pooling*, I mean any model in which

multiple groups' data enter directly into the formulas used to estimate any of the para-
meters ultimately included in each group's model. As examples, for each hospital i, the
health statisticians might estimate the regression coefficients of the risk factors, β_i, by
using one logistic regression model that posits $\beta_i = \beta$ for all i; and, the statistical agency
might smooth the survey estimates of population shares across states with a hierarchical
regression model (e.g., Ghosh and Rao 1994). Estimates obtained from explicit data
pooling generally have smaller variances than estimates obtained from fitting models
separately in each group, although they are accompanied with an introduction of bias.
When explicit data pooling reduces variance by more than it increases squared bias, the
mean squared errors of the estimates are decreased.

Sometimes, however, statisticians must operate under external constraints that prohibit
explicit data pooling. For simplicity, I refer to these constraints as *legal constraints*. Such
legal constraints can arise from a mandate that each group's estimates ultimately be deter-
mined from only that group's data, or they can result from the concern that some of the
groups will not accept the estimates if they have been derived from explicit data pooling.
Naturally, the statisticians can respect the constraints by fitting models separately in each
group without any consideration of the other groups, but this forfeits the potential gains
from using across-group information. Thus, the statisticians are faced with a dilemma:
how can the groups' estimates be improved while respecting the constraints?

This article describes techniques that serve as potential solutions to this constrained
estimation dilemma. In Section 2, I describe the basic idea underpinning these techniques:
use information obtained from explicit data pooling to help specify models in each group,
but estimate the parameters ultimately included in each group's model using just that
group's data. This may satisfy the legal constraints because data pooling is used only
for model selection and not for parameter estimation, and it may increase accuracy
because it takes advantage of across-group information. In Section 3, I present examples
of settings where this approach could be useful. In Section 4, I describe several procedures
that utilize across-group information to specify separate, normal linear regression models.
In Section 5, I suggest and illustrate methods for evaluating these procedures empirically.
In Section 6, I provide additional comments on utilizing across-group information when
smoothing survey estimates. Finally, in Section 7, I end with some general comments.

2. Medial Information Pooling

Explicit data pooling is one set of techniques in a more general class of approaches to
increasing estimation accuracy that I call *information pooling*. I define information pool-
ing to be using both a group's data and knowledge not contained in that group's data
to make estimates in that group. This is a broad definition, and nearly every statistical
analysis employs some form of information pooling. For example, an essential form of
information pooling is relying on past experience to design the data collection mechan-
ism and to build the statistical models for an estimation task.

Conceptually, information pooling techniques exist on a continuum ordered by how
directly each relies on multiple groups' data to make estimates. At one extreme of the
continuum is solely using past experience, which I call *minimal information pooling*. At
the other extreme of the continuum is the use of explicit data pooling, which I call *maximal*

information pooling. In between these two extremes is a host of information pooling techniques that use multiple groups' data somewhat indirectly, such as a strategy that utilizes information from multiple groups' parameter estimates to specify the model in each group. Since such a model specification strategy relies on estimates of parameters from multiple groups to specify each group's model, it uses multiple groups' data more directly than does solely using past experience. Since it estimates the parameters included in each group's model from just the data in that group, it does not use multiple groups' data as directly as explicit data pooling. I refer to techniques that lie in between minimal and maximal information pooling as *medial information pooling.*

It seems clear that a prohibition of *all* information pooling strategies would severely restrict, if not eliminate, the ability of the statisticians to provide accurate estimates. However, just because one form of information pooling is unacceptable under certain constraints, namely explicit data pooling, it is not necessarily the case that all forms of information pooling are unacceptable. Thus, in a constrained estimation setting, the pertinent question the statisticians should consider is not *whether* information pooling is permissible; rather, it is, ''*how much* information pooling is permissible?'' In answering this question, the statisticians need to determine how far along the information pooling continuum they are allowed to travel. Undoubtedly, as the techniques move towards maximal information pooling, they are more likely to be perceived as too similar to explicit data pooling, and hence less likely to be acceptable to those who disapprove of explicit data pooling.

If there is little reduction in estimation errors when a potentially controversial form of information pooling is employed, then it is not worthwhile to argue for the ''legality'' of that strategy. Thus, a second pertinent question the statisticians must answer is, ''does using an information pooling strategy reduce estimation errors by a sufficient amount?'' Implicit in the answers to this question is a tradeoff between accuracy and acceptability. We expect the strategies that rely more directly on multiple groups' data to produce estimates with smaller mean squared errors, yet these strategies will be more controversial. Those strategies that look less like explicit data pooling will be easier to justify legally, but they will not give as large a payoff in estimation accuracy. Therefore, to find a strategy with a satisfactory balance between accuracy and acceptability, it may be necessary to consider strategies at many locations on the information pooling continuum.

3.   Settings Where Medial Information Pooling Could Be Useful

In this section, I describe settings where medial information pooling could help statisticians improve estimates without explicit data pooling.

3.1.   Potential constraints in the 2000 U.S. Census

Under the proposed design of the year 2000 U.S. population census that includes the Integrated Coverage Measurement (ICM) survey, the U.S. Census Bureau uses sampling to adjust enumerated counts for undercoverage and overcoverage in each state (United States Bureau of the Census 1997). Direct estimates of such coverage factors for sub-state demographic and geographic domains (i.e., estimates made without combining data across domains) are likely to be highly variable (Griffin and Vacca 1998). To reduce the average

mean squared errors of the estimates of these coverage factors, a desirable approach is to borrow strength across states by using a hierarchical regression model that smoothes the direct estimates across states.

However, the U.S. Census Bureau has expressed the desire to avoid explicitly pooling data across states when estimating coverage factors from the ICM survey (Mulry 1996; Thompson and Fay 1998). Based on experience with previous censuses, the Census Bureau believes that members of Congress and other census clients will not want the estimates of population counts in other states to factor in to the determination of population counts in their own state (Fay and Thompson 1993). Additionally, the Census Bureau is concerned that census clients might challenge the legality of explicitly pooling data across states (Thompson and Fay 1998). Nonetheless, there may be across-state information that, if tapped, could improve the accuracy of the within-state estimates. Medial information pooling procedures could help the Census Bureau extract such information and construct separate smoothing models in each state.

The legal constraints faced by the Census Bureau are generated by the use of sampling for the purpose of apportioning seats in the House of Representatives. In January 1999, the Supreme Court ruled that such sampling is illegal under the Census Act. Consequently, the Census Bureau redesigned the census and eliminated the ICM survey (United States Bureau of the Census 1999). As of this writing, it is not clear whether the above legal constraints exist for the redesigned plan.

3.2.   Potential constraints in audits

During the 1980's and early 1990's, the U.S. federal government audited states' welfare payments (e.g., AFDC, Medicaid, and Food Stamps payments) to determine if the states were overpaying benefits. The audits were conducted on samples of each state's welfare cases. If the federal auditors determined that the estimated average amount of overpayments in a state exceeded a threshold, the state was fined a multi-million dollar penalty by the federal government. The federal audits and resultant fines were extremely controversial, and many states challenged the penalties in the courts (Kramer 1990).

One approach to estimating overpayment rates, suggested by Fairley et al. (1990), is with a hierarchical model that pools data across different states and different time periods. This model was rejected by the federal auditors in part because the potential biases in some individual states were too large to be acceptable (Hansen and Tepping 1990). Instead, state-specific regression estimators were used (Hansen and Tepping 1990).

This statistical rationale for rejecting explicit data pooling is related to a political rationale: explicit data pooling might inject even more controversy into an already controversial program. Consider a scenario where a state is penalized if its estimate of the average amount of overpayments is computed via data pooling methods, but it is not penalized if this estimate is computed by separate estimation. Faced with millions of dollars in fines that appear only if its data are pooled with other states' data, the penalized state is likely to challenge the federal auditors' use of explicit data pooling.

It seems clear that the federal auditors want to avoid the additional controversy that comes with explicit data pooling. But, they might be willing to use less controversial forms of information pooling in hopes of improving estimates. For example, if the federal

auditors consider regression estimators that potentially include measures of case difficulty and strata for monthly effects, medial information pooling could help the auditors specify the form of these models.

This auditing example is representative of a general setting where constraints on explicit data pooling may exist, namely when statistical analyses are used to compare groups' performances against a standard or against each other. If some groups believe that explicit data pooling may shrink their estimates in a way that makes them look worse relative to separate analyses, these groups may challenge the use of explicit data pooling. Settings where this concern could be relevant include assessments of schools' educational achievements and evaluations of hospitals' performances.

3.3.   Potential constraints on use of data from prior time periods

Multiple groups also can be defined as different time periods, for example separate years of data from a repeated survey. To improve predictions for current or future time periods, statistical agencies may consider explicitly pooling data across time periods (Bell and Hillmer 1990). However, some agencies may be reluctant to do so because they do not want prior data to be used for current estimates. With medial information pooling, these agencies may be able to take advantage of similarities in the relationships among variables across years to help specify a model that is fit using only the current data.

3.4.   Other uses of medial information pooling

Medial information pooling also can be useful when explicit data pooling is legal but practically infeasible, such as in the following three settings.

3.4.1.   Statistical agency seeks transparent procedures
When working with data users whose statistical background is limited, statistical agencies may want to avoid complicated techniques if simpler techniques yield similar results. In a setting where a hierarchical model is potentially too complicated, a nonhierarchical model constructed with the aid of medial information pooling might be a sufficiently accurate and explainable substitute.

3.4.2.   National statistical agency advises local statistical agencies
In some settings, groups seek to construct their own models for an estimation task. A statistical agency with access to all the groups' data, and perhaps more statistical expertise, can help these groups specify their own individual models. Specifically, this agency could use the results of explicit data pooling to determine which of many covariates have useful predictive power, and then advise the groups to include these predictors when they fit their own models.

3.4.3.   Evaluation of pre-specified models
In some settings, models must be specified before the data are collected. For example, models used in the Post-Enumeration Survey of the 1990 U.S. census had to be specified before the survey was conducted (Hogan 1993). Medial information pooling procedures can be used to evaluate these pre-specified models, for example to determine the

maximum amount of information that could have been extracted from multiple groups' data without explicit data pooling. This idea is related to Zaslavsky's (1993) suggestion to use results from a hierarchical model to check the accuracy of competing estimators of population shares in the 1990 U.S. census.

4. Approaches to Medial Information Pooling

Medial information pooling can have utility in any setting where statisticians want to use a separate regression model in each group to predict a future outcome (e.g., predicting patient recoveries in a hospital), or to smooth many estimates (e.g., smoothing sub-population counts in a state), because they cannot explicitly pool data across the groups. In this section, I develop several medial information pooling procedures. For ease of explanation, my discussion of these procedures is based on prediction rather than smoothing settings. As discussed in Section 6, these procedures also are relevant for smoothing settings.

In each of $i = 1, \ldots, G$ groups, let $\mathcal{Y}_i$ be an unobserved, $m_i \times 1$ vector of dependent variables, and let $\mathbf{Y}_i$ be an observed, $n_i \times 1$ vector of the same dependent variables. Let $\mathcal{X}_i$ be an $m_i \times p$ matrix of the values of p predictors associated with $\mathcal{Y}_i$, and let $\mathbf{X}_i$ be an $n_i \times p$ matrix of the values of the same predictors associated with $\mathbf{Y}_i$. The p predictors in the $(\mathcal{X}_i, \mathbf{X}_i)$ are the same for all i, although the values of the variables generally differ across groups. As an example of this notation, consider the following situation. In each hospital i of $G = 100$ hospitals, consider predicting m_i current patients' recovery times, $\mathcal{Y}_i$, from $p = 4$ background variables, $\mathcal{X}_i$ (e.g., an intercept, age, sex, and some "healthiness" score), by using n_i previous patients' recovery times, $\mathbf{Y}_i$, and their values for the same four background variables, $\mathbf{X}_i$.

For each group, we assume that

$$\mathcal{Y}_i \sim N(\mathcal{X}_i \boldsymbol{\beta}_i, \Sigma_{\mathcal{Y},i}) \tag{1}$$

where β_i is a $p \times 1$ vector of regression coefficients and $\Sigma_{\mathcal{Y},i}$ is an $m_i \times m_i$ covariance matrix. Given some estimate of β_i, which we call $\tilde{\beta}_i$, we estimate $\mathcal{Y}_i$ as

$$\tilde{\mathcal{Y}}_i = \mathcal{X}_i \tilde{\boldsymbol{\beta}}_i \tag{2}$$

The goal of medial information pooling procedures is to use across-group information to help determine each group's $\tilde{\boldsymbol{\beta}}_i$ so that the $\tilde{\mathcal{Y}}_i$ are as accurate as possible, while respecting any legal constraints.

4.1. *Motivation for and description of strategies*

A typical method for determining the $\tilde{\boldsymbol{\beta}}_i$ of (2) separately in each group is to assume that

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \boldsymbol{\beta}_i, \Sigma_i) \tag{3}$$

where Σ_i is an $n_i \times n_i$ covariance matrix, and estimate $\boldsymbol{\beta}_i$ by weighted least-squares. The weighted least-squares estimate, labeled as $\tilde{\boldsymbol{\beta}}_i^{\mathsf{F}}$, where the superscript F indicates that the regression is fit with the full set of predictors in $\mathbf{X}_i$, is

$$\tilde{\boldsymbol{\beta}}_i^{\mathsf{F}} = (\mathbf{X}_i^t [\tilde{\Sigma}_i^{\mathsf{F}}]^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i^t [\tilde{\Sigma}_i^{\mathsf{F}}]^{-1} \mathbf{Y}_i \tag{4}$$

where $\tilde{\Sigma}_i^{\mathsf{F}}$ is an estimate of Σ_i. In typical applications of weighted least-squares, and in this

section, it is assumed that $\Sigma_i = \sigma_i \mathbf{C}_i$, where $\mathbf{C}_i$ is some known positive definite matrix and σ_i is an unknown scalar quantity.

Sometimes, because of sampling variability, the $\tilde{\boldsymbol{\beta}}_i^{\mathsf{F}}$ are far from the $\boldsymbol{\beta}_i$. When this occurs, the $\tilde{\boldsymbol{y}}_i^{\mathsf{F}} = \boldsymbol{X}_i \tilde{\boldsymbol{\beta}}_i^{\mathsf{F}}$ may inaccurately estimate the $\boldsymbol{y}_i$. To mitigate the effects of random variability, statisticians have developed alternative methods of estimating the $\boldsymbol{\beta}_i$, such as ridge regression (Hoerl and Kennard 1970), robust regression (see Chapter 7 of Huber 1981; Chapter 12 of Gelman et al. 1996a), and semi-automatic model selection procedures based on penalized likelihood functions (e.g., Akaike 1974; Schwarz 1978; Raftery 1995). In all these approaches, estimation of the $\boldsymbol{\beta}_i$ is based on only the information from group i.

When regressing in multiple groups, information from all the groups can be utilized to help estimate the β_i. Explicit data pooling does this, but we are assuming that this approach is prohibited by legal constraints and that other, less pooling-intensive approaches are needed. What kinds of medial strategies are located on the information pooling continuum? To address this question, it is helpful to construct a wish-list of information that, if known, might aid modelers to improve predictions. The goal of a medial information pooling strategy is to use multiple groups' data to make one of these wishes come true, or at least approximately true, while respecting the constraints.

The first item we might wish for is knowledge of which predictors' estimated coefficients can be set to zero to reduce the mean squared errors of the $\tilde{\boldsymbol{y}}_i$. I refer to such predictors as *unimportant predictors*. Identifying unimportant predictors is generally the goal of traditional model selection strategies, such as choosing the model in each group that minimizes the AIC (Akaike 1974) or the BIC (Schwarz 1978; Raftery 1995). Utilizing across-group information might allow us to identify these predictors more effectively. This idea is a medial information pooling strategy: use multiple groups' data to help determine which predictors should be removed from a group's model, and then estimate the coefficients ultimately included in each group's model using only that group's data. I call this strategy a *Predictor Exclusion* strategy.

If knowledge for the Predictor Exclusion strategy is not available, it would be helpful to know whether a traditional model selection procedure, such as a stepwise regression, improves the estimate of $\boldsymbol{y}_i$. Given this knowledge, we can use the traditional model selection procedure when it results in a more accurate estimate of $\boldsymbol{y}_i$ than the full model does, and we can avoid the traditional model selection procedure when it results in a less accurate estimate of $\boldsymbol{y}_i$ than the full model does. Since multiple groups' data can help us estimate coefficients more accurately, they may be able to help us gain this knowledge. This leads to a second medial information pooling strategy: use multiple groups' data to choose either the full model or the model produced by a traditional model selection procedure. I call this a *Stepwise/Full Selection* strategy.

Moving down the wish-list, another item is knowledge of the number of predictors in each group that make an important contribution to the predictive ability of the model. With knowledge of the number of important predictors, we can use standard techniques, like selecting the model with that number of predictors that maximizes R^2, in hopes of using those important predictors, and, in the process, excluding unimportant predictors. This idea forms a third medial information pooling strategy: use multiple groups' data to determine the number of predictors to include in each group's model, and then

determine those predictors and estimate their regression coefficients separately in each group. I call this a *Dimension Selection* strategy.

Another item on the wish-list is to know which units, if any, have data values that cause the $\tilde{\boldsymbol{\beta}}_i^{\mathsf{F}}$ to be far from the $\boldsymbol{\beta}_i$. When such influential units are identified, we can estimate $\boldsymbol{\beta}_i$ without these units. The resultant estimate of $\boldsymbol{\beta}_i$ is closer to the truth, and the resultant $\tilde{\boldsymbol{y}}_i$ is likely to estimate $\boldsymbol{y}_i$ more accurately. Because multiple groups' data can help us estimate the $\boldsymbol{\beta}_i$ more accurately, they also may help us identify influential units. This idea is a fourth medial information pooling strategy: use multiple groups' data to determine whether individual units' observations are distorting estimates of regression coefficients, and, if so, remove those units when fitting the separate models. I call this strategy a *Unit Exclusion* strategy.

Two other strategies include using across-group information: 1) to find appropriate polynomial functions of or interactions among predictors; and 2) to identify transformations of the dependent or independent variables. The former strategy can be viewed as a subset of the Predictor Exclusion strategy; that is, we simply treat the polynomial functions or interactions as we do other predictors and decide if they are unimportant. The latter strategy can be implemented by fitting and checking the explicit data pooling model: the best-fitting transformations in the pooled model are used in each group's separate model. To check the adequacy of models that explicitly pool data, I recommend the use of posterior predictive checks (Rubin 1984; Gelman, Meng, and Stern 1996b) and the techniques in Hodges (1998). I do not pursue these two strategies further in this article.

4.2.   Description of procedures

To create viable procedures that implement these four strategies, it is necessary to mine extra information that exists in the groups' data. As shown in many applications, good-fitting hierarchical models effectively take advantage of such information (e.g., Rubin 1980; Morris 1983). It makes sense, then, to use hierarchical models as tools for extracting across-group information. I call a procedure that uses a hierarchical model to mine information a *hierarchical model selection procedure* (HMSP). The HMSPs examined here use the following hierarchical normal regression model, labeled an HNRM, as a baseline for information extraction:

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i), \qquad \boldsymbol{\beta}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\beta) \tag{5}$$

The HMSPs are based on the maximum likelihood estimates of the parameters in (5), which we find via the EM algorithm (Dempster, Laird, and Rubin 1977a). The HNRM is used here because of its wide applicability and ease of simulation. In actual applications, to improve model fit, it is generally advisable to incorporate group-level covariates in the distribution of the $\boldsymbol{\beta}_i$ in (5). The procedures remain appropriate for such models. Reiter (1999) develops medial information pooling procedures that use nonhiearchical, explicit data pooling models as baselines for information extraction. Such procedures are not presented here.

Most medial information pooling procedures can be implemented in two distinct manners: *local* or *global* implementation. A local procedure makes decisions about model specification separately in each group, whereas a global procedure makes the

same decisions about model specification in all the groups. For the Predictor Exclusion strategy, local procedures select predictors separately in each group's model, whereas global procedures force all the groups' models to have the same predictors. For the Stepwise/Full Selection strategy, local procedures decide separately in each group whether or not to perform a traditional model selection procedure, whereas global procedures decide to perform a traditional model selection procedure in all the groups or none of the groups. For the Dimension Selection strategy, local procedures select the dimension of the model separately in each group, whereas global procedures force all the groups' models to have the same dimension. In both cases, the predictors are selected independently in each group. For the Unit Exclusion strategy, local procedures identify the units that should be dropped in each group. There are no global Unit Exclusion procedures, since ''using the same units in each group'' is, of course, not possible.

The decision to use local or global procedures is driven by the constraints in a particular setting. In some settings, global procedures may be more acceptable than local procedures. For example, in some settings, it may reduce controversy to include the same predictors in each group's model. In other settings, global procedures might be perceived as more reliant on multiple groups' data than local procedures, since the same operations are performed in each group. To provide users with local and global options, I present both local and global HMSPs.

Each procedure attempts to minimize some criterion over subsets of included predictors or units. To save computational effort, subsets are searched via a backward-forward stepwise procedure that starts from the full set of predictors or, in Unit Exclusion procedures, from the full set of units. Procedures' criteria also can be minimized by searching all subsets with the branching techniques of Furnival and Wilson (1974). To describe these criteria, we use the following notation:

- S denotes a set of predictors. An intercept term is always included in S. When S contains all the predictors under consideration for inclusion in a regression model, the set is denoted by F. When referring to the set of predictors in the regression model for group i, we use S_i.
- $\mathcal{K}_i$ denotes a set of units in group i. When referring to the set of units used to estimate β_i, we use K_i.
- $|\mathsf{S}|$ denotes the number of predictors in set S. $|\mathcal{K}_i|$ denotes the number of units in set $\mathcal{K}_i$.
- $\mathbf{X}_{i:\mathcal{K}_i}^{\mathsf{S}}$ is a $|\mathcal{K}_i| \times |\mathsf{S}|$ matrix of predictors, with rows corresponding to the subset of units in $\mathcal{K}_i$ and columns corresponding to the subset of predictors in S. To ease notation, when $\mathsf{S} = \mathsf{F}$ we write simply $\mathbf{X}_{i:\mathcal{K}_i}$, and when $\mathcal{K}_i = n_i$ we write $\mathbf{X}_i^{\mathsf{S}}$ or $\mathbf{X}_i$. In general, superscript letters reference subsets of predictors, and subscript letters reference subsets of units within groups.
- $\tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{S}} = \mathbf{X}_i^{\mathsf{S}} \tilde{\boldsymbol{\beta}}_{i:\mathcal{K}_i}^{\mathsf{S}}$ where

$$\tilde{\boldsymbol{\beta}}_{i:\mathcal{K}_i}^{\mathsf{S}} = (\mathbf{X}_{i:\mathcal{K}_i}^{\mathsf{S}^t} \mathbf{C}_{i:\mathcal{K}_i}^{-1} \mathbf{X}_{i:\mathcal{K}_i}^{\mathsf{S}})^{-1} \mathbf{X}_{i:\mathcal{K}_i}^{\mathsf{S}^t} \mathbf{C}_{i:\mathcal{K}_i}^{-1} \mathbf{Y}_{i:\mathcal{K}_i} \tag{6}$$

Note that $\mathbf{Y}_{i:\mathcal{K}_i}$ has length $|\mathcal{K}_i| \times 1$, and $\tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{S}}$ always has length $n_i \times 1$. When $\mathcal{K}_i = n_i$, we write $\tilde{\mathbf{Y}}_i^{\mathsf{S}}$ and $\tilde{\boldsymbol{\beta}}_i^{\mathsf{S}}$. In general, the use of a ''~'' indicates that a parameter is estimated by using a separate model.

- $\hat{\mathbf{Y}}_i^S = \mathbf{X}_i^S \hat{\boldsymbol{\beta}}_i^S$, where $\hat{\boldsymbol{\beta}}_i^S$ is an estimate of the coefficients whose predictors are in the set S that utilizes the estimated parameters of the hierarchical model. Specifically, we first fit an HNRM that includes the predictors in the set F. Let $\hat{\boldsymbol{\beta}}_i^F$, $\hat{\boldsymbol{\Sigma}}_i$, and $\hat{\boldsymbol{\Sigma}}_\beta$ be the maximum likelihood estimates of the corresponding parameters of the HNRM. In the ith group, let $\mathbf{b}_i^F \sim N(\hat{\boldsymbol{\beta}}_i^F, \hat{\boldsymbol{\Lambda}}_i^F)$, where $\hat{\boldsymbol{\Lambda}}_i^F = (\mathbf{X}_i^t \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{X}_i + \hat{\boldsymbol{\Sigma}}_\beta^{-1})^{-1}$ is an estimate of the variance of $\mathbf{b}_i^F$. Then, $\hat{\boldsymbol{\beta}}_i^S$ is the mode of the distribution of the coefficients whose predictors are in the set S, conditional on the coefficients in the set $F - S$ equaling zero (i.e., the mode of $f(\mathbf{b}_i^S | \mathbf{b}_i^{F-S} = \mathbf{0})$). When $S = F$, we simply use $\hat{\mathbf{Y}}_i^F = \mathbf{X}_i \hat{\beta}_i^F$. In general, the use of a ''^'' indicates that a parameter is estimated by using a hierarchical model.
- We define two estimates of conditional variances:

$$\tilde{\Sigma}_i^S = \tilde{\sigma}_i^S \mathbf{C}_i = ((\mathbf{Y}_i - \tilde{\mathbf{Y}}_i^S)^t \mathbf{C}_i^{-1} (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i^S)/n_i) \mathbf{C}_i \tag{7}$$

$$\hat{\Sigma}_i^S = \hat{\sigma}_i^S \mathbf{C}_i = ((\mathbf{Y}_i - \hat{\mathbf{Y}}_i^S)^t \mathbf{C}_i^{-1} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^S)/n_i) \mathbf{C}_i \tag{8}$$

- We also make use of the following quantities in some procedures:

$$\mathbf{V}_i = \hat{\boldsymbol{\Sigma}}_i^F + \mathbf{X}_i \hat{\boldsymbol{\Lambda}}_i^F \mathbf{X}_i^t \tag{9}$$

$$\mathbf{E}_i^S = \mathbf{X}_i^S (\mathbf{X}_i^{S^t} [\tilde{\Sigma}_i^F]^{-1} \mathbf{X}_i^S)^{-1} \mathbf{X}_i^{S^t} [\tilde{\Sigma}_i^S]^{-1} \hat{\mathbf{Y}}_i^F \tag{10}$$

We label the HMSPs with a series of letters. The labels for local and global HMSPs begin with *HL* and *HG*, respectively. The next letter in a procedure's name corresponds to the first letter of the name of the strategy that the procedure implements (e.g., *P* indicates the Predictor Exclusion strategy, *S* indicates the Stepwise/Full Selection strategy, and so on). The final letters in each procedure's name describe characteristics of the procedure's criterion. For example, hierarchical, local Predictor Exclusion procedures based on AIC-like and BIC-like criteria are called HLP-AIC and HLP-BIC, respectively.

We also define four other procedures-FULL, HMFULL, MAIC, and MBIC – which are useful for comparisons. For FULL, $\tilde{\mathcal{Y}}_i = \mathcal{X}_i \tilde{\boldsymbol{\beta}}_i^F$, and for HMFULL, $\tilde{\mathcal{Y}}_i = \mathcal{X}_i \hat{\boldsymbol{\beta}}_i^F$. MAIC and MBIC find the subset of predictors in each group that minimizes the AIC and the BIC, respectively. The criteria for MAIC and MBIC are the same as the criteria for HLP-AIC and HLP-BIC shown in Table 1, except that $\tilde{\mathbf{Y}}_i^S$ and $\tilde{\Sigma}_i^S$ replace $\hat{\mathbf{Y}}_i^S$ and $\hat{\Sigma}_i^S$. For MAIC, MBIC, and all HMSPs, once the S_i and K_i are selected, we estimate the $\mathcal{Y}_i$ with:

$$\tilde{\mathcal{Y}}_i = \mathcal{X}_i^{S_i} \tilde{\boldsymbol{\beta}}_{i:K_i}^{S_i} \tag{11}$$

The criteria for the local HMSPs are presented in Table 1. To save space, the five local Dimension Selection procedures – labeled HLD-ED, HLD-VWD, HLD-AIC, HLD-BIC, and HLD-MSE – are not shown. For each of these five procedures, in each group we first apply the corresponding Predictor Exclusion procedure (e.g., for HLD-ED, apply HLP-ED) and determine a temporary S_i. Then, in each group, we let $d_i = |S_i|$ and find a new set of predictors S_i such that:

$$S_i = \arg \max_{S:|S| = d_i} R_i^2(S) \tag{12}$$

Table 1. *Criteria for the local HMSPs*

Procedure	Select S_i or K_i such that:
Pred. Excl.	
HLP-ED	$\mathsf{S}_i = \arg \min_{\mathsf{S}} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}}))$
HLP-VWD	$\mathsf{S}_i = \arg \min_{\mathsf{S}} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t \mathbf{V}_i^{-1} (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}}))$
HLP-AIC	$\mathsf{S}_i = \arg \min_{\mathsf{S}} (\log(|\hat{\Sigma}_i^{\mathsf{S}}|) + (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}})^t [\hat{\Sigma}_i^{\mathsf{S}}]^{-1} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}}) + 2|\mathsf{S}|)$
HLP-BIC	$\mathsf{S}_i = \arg \min_{\mathsf{S}} (\log(|\hat{\Sigma}_i^{\mathsf{S}}|) + (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}})^t [\hat{\Sigma}_i^{\mathsf{S}}]^{-1} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}}) + |\mathsf{S}| \log(n_i))$
HLP-MSE	$\mathsf{S}_i = \arg \min_{\mathsf{S}} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \mathbf{E}_i^{\mathsf{S}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \mathbf{E}_i^{\mathsf{S}}) + \mathrm{tr}(\mathbf{X}_i^{\mathsf{S}} (\mathbf{X}_i^{\mathsf{S}^t} [\tilde{\Sigma}_i^{\mathsf{F}}]^{-1} \mathbf{X}_i^{\mathsf{S}})^{-1} \mathbf{X}_i^{\mathsf{S}^t}))$
Step./Full Sel.	
HLS-AIC	$\mathsf{S}_i = \arg \min_{\mathsf{S} \in \{\mathsf{F}, \mathsf{S}_i^{\mathsf{A}}\}} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}}))$
	where $\mathsf{S}_i^{\mathsf{A}}$ is the set that minimizes the AIC in group i
HLS-BIC	$\mathsf{S}_i = \arg \min_{\mathsf{S} \in \{\mathsf{F}, \mathsf{S}_i^{\mathsf{B}}\}} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}}))$
	where $\mathsf{S}_i^{\mathsf{B}}$ is the set that minimizes the BIC in group i
Unit Excl.	
HLU-ED	$\mathsf{K}_i = \arg \min_{\mathcal{K}_i} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{F}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{F}}))$
HLU-VWD	$\mathsf{K}_i = \arg \min_{\mathcal{K}_i} ((\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{F}})^t \mathbf{V}_i^{-1} (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{F}}))$

where

$$R_i^2(\mathsf{S}) = 1 - (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i^{\mathsf{S}})/(\mathbf{Y}_i - \bar{\mathbf{Y}}_i)^t (\mathbf{Y}_i - \bar{\mathbf{Y}}_i) \tag{13}$$

HLP-ED and HLP-VWD treat $\hat{\mathbf{Y}}_i^{\mathsf{F}}$ as a gold standard and search for the separate model whose $\tilde{\mathbf{Y}}_i^{\mathsf{S}}$ has the smallest Euclidean distance (ED) or variance weighted distance (VWD) from this standard. These criteria are designed so that predictors whose estimated coefficients in FULL are far from those in the HNRM are excluded, which can improve estimates of the $\boldsymbol{y}_i$ on average since the $\hat{\boldsymbol{\beta}}_i^{\mathsf{F}}$ tend to be closer to the $\boldsymbol{\beta}_i$ than the $\tilde{\boldsymbol{\beta}}_i^{\mathsf{F}}$ are. The $\mathbf{V}_i$ in HLP-VWD estimates the variance of new predictions of the outcomes at $\mathbf{X}_i$, using $\mathbf{X}_i \hat{\boldsymbol{\beta}}_i^{\mathsf{F}}$ as the estimator of the new predictions. The $\mathbf{V}_i$ discounts the effects on the criterion of units whose estimates in $\hat{\mathbf{Y}}_i^{\mathsf{F}}$ have large variances.

HLP-AIC and HLP-BIC are analogues of MAIC and MBIC. When $\hat{\mathbf{Y}}_i^{\mathsf{S}}$ is sufficiently close to $\hat{\mathbf{Y}}_i^{\mathsf{F}}$, thereby suggesting that the predictors in $\mathsf{F} - \mathsf{S}$ are unimportant, the values of these criteria decrease because of the penalty, and the submodel is preferred to the full model. An advantage of HLP-AIC and HLP-BIC over MAIC and MBIC is that they use generally more reliable estimates of coefficients and so may lead to better decisions about excluding predictors. For example, consider a scenario where sampling variability results in a spuriously low estimate of a coefficient in FULL, but the estimate from the HNRM is appropriately large. MAIC or MBIC may drop the coefficient's predictor, whereas the HMSPs are more likely to keep it in the model.

The criterion in HLP-MSE is derived from the formula for the total mean squared error of $\tilde{\mathbf{Y}}_i^{\mathsf{S}}$, assuming the ''true regression model'' includes all predictors in the set F. We use $\hat{\mathbf{Y}}_i^{\mathsf{F}}$ to estimate $\mathbf{X}_i \beta_i$, and $\mathbf{E}_i^{\mathsf{S}}$ to estimate $E(\mathbf{X}_i^{\mathsf{S}} \tilde{\boldsymbol{\beta}}_i^{\mathsf{S}})$. Thus, HLP-MSE attempts to find a model that approximately minimizes the prediction mean squared errors.

The Stepwise/Full Selection procedures find the separate models whose fitted values are closest to those of the HNRM, but only two models in each group's separate model space are considered. The Unit Exclusion procedures also treat $\hat{\mathbf{Y}}_i^{\mathsf{F}}$ as a gold standard, so that finding the $\tilde{\mathbf{Y}}_{i:\mathcal{K}_i}^{\mathsf{F}}$ that is closest to the $\hat{\mathbf{Y}}_i^{\mathsf{F}}$ seeks to improve on average the estimates of the $\boldsymbol{\beta}_i$.

To develop global HMSPs, we use criteria that are unweighted sums across groups of the criteria developed for the local HMSPs. It is also possible to create procedures that employ weighted averages of the locally evaluated criteria. Such procedures are not pursued in this article.

The global HMSPs are summarized in Table 2. Again, to save space, the corresponding five Dimension Selection HMSPs – labeled HGD-ED, HGD-VWD, HGD-AIC, HGD-BIC, and HGD-MSE – are not listed on the table. For each of these five procedures, we first apply the corresponding Predictor Exclusion procedure to select a temporary, common set of predictors, S^*. Let $d^* = |\mathsf{S}^*|$. Then, we find an S_i in each

Table 2.   Criteria for the global HMSPs

Procedure	*Select S_i such that:*
Pred. Excl.	
HGP-ED	$\mathsf{S}_i = \mathsf{S}^* = \arg \min\limits_{\mathsf{S}} \left(\sum\limits_i (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}}) \right)$
HGP-VWD	$\mathsf{S}_i = \mathsf{S}^* = \arg \min\limits_{\mathsf{S}} \left(\sum\limits_i (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}})^t \mathbf{V}_i^{-1} (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathsf{S}}) \right)$
HGP-AIC	$\mathsf{S}_i = \mathsf{S}^* = \arg \min\limits_{\mathsf{S}} \left(\sum\limits_i \log(|\hat{\Sigma}_i^{\mathsf{S}}|) + (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}})^t [\hat{\Sigma}_i^{\mathsf{S}}]^{-1} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}}) + 2|\mathsf{S}| \right)$
HGP-BIC	$\mathsf{S}_i = \mathsf{S}^*$
	$= \arg \min\limits_{\mathsf{S}} \left(\sum\limits_i \log(|\hat{\Sigma}_i^{\mathsf{S}}|) + (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}})^t [\hat{\Sigma}_i^{\mathsf{S}}]^{-1} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{\mathsf{S}}) + |\mathsf{S}| \log(n_i) \right)$
HGP-MSE	$\mathsf{S}_i = \mathsf{S}^*$
	$= \arg \min\limits_{\mathsf{S}} \left(\sum\limits_i (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \mathbf{E}_i^{\mathsf{S}})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \mathbf{E}_i^{\mathsf{S}}) + \mathrm{tr}(\mathbf{X}_i^{\mathsf{S}} (\mathbf{X}_i^{\mathsf{S}'} [\tilde{\Sigma}_i^{\mathsf{F}}]^{-1} \mathbf{X}_i^{\mathsf{S}})^{-1} \mathbf{X}_i^{\mathsf{S}'}) \right)$
Step./Full Sel.	
HGS-AIC	$\mathsf{S}_i = \mathcal{S}^*[i]$, where $\mathcal{S}^* = \arg \min\limits_{\mathcal{S} \in \{\mathcal{F}, \mathcal{S}^{\mathsf{A}}\}} \left(\sum\limits_i (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathcal{S}[i]})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathcal{S}[i]}) \right)$
	where $\mathcal{F} = \{\mathsf{F}, \mathsf{F}, \ldots, \mathsf{F}\}$ with $|\mathcal{F}| = \mathcal{G}$, and $\mathcal{S}^{\mathsf{A}} = \{\mathsf{S}_1^{\mathsf{A}}, \mathsf{S}_2^{\mathsf{A}}, \ldots, \mathsf{S}_{\mathcal{G}}^{\mathsf{A}}\}$
HGS-BIC	$\mathsf{S}_i = \mathcal{S}^*[i]$, where $\mathcal{S}^* = \arg \min\limits_{\mathcal{S} \in \{\mathcal{F}, \mathcal{S}^{\mathsf{B}}\}} \left(\sum\limits_i (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathcal{S}[i]})^t (\hat{\mathbf{Y}}_i^{\mathsf{F}} - \tilde{\mathbf{Y}}_i^{\mathcal{S}[i]}) \right)$
	where $\mathcal{F} = \{\mathsf{F}, \mathsf{F}, \ldots, \mathsf{F}\}$ with $|\mathcal{F}| = \mathcal{G}$, and $\mathcal{S}^{\mathsf{B}} = \{\mathsf{S}_1^{\mathsf{B}}, \mathsf{S}_2^{\mathsf{B}}, \ldots, \mathsf{S}_{\mathcal{G}}^{\mathsf{B}}\}$

Note: $\mathcal{S}^*[i]$ is the set of predictors in $\mathcal{S}^*$ for group i.

group such that:

$$S_i = \arg \max_{S:|S|=d^*} R_i^2(S) \tag{14}$$

Like their local counterparts, global Dimension Selection HMSPs select predictors separately in each group.

In the global Predictor Exclusion procedures, the summations across groups imply that decisions to include or exclude a predictor depend on the estimate of its coefficient in all the groups. For example, when a predictor is estimated as important in most groups and unimportant in the rest of the groups, the predictor is likely to be included in the groups' models. Similarly, when the predictor is estimated as unimportant in most groups and important in the rest of the groups, the predictor is likely to be excluded from all the models.

5. Empirical Evaluations of the Procedures' Accuracies

Typically, statisticians evaluate the performances of model selection procedures in finite samples via simulation studies (e.g., Dempster, Schatzoff, and Wermuth 1977b; Miller 1990). As noted by Dempster et al. (1977b), simulation studies are especially useful since they can provide empirical evidence of a procedure's operating characteristics in data sets that typify realistic settings. In this section, I present two simulation studies of the procedures using generated data. These studies serve as blueprints for evaluating medial information pooling procedures in real settings.

Study I is an assessment of the procedures' accuracies ''on average'' across many replicated data sets with a variety of characteristics. This type of study is especially useful when it is desired, or necessary, to specify the statistical procedures that will be used *before* actually collecting the data, as such assessments can provide some indication of whether the procedures have the potential to increase accuracy sufficiently so as to make arguing for their legality worthwhile. The study also suggests general conclusions about medial information pooling procedures.

Study II illustrates a method for predicting the procedures' accuracies in any particular data set using observable characteristics of that data set. Specifically, we build a regression estimator that predicts the procedures' performances from observable measures of the degree of collinearity among the predictors, the amount of across-group information, and the distances of the estimated coefficients from zero. This type of study is useful when it is possible to specify statistical procedures *after* collecting the data, since the observable data can be used to obtain more precise predictions of the procedures' accuracies than predictions from ''on average'' studies.

To describe the simulations of both studies, we adapt the notation of Section 4 by writing a subscript ''(r)'' to refer to the variables for each replicated data set r. Thus, for replication r, $\mathcal{G}_{(r)}$ is the number of groups, $\mathsf{F}_{(r)}$ is the full set of predictors, $\mathcal{Y}_{(r)i}$ and $\mathcal{X}_{(r)i}$ are, respectively, the unobserved, $m_{(r)i} \times 1$ vector of dependent variables for group i and the $m_{(r)i} \times |\mathsf{F}_{(r)}|$ matrix of the values of the predictors associated with $\mathcal{Y}_{(r)i}$, and $\mathbf{Y}_{(r)i}$ and $\mathbf{X}_{(r)i}$ are, respectively, the observed, $n_{(r)i} \times 1$ vector of dependent variables for group i and the $n_{(r)i} \times |\mathsf{F}_{(r)}|$ matrix of the values of the predictors associated with $\mathbf{Y}_{(r)i}$.

For all replications, data sets are simulated from a common data generation model: for each group $i = 1, \ldots, \mathcal{G}_{(r)}$,

$$\mathbf{Y}_{(r)i} \sim N(\mathbf{X}_{(r)i}\boldsymbol{\beta}_{(r)i}, \sigma_{(r)i}\mathbf{I}_{n_{(r)i}}) \tag{15}$$

$$\mathcal{Y}_{(r)i} \sim N(\mathcal{X}_{(r)i}\boldsymbol{\beta}_{(r)i}, \sigma_{(r)i}\mathbf{I}_{m_{(r)i}}) \tag{16}$$

For convenience, in all replications, $n_{(r)i} = m_{(r)i} = 40$ for $i = 1, \ldots, \mathcal{G}_{(r)}$. Using different $n_{(r)i}$ is roughly equivalent to estimating the regression coefficients with different precisions. We incorporate such differences in precisions by varying the $\sigma_{(r)i}$. We set $\mathcal{G}_{(r)} = 50$ and $|\mathsf{F}_{(r)}| = 15$ for all r. The values of the $\mathbf{X}_{(r)i}$, $\mathcal{X}_{(r)i}$, $\boldsymbol{\beta}_{(r)i}$, and $\sigma_{(r)i}$ vary across replications.

We refer to variables estimated by using procedure p by writing a superscript ''(p)''. For example, to refer to the estimates of $\mathcal{Y}_{(r)i}$ determined by using procedure p, we write $\tilde{\mathcal{Y}}_{(r)i}^{(p)}$. As a criterion for evaluating procedure p in group i in replication r, we define the relative squared prediction error:

$$RSPE_{(r)i}^{(p)} = (\mathcal{Y}_{(r)i} - \tilde{\mathcal{Y}}_{(r)i}^{(p)})^t(\mathcal{Y}_{(r)i} - \tilde{\mathcal{Y}}_{(r)i}^{(p)})/(\mathcal{Y}_{(r)i} - \tilde{\mathcal{Y}}_{(r)i}^{(FULL)})^t(\mathcal{Y}_{(r)i} - \tilde{\mathcal{Y}}_{(r)i}^{(FULL)}) \tag{17}$$

When $RSPE_{(r)i}^{(p)} < 1$, procedure p predicts $\mathcal{Y}_{(r)i}$ more accurately than FULL does; when $RSPE_{(r)i}^{(p)} > 1$, procedure p predicts $\mathcal{Y}_{(r)i}$ less accurately than FULL does.

For each procedure in each replication, we compute three summaries of the $RSPE_{(r)i}^{(p)}$:

$$\mathcal{A}_{(r)}^{(p)} = (1/50) \sum_i RSPE_{(r)i}^{(p)} \tag{18}$$

$$\mathcal{M}_{(r)}^{(p)} = \max_i RSPE_{(r)i}^{(p)} \tag{19}$$

$$\mathcal{P}_{(r)}^{(p)} = (1/50) \sum_i I[RSPE_{(r)i}^{(p)} > 1] \tag{20}$$

where $I[\ldots]$ is an indicator function that equals one when the expression in the square brackets is true and equals zero when this expression is false. These criteria assess different aspects of the procedures' performances. For example, suppose that in some data set r, procedure p has $\mathcal{A}_{(r)}^{(p)} = 0.90$, $\mathcal{M}_{(r)}^{(p)} = 1.40$, and $\mathcal{P}_{(r)}^{(p)} = 0.30$. Statisticians considering the use of this procedure instead of FULL must decide whether the 10% reduction in average squared prediction error outweighs the 40% increase in squared prediction errors in one group and increased squared prediction errors in 30% of the groups.

It is possible to conceive of other evaluation criteria. For example, for each $RSPE_{(r)i}^{(p)}$, one could use variance-weighted quadratic forms instead of the unweighted quadratic forms in (17). Such variance-weighted forms adjust for scale differences in the groups' dependent variables. In the applications of medial information pooling considered in this article, variables are expected to be on the same scale in each group, so that an unweighted quadratic form is sensible. When variance-weighted or other criteria are more relevant to some potential users' setting, ideally simulation studies based on these criteria would be undertaken. Additionally, different evaluation criteria may be optimized by procedures other than those presented here.

Table 3.   Design parameters for simulating data sets in Study I

Design Parameters
$\beta_{(r)i}^0 = \beta^0 = 0$
$\mathbf{X}_{(r)ij}^{\mathsf{F}-1} \sim N(\mathbf{0}, \mathbf{\Delta}_{(r)})$
$\mathcal{X}_{(r)ij}^{\mathsf{F}-1} \sim N(\mathbf{0}, \mathbf{\Delta}_{(r)})$
$\sigma_{(r)i} \sim \chi_{v_{(r)}}^2$, where $v_{(r)} \in \{10, 40, 90, 160, 250, 360, 640\}$

5.1.   Study I: Averaging across data sets

In general, to maximize the usefulness of ''on average'' studies, users should design them so that the replicated data sets have characteristics that resemble those likely to exist in the real populations. In this subsection, I illustrate an ''on average'' simulation study using the design shown in Table 3.

Because the procedures are invariant under linear transformations, the intercept and the mean vector of the nonintercept predictors for each unit ij are irrelevant and are set to zero. In each replication r, the $\sigma_{(r)i}$ are drawn from a chi-squared distribution with randomly drawn degrees of freedom $v_{(r)}$, so that the values of the $\sigma_{(r)i}$, which have expectation $v_{(r)}$, vary across replications.

The variance matrix of the predictors, $\mathbf{\Delta}_{(r)}$, is drawn randomly in each replication by the following method: 1) draw a continuous value $h_{(r)}$ from a uniform distribution on $[0, .5]$; 2) create a 14×14 scale matrix, $S_{(r)}$, with $h_{(r)}$ off the diagonals and one on the diagonals; 3) draw an integer value $df_{(r)}$ with $P(df_{(r)} = 15) = 2/3$ and $P(df_{(r)} = 500) = 1/3$; 4) draw $\mathbf{\Delta}_{(r)}$ from a scaled Wishart distribution with scale parameter equal to $S_{(r)}$ and degrees of freedom equal to $df_{(r)}$; and, 5) standardize $\mathbf{\Delta}_{(r)}$ to have variances of one on the diagonals. When $df_{(r)} = 500$, the drawn $\mathbf{\Delta}_{(r)} \approx S_{(r)}$. When $df_{(r)} = 15$, $\mathbf{\Delta}_{(r)}$ typically contains a variety of off-diagonal correlations, including some that are near one. The $df_{(r)}$ is drawn as 15 with more probability to increase the variety of correlation structures. The value of $h_{(r)}$ is bounded by 0.5 because larger values frequently produce nearly singular matrices.

We draw values for the true regression coefficients in each replication by the following process. First, we draw independently two mean values, $\mu_{(r)}^A$ and $\mu_{(r)}^B$, from $\{0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 8.0, 16.0\}$, and draw independently two variance values, $\tau_{(r)}^C$ and $\tau_{(r)}^D$, from $\{0.01, 0.25, 1.00, 2.25, 4.00, 6.25, 9.00, 12.25, 16.00\}$. Then, we randomly partition the 14 nonintercept predictors into four sets: $AC_{(r)}$, $AD_{(r)}$, $BC_{(r)}$, and $BD_{(r)}$. Some of these sets may be empty. In each group i, the predictors' coefficients in these four sets are drawn independently from the following four distributions:

$$\beta_{(r)i}^s \sim N(\mu_{(r)}^A, \tau_{(r)}^C), \text{ for } s \in AC_{(r)} \tag{21}$$

$$\beta_{(r)i}^s \sim N(\mu_{(r)}^A, \tau_{(r)}^D), \text{ for } s \in AD_{(r)} \tag{22}$$

$$\beta_{(r)i}^s \sim N(\mu_{(r)}^B, \tau_{(r)}^C), \text{ for } s \in BC_{(r)} \tag{23}$$

$$\beta_{(r)i}^s \sim N(\mu_{(r)}^B, \tau_{(r)}^D), \text{ for } s \in BD_{(r)} \tag{24}$$

This method varies the coefficients' similarities across groups and distances from zero

in the replications, since up to four different coefficient structures can exist in the same data set. For example, when $\mu^A_{(r)} = 0.1$, $\mu^B_{(r)} = 8.0$, $\tau^C_{(r)} = .01$, and $\tau^D_{(r)} = 16.00$, the coefficients in $AC_{(r)}$ are typically very similar and near zero across all groups, those in $AD_{(r)}$ typically differ widely across groups with an average near zero, those in $BC_{(r)}$ are typically very similar and far from zero across all groups, and those in $BD_{(r)}$ typically differ widely across groups with an average far from zero.

Using this design, we draw 250 replicated data sets. Across the data sets, the correlation structures of the predictors range fairly uniformly from near-independence to severe multicollinearity. The amounts of across-group information in the data sets, as measured by the values of the $\mathcal{A}^{(\mathrm{HMFULL})}_{(r)}$, range fairly uniformly from practically none to an amount consistent with populations where $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i, j, although there are relatively few data sets with no across-group information. About 80% of the regression coefficients are within two standard deviations from zero, with a wide variety of variances and locations across and within data sets. Thus, for $|\mathsf{F}| = 15$ and $\mathcal{G} = 50$, these simulations appear to cover a wide range of settings. See Reiter (1999) for more descriptions of the characteristics of the simulated data sets.

For each procedure p, Table 4 shows the averages of the $\mathcal{A}^{(p)}_{(r)}$, $\mathcal{M}^{(p)}_{(r)}$, and $\mathcal{P}^{(p)}_{(r)}$ across the 250 replications, which we abbreviate for each procedure p as $\bar{\mathcal{A}}^{(p)}$, $\bar{\mathcal{M}}^{(p)}$, and $\bar{\mathcal{P}}^{(p)}$, respectively. The standard error of each mean in Table 4 is less than .01. Also shown for each procedure p are the minimum and maximum $\mathcal{A}^{(p)}_{(r)}$ and the number of times each procedure has $\mathcal{A}^{(p)}_{(r)} > 1$. From such results, potential users can decide whether these procedures are worthwhile. For example, in some settings, HLP-ED may be so controversial that the expected .13 reduction in average squared prediction error is not worth the extra controversy, whereas HLU-ED may be relatively uncontroversial so that the expected .04 reduction is worth it. In other settings, users may decide HLP-ED is worthwhile, since it potentially can reduce $\mathcal{A}^{(p)}_{(r)}$ by almost 30% without much risk of producing an $\mathcal{A}^{(p)}_{(r)} > 1$.

This study also suggests some general conclusions about these HMSPs. First, all HMSPs have an $\bar{\mathcal{A}}^{(p)} < 1$, showing that there are ''on average'' gains when using medial information pooling relative to full, separate regressions. Second, $\bar{\mathcal{M}}^{(p)} > 1$ and $\bar{\mathcal{P}}^{(p)} > 0$ for all HMSPs, thus indicating that such ''on average'' gains are accompanied with reductions in accuracy in some groups. Third, global implementations of the strategies typically are not as effective as local implementations. Thus, when local or global implementations of a strategy are equally acceptable, I suggest using the local implementations. Fourth, the reductions in average relative squared prediction error are largest for the Predictor Exclusion strategy, next largest for the Stepwise/Full Selection and Unit Exclusion strategies, and smallest for the Dimension Selection strategy. This is expected given the strategies' decreasing reliance on multiple groups' data for model specification. Fifth, all HMSPs outperform MAIC and MBIC on all evaluation criteria. In fact, as shown by Reiter (1999), compared to MAIC and MBIC, all HMSPs avoid big increases in average squared prediction error when many predictors are important, and most reduce average squared prediction error by larger amounts when many predictors are unimportant. Finally, all procedures lag significantly behind HMFULL. This is evidence that adherence to legal constraints that prohibit explicit data pooling generally sacrifices accuracy.

Upon closer examination, we find, not surprisingly, that the procedures' performances

Table 4. Summaries of the $\mathcal{A}_{(r)}^{(p)}$, $\mathcal{M}_{(r)}^{(p)}$, and $\mathcal{P}_{(r)}^{(p)}$ across the 250 data sets

Strategy (scope)	Procedure	$\bar{\mathcal{A}}^{(p)}$	> 1.0*	$\mathcal{A}^{(p)}$		$\bar{\mathcal{M}}^{(p)}$	$\bar{\mathcal{P}}^{(p)}$
				min	max		
	HMFULL	.77	0	.61	.96	1.16	.09
	MAIC	1.01	138	.83	1.25	1.51	.48
	MBIC	1.03	154	.77	1.57	1.77	.50
Pred. Excl. (local)	HLP-ED	.87	6	.71	1.05	1.29	.17
	HLP-VWD	.87	4	.70	1.05	1.29	.17
	HLP-AIC	.93	18	.80	1.03	1.26	.28
	HLP-BIC	.93	47	.76	1.10	1.39	.32
	HLP-MSE	.90	11	.72	1.02	1.31	.26
Step./Full Sel. (local)	HLS-AIC	.96	68	.83	1.01	1.19	.09
	HLS-BIC	.96	63	.77	1.01	1.19	.07
Dim. Sel. (local)	HLD-ED	.98	15	.79	1.06	1.23	.38
	HLD-VWD	.98	11	.88	1.06	1.23	.38
	HLD-AIC	.98	81	.84	1.06	1.26	.41
	HLD-BIC	.98	100	.80	1.13	1.39	.43
	HLD-MSE	.96	96	.75	1.11	1.35	.40
Unit Excl. (local)	HLU-ED	.96	3	.92	1.04	1.21	.17
	HLU-VWD	.96	5	.92	1.03	1.21	.18
Pred. Excl. (global)	HGP-ED	.92	10	.69	1.03	1.23	.23
	HGP-VWD	.92	3	.69	1.01	1.19	.20
	HGP-AIC	.95	3	.70	1.02	1.13	.15
	HGP-BIC	.94	26	.70	1.17	1.36	.24
	HGP-MSE	.91	20	.68	1.05	1.30	.25
Step./Full Sel. (global)	HGS-AIC	.97	0	.83	1.00	1.10	.09
	HGS-BIC	.97	0	.77	1.00	1.09	.06
Dim. Sel. (global)	HGD-ED	.97	48	.75	1.18	1.17	.33
	HGD-VWD	.98	40	.75	1.07	1.15	.31
	HGD-AIC	.98	23	.70	1.03	1.08	.23
	HGD-BIC	.97	44	.70	1.18	1.25	.30
	HGD-MSE	.97	57	.70	1.18	1.22	.35

*The number of replications out of 250 with $\mathcal{A}_{(r)}^{(p)} > 1.0$.

are sensitive to characteristics of the data. All HMSPs are more effective when the HNRM improves estimates dramatically relative to FULL and when there are only small correlations among the predictors in the $\mathbf{X}_i$. Also, all procedures, except the Unit Exclusion procedures, increase in effectiveness as the $\beta_{(r)i}^{\mathsf{F}}$ move closer to zero, where distance is measured in terms of standard deviations of the $\tilde{\beta}_i$. The Unit Exclusion procedures do not depend on the locations of the coefficients.

The specific values in Table 4 are not indicative of these procedures' performances in data sets with different characteristics than those covered by the simulation design in Table 3. In fact, Reiter (1999) shows that for larger values of $\mathcal{G}$ and $|\mathsf{F}|$, the HMSPs reduce average squared prediction errors by larger amounts at the cost of increasing squared prediction errors in some groups by larger amounts. Additionally, when a hierarchical

model that does not fit the data well is used for information extraction, medial information pooling procedures can lose their effectiveness. For example, when mis-specification of the distribution of the $\boldsymbol{\beta}_i$ in (5) results in some groups' estimated coefficients being pulled close to zero and far from their true values, the predictors associated with these coefficients are incorrectly excluded by those procedures that drop predictors. In such groups, estimates from the medial information pooling procedures are less accurate than those from FULL. Excessive shrinkage towards values far from zero can attenuate potential gains from medial information pooling, since procedures are likely to keep predictors with spuriously large estimated coefficients in the models.

5.2.   Study II: Predicting performances from observable data

A shortcoming of using ''on average'' simulation studies for deciding whether or not to employ medial information pooling is that they do not utilize fully information that exists in the observed data set. For example, when $\tilde{\mathbf{Y}}_i^{\mathsf{F}} \approx \hat{\mathbf{Y}}_i^{\mathsf{F}}$ for all i, the procedures do not give a substantial payoff relative to FULL (e.g., HLP-ED selects F). In contrast, when each 14×1 vector $\boldsymbol{\beta}_i = \mathbf{0}$ for $i = 1, \ldots, 50$, $\mathcal{A}_{(r)}^{(p)} < .75$ for all HMSPs except the Unit Exclusion procedures. When decisions about information pooling can be made after data are collected, we can utilize the data to get better estimates of the performances of the procedures.

In this section, we predict the procedures' $\mathcal{A}_{(r)}^{(p)}$s using regressions of the $\mathcal{A}_{(r)}^{(p)}$s on observable measures of relevant data characteristics. We focus on predicting $\mathcal{A}_{(r)}^{(p)}$s rather than $\mathcal{M}_{(r)}^{(p)}$s or $\mathcal{P}_{(r)}^{(p)}$s because a procedure that fails to reduce $\mathcal{A}_{(r)}^{(p)}$ is not worth using, regardless of the values of the other criteria. The process of undertaking the study – which serves as a template for other studies of this kind – includes simulating representative data sets, developing estimators of performance measures, evaluating these estimators, and examining these estimators' performances under any proposed decision rules.

For simulated data sets, we use the 250 replications from the previous study. Additionally, we add twenty replications with little across-group information, since relatively few of the 250 data sets have $\mathcal{A}_{(r)}^{(\mathrm{HMFULL})} \approx 1$. These twenty additional data sets assist us in predicting $\mathcal{A}_{(r)}^{(p)}$s when there is little across-group information.

Investigations indicated that the most relevant factors for predicting $\mathcal{A}_{(r)}^{(p)}$ are the degree of collinearity among the predictors in the $\mathbf{X}_{(r)i}$, the amount of across-group information, and the distances of coefficients from zero. Thus, we require observable measures of these factors to use in the regression models. We measure the degree of collinearity among the predictors in replicated data set r by averaging the condition numbers, $\kappa_{(r)i}$, across groups:

$$\kappa_{(r)} = (1/50) \sum_i \kappa_{(r)i} = (1/50) \sum_i \max \mathrm{SVD}_{(r)i} / \min \mathrm{SVD}_{(r)i} \tag{25}$$

where $\mathrm{SVD}_{(r)i}$ is the vector of singular values of a re-scaled version of $\mathbf{X}_{(r)i}$ with columns of unit length. The quantity $\kappa_{(r)i}$ is recommended as a measure of the degree of collinearity by Belsley (1990). As an observable measure of the effectiveness of explicit data pooling, we use the ratio of the summed squared residuals from the HNRM to the summed squared residuals from FULL:

$$\hat{R}_{(r)} = \sum_i (\mathbf{Y}_{(r)i} - \hat{\mathbf{Y}}_{(r)i}^{\mathsf{F}})^t (\mathbf{Y}_{(r)i} - \hat{\mathbf{Y}}_{(r)i}^{\mathsf{F}}) / \sum_i (\mathbf{Y}_{(r)i} - \tilde{\mathbf{Y}}_{(r)i}^{\mathsf{F}})^t (\mathbf{Y}_{(r)i} - \tilde{\mathbf{Y}}_{(r)i}^{\mathsf{F}}) \tag{26}$$

In the 270 simulated data sets, $Cor(\hat{R}, \mathcal{A}^{(\text{HMFULL})}) = -0.92$, showing that $\hat{R}$ is a useful proxy for $\mathcal{A}^{(\text{HMFULL})}$, which itself is a measure of the amount of across-group information. A shortcoming of $\hat{R}$ is that it is large also when the HNRM fits poorly, but this is not an issue in our simulation design. As an observable measure of the distances of the true regression coefficients from zero, we use standardized, estimated non-intercept coefficients of the HNRM. Specifically, these standardized coefficients, which we denote as $|b^s_{(r)i}|$, are:

$$|b^s_{(r)i}| = |\hat{\beta}^s_{(r)i}|/(\widetilde{Var}(\tilde{\beta}^s_{(r)i}))^{\frac{1}{2}} \tag{27}$$

where $\widetilde{Var}(\tilde{\beta}^s_{(r)i})$ is the sth diagonal element of $\tilde{\sigma}^{\text{F}}_{(r)i}(\mathbf{X}^t_{(r)i}\mathbf{X}_{(r)i})^{-1}$. For each replication r, we average these $|b^s_{(r)i}|$ across groups and coefficients to create four categorical predictors:

$$B.5_{(r)} = (1/(50)(14)) \sum_{i,s} I[0.0 \le |b^s_{(r)i}| < .50] \tag{28}$$

$$B1_{(r)} = (1/(50)(14)) \sum_{i,s} I[.50 \le |b^s_{(r)i}| < 1.0] \tag{29}$$

$$B2_{(r)} = (1/(50)(14)) \sum_{i,s} I[1.0 \le |b^s_{(r)i}| < 2.0] \tag{30}$$

$$B3_{(r)} = (1/(50)(14)) \sum_{i,s} I[2.0 \le |b^s_{(r)i}| < 3.0] \tag{31}$$

Using these measures, for each procedure p in data set r we predict $\mathcal{A}^{(p)}_{(r)}$ with the estimator

$$\tilde{A}^{(p)(\delta)}_{(r)} = \tilde{\delta}^{(p)}_0 + \tilde{\delta}^{(p)}_1 \log(\kappa_{(r)}) + \tilde{\delta}^{(p)}_2 B.5_{(r)} + \tilde{\delta}^{(p)}_3 B1_{(r)} + \tilde{\delta}^{(p)}_4 B2_{(r)} + \tilde{\delta}^{(p)}_5 B3_{(r)}$$
$$+ \tilde{\delta}^{(p)}_6 \log(\hat{R}_{(r)}) \tag{32}$$

where the $\tilde{\delta}^{(p)}$ are estimated by ordinary least-squares using the data from the 270 replications. Including quadratic terms of or interactions among the predictors in (32) does not improve the predicted $\mathcal{A}^{(p)}_{(r)}$s significantly.

To evaluate these estimators, we simulate 100 additional data sets using the design in Table 3. For each procedure p, let $\mathcal{J}^{(p)}$ be the subset of these 100 replications for which procedure p does not select the full models in all groups. For each $r \in \mathcal{J}^{(p)}$, we compute the difference between the true $\mathcal{A}^{(p)}_{(r)}$ and the predicted $\mathcal{A}^{(p)}_{(r)}$

$$\epsilon^{(p)(\delta)}_{(r)} = \mathcal{A}^{(p)}_{(r)} - \tilde{A}^{(p)(\delta)}_{(r)} \tag{33}$$

For each r not in $\mathcal{J}^{(p)}$, we observe that $\mathcal{A}^{(p)}_{(r)} = 1$, so that there is no need to use the estimator in (32). Summing over the $r \in \mathcal{J}^{(p)}$, we compute the root mean squared error,

$$\sqrt{MSE^{(p)(\delta)}} = \sqrt{(1/|\mathcal{J}^{(p)}|) \sum_{r \in \mathcal{J}^{(p)}} (\epsilon^{(p)(\delta)}_{(r)})^2} \tag{34}$$

Values of the root mean squared error near zero indicate that the estimator accurately predicts average relative squared prediction errors.

We use (32) to predict the average relative squared prediction errors for the HMSPs with the smallest $\bar{\mathcal{A}}^{(p)}$s, namely HLP-ED, HLS-BIC, HLD-MSE, HLU-ED, HGP-MSE,

Table 5. Summary of the $\epsilon_{(r)}^{(p)(\delta)}$ for selected procedures

Procedure	$\sqrt{MSE}^{(p)(\delta)}$	Median $(|\epsilon_{(r)}^{(p)(\delta)}|)$	max $(|\epsilon_{(r)}^{(p)(\delta)}|)$	Number of $|\epsilon_{(r)}^{(p)(\delta)}|$	
				$> .05, < .10$	$> .10$
MAIC	.04	.03	.11	15	1
MBIC	.15	.04	1.17	31	12
HLP-ED	.03	.02	.09	15	0
HLS-BIC	.03	.02	.07	9	0
HLD-MSE	.03	.02	.10	9	1
HLU-ED	.02	.01	.04	0	0
HGP-MSE	.04	.03	.16	21	2
HGS-BIC	.05	.03	.09	9	0
HGD-MSE	.04	.02	.16	10	3

Note: Results based on replications with $\mathcal{A}_{(r)}^{(p)} \neq 1$.

HGS-BIC, and HGD-MSE. For comparisons, we also use (32) to predict the average relative squared prediction errors for MAIC and MBIC, although we set $\tilde{\delta}_6^{(p)} = 0$ since across-group information is unrelated to the performances of MAIC and MBIC. Summaries of the absolute errors for the predicted $\mathcal{A}_{(r)}^{(p)}$s are presented in Table 5. For the HMSPs, all root mean squared errors are .05 or less, and all median absolute errors are .03 or less. Also, for each HMSP, at least 75% of the absolute errors are smaller than .05, and absolute errors larger than .10 are rare. Thus, if we use (32) to predict the HMSPs' $\mathcal{A}_{(r)}^{(p)}$ in some observed data set with characteristics like those of these simulated data sets, the predictions are likely to be within .05 of the true values.

The maximum $|\epsilon_{(r)}^{(p)(\delta)}|$ for the HMSPs are relatively large, indicating that (32) inaccurately predicts $\mathcal{A}_{(r)}^{(p)}$s for some data sets. Five of the six predictions with $|\epsilon_{(r)}^{(p)(\delta)}| > .10$ have $\mathcal{A}_{(r)}^{(p)} < \tilde{A}_{(r)}^{(p)} < .90$ and $B.5_{(r)} > 8$. Thus, the largest absolute errors tend to occur when the true $\mathcal{A}_{(r)}^{(p)}$s are small and many estimated coefficients are close to zero. Such errors are not especially problematic since the $\tilde{A}_{(r)}^{(p)}$ suggest reductions in $\mathcal{A}_{(r)}^{(p)}$s of at least .10, which is conservative relative to the true reductions. A large, positive error occurs in only one out of 100 replications. Thus, we can feel confident that the estimator in (32) for the HMSPs rarely underestimates $\mathcal{A}_{(r)}^{(p)}$s by more than .10.

Once predicted $\mathcal{A}_{(r)}^{(p)}$s have been obtained, we then decide whether the estimated gains in accuracy are large enough to argue for the use of the medial information pooling procedures. To demonstrate how these predictions can be used to assist such decisions, I adopt a simple decision rule for the study: we use Predictor Exclusion procedures when they reduce $\mathcal{A}_{(r)}^{(p)}$ by at least .10, and we use Stepwise/Full Selection, Dimension Selection, and Unit Exclusion procedures when they reduce $\mathcal{A}_{(r)}^{(p)}$ by at least .05. This decision rule is only for illustrations; it is not meant in any way as a recommended rule. For the demonstration, we examine three of the 100 evaluation replications: r_{min}, which is the replication for which $\mathcal{A}_{(r)}^{(MBIC)}$ is minimized, r_{max}, which is the replication for which $\mathcal{A}_{(r)}^{(MBIC)}$ is maximized, and r_{rand}, which is randomly chosen from the 100 replications. The actual and predicted $\mathcal{A}_{(r)}^{(p)}$s for the three replications are shown in Table 6.

In r_{min}, after fitting the HNRM we find that $B.5 \approx 13$, so that there is an increased potential that the $\tilde{A}_{(r)}^{(p)(\delta)}$s will overestimate the true $\mathcal{A}_{(r)}^{(p)}$s for the HMSPs. As the table

Table 6. *Predictions for selected procedures for r_{min}, r_{max}, and r_{rand}*

Procedure	$r = r_{min}$		$r = r_{max}$		$r = r_{rand}$	
	$\mathcal{A}_{(r)}^{(p)}$	$\tilde{\mathcal{A}}_{(r)}^{(p)(\delta)}$	$\mathcal{A}_{(r)}^{(p)}$	$\tilde{\mathcal{A}}_{(r)}^{(p)(\delta)}$	$\mathcal{A}_{(r)}^{(p)}$	$\tilde{\mathcal{A}}_{(r)}^{(p)(\delta)}$
MAIC	.82	.87	1.07	1.06	.98	1.00
MBIC	.77	.80	2.26	1.09	.98	.99
HLP-ED	.70	.77	.99	.99	.78	.83
HLS-BIC	.79	.86	1.00	.99	.92	.98
HLD-MSE	.79	.83	1.01	1.01	.94	.98
HLU-ED	.91	.95	.99	.98	.94	.95
HGP-MSE	.67	.77	1	1	.88	.90
HGS-BIC	.77	.85	1	1	1	1
HGD-MSE	.73	.84	1	1	.93	1.00

Notes: $r_{min} = \arg\min (\mathcal{A}_{(r)}^{(\text{MBIC})})$, $r_{max} = \arg\max (\mathcal{A}_{(r)}^{(\text{MBIC})})$.

shows, this is exactly what happens. Despite these errors, the predicted values still indicate that the HMSPs reduce $\mathcal{A}_{(r)}^{(p)}$s substantially. The estimators also correctly predict substantial reductions in $\mathcal{A}_{(r)}^{(p)}$s from MAIC and MBIC, which in this replication are dominated only by the Predictor Exclusion procedures. Based on the predicted $\mathcal{A}_{(r)}^{(p)}$s in this data set and our decision rules, we would use any of the medial information pooling procedures over FULL in this data set.

In r_{max}, there are practically no payoffs to using medial information pooling instead of FULL. In fact, all the global procedures select the full model in each group. Appropriately, the predicted $\mathcal{A}_{(r)}^{(p)}$s for all medial information pooling procedures are near one. The predictions also reveal that MAIC and MBIC worsen estimates, although the estimates for MBIC underestimate the extremely high $\mathcal{A}_{(r)}^{(\text{MBIC})}$. Based on the predicted $\mathcal{A}_{(r)}^{(p)}$s in this data set and our decision rules, we would not use the medial information pooling procedures in this data set.

In r_{rand}, the estimators predict moderate reductions in $\mathcal{A}_{(r)}^{(p)}$s for all procedures except the Predictor Exclusion procedures, which have relatively large predicted reductions in $\mathcal{A}_{(r)}^{(p)}$. The predictions tend to be overestimates, but they are all within .06 of the true values. Based on the predicted $\mathcal{A}_{(r)}^{(p)}$s in this data set and our decision rules, we would use HLP-ED, HLU-ED, and HGP-MSE. But, the predicted reductions for the other procedures are not large enough under our decision rules to justify using them over FULL, even though the actual $\mathcal{A}_{(r)}^{(p)}$s are reduced in some of these procedures by more than .05.

Although the estimator in (32) typically yields predictions within .05 of the true $\mathcal{A}_{(r)}^{(p)}$s in these simulations, it may be unreliable for data sets with characteristics that differ from those of this study's design. However, the process of replicating data sets, developing and evaluating estimators, and utilizing such estimators that is demonstrated in this study is recommended for similar studies with other data characteristics.

6. Additional Comments on Smoothing Settings

In smoothing settings, the ultimate goal is to estimate a population quantity, such as a mean or total, in each of several domains that are common to each of many groups.

From data on sampled units, each domain's population quantity can be directly esti-
mated using only the data from that domain. There also exist data on domain-level pre-
dictors that are potentially relevant for estimating the population quantities. Using these
predictors, a separate model for each group is built that is used to smooth the direct
estimates.

The procedures from Section 4 can be adapted for smoothing settings by letting the $\mathcal{Y}_i$
be $m \times 1$ vectors of population quantities, the $\mathbf{Y}_i$ be $m \times 1$ vectors of direct estimates of
these quantities, and the $\mathcal{X}_i = \mathbf{X}_i$ be $m \times |\mathsf{F}|$ matrices of values of domain-level predictors.
As an example of this notation, consider estimating coverage factors, like those from the
proposed design of the 2000 U.S. census, for a hypothetical country with the same $m = 10$
domains in each state i of $\mathcal{G} = 50$ states. Estimate the 10×1 vector of true coverage
factors, $\mathcal{Y}_i$, by smoothing the direct estimates of the coverage factors, $\mathbf{Y}_i$, using $p = 3$
variables, $\mathcal{X}_i$ (e.g., an intercept, number of houses in the domain, and mail return rates
in the domain). When there are no predictors available other than domain indicators,
each $\mathcal{X}_i$ can be composed of the same $p = m$ dummy variables corresponding to the
domains. The $\mathbf{Y}_i$ can be smoothed by setting one or more of the estimated coefficients
of the main effects and interactions in $\tilde{\boldsymbol{\beta}}_i$ to zero.

Another method of estimating $\mathcal{Y}_i$ is to use some weighted average of the direct estimate,
$\mathbf{Y}_i$, and the $\tilde{\mathcal{Y}}_i$ in (3):

$$\tilde{\mathcal{Y}}_i^* = \mathbf{W}_i \mathbf{Y}_i + (\mathbf{1} - \mathbf{W}_i) \tilde{\mathcal{Y}}_i \tag{35}$$

where $\mathbf{W}_i$ is an $m \times 1$ vector of weights (Ghosh and Rao 1994). Since the accuracy of (35)
depends on the accuracy of $\tilde{\mathcal{Y}}_i$, we may be able to improve $\tilde{\mathcal{Y}}_i^*$ by using medial information
pooling procedures to determine $\tilde{\mathcal{Y}}_i$. Once $\tilde{\mathcal{Y}}_i$ is determined, the weights then can be
specified.

A related method for estimating the $\mathcal{Y}_i$ is to use the modal estimates of the $\mathcal{Y}_i$ from
within-group hierarchical models, such as

$$\mathbf{Y}_i \sim N(\mathcal{Y}_i, \Sigma_i) \qquad \mathcal{Y}_i \sim N(\mathcal{X}_i^{\mathsf{S}} \boldsymbol{\beta}_i^{\mathsf{S}}, \Sigma_{\mathcal{Y},i}^{\mathsf{S}}) \tag{36}$$

where Σ_i is estimated from resampling techniques (Wolter 1985) but treated as known, and
$\Sigma_{\mathcal{Y},i}^{\mathsf{S}}$ is estimated from the data. Medial information pooling can be used to select the set S
for each group's hierarchical model. Such models can use a three-level hierarchical model
for mining across-group information:

$$\mathbf{Y}_i \sim N(\mathcal{Y}_i, \Sigma_i) \qquad \mathcal{Y}_i \sim N(\mathcal{X}_i \boldsymbol{\beta}_i, \Sigma_{\mathcal{Y},i}) \qquad \boldsymbol{\beta}_i \sim N(\boldsymbol{\mu}, \Sigma_{\boldsymbol{\beta}}) \tag{37}$$

where $|\mathsf{F}| < m$. To do so, we let $\hat{\mathbf{Y}}_i^{\mathsf{F}}$ be the modal estimates of $\mathcal{Y}_i$ from (37), and, for each
set $\mathsf{S} \neq \mathsf{F}$ under consideration in the procedures' criteria, we let $\tilde{\mathbf{Y}}_i^{\mathsf{S}}$ be the modal estimates
of $\mathcal{Y}_i$ as determined from (36). For more details, see Reiter (1999).

7.   Concluding Remarks

The investigation of techniques for estimation in the presence of external constraints
that prohibit explicit data pooling has been driven by two questions: ''how much infor-
mation pooling is permissible?'' and ''does using an information pooling strategy reduce
estimation errors by a sufficient amount?'' These two questions cannot be answered

independently. Controversial strategies that reduce errors only slightly are not worth using, particularly when they require an arduous defense of their legality.

To assess procedures' accuracies, I recommend undertaking simulation studies like those of Section 5. ''On average'' studies like the one in Section 5.1 can be especially useful when it is desired, or necessary, to specify the statistical procedures that will be used before actually collecting the data. In settings where statisticians are able to specify the techniques for analyzing the data after collecting the data, they may be able to predict more accurately the medial information pooling procedures' performances from observable data. The process outlined in the study of Section 5.2 is especially relevant in such settings.

These new forms of information pooling have not yet been applied in real settings; thus, I can not offer any evidence of how such procedures are viewed by data users. Nonetheless, I can roughly order the acceptabilities of the procedures based on how directly the procedures rely on multiple groups' data when specifying each group's separate model. Within a strategy, I expect that all procedures are equally acceptable. It is also likely that the strategies' ordering is the same given a local or global scope although one scope may be more acceptable than another in a particular setting. I expect the ordering of the strategies' acceptabilities from most to least acceptable to be Dimension Selection, Stepwise/Full Selection, and Predictor Exclusion, with Unit Exclusion falling somewhere among these strategies. In conjunction with the ordering of the strategies' accuracies from Section 5.1, this confirms the hypothesis that potential users can expect an inverse relationship between accuracy and acceptability when considering these medial information pooling strategies.

Of course, the worth of medial information pooling procedures in a particular setting depends on the constraints in that setting. In a sense, the role of statisticians when confronted with legal constraints is to bring information pooling strategies to the table and work with the constraint-makers to find an acceptable strategy that increases accuracy. As shown in this article, medial information pooling procedures can substantially increase accuracy relative to full, separate regressions and traditional, stepwise model selection procedures, so that they are worth bringing to the table as potential solutions to the constrained estimation dilemma.

8. References

Akaike, H. (1974). A New Look at the Statistical Identification Model. IEEE Transactions on Automatic Control, 19, 716–723.

Bell, W.R. and Hillmer, S.C. (1990). The Time Series Approach to Estimation for Repeated Surveys. Survey Methodology, 16, 195–215.

Belsley, D.A. (1990). Conditioning Diagnostics: Collinearity and Weak Data in Regression. New York: John Wiley & Sons.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977a). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1–38.

Dempster, A.P., Schatzoff, M., and Wermuth, N. (1977b). A Simulation Study of Alternatives to Ordinary Least Squares (with discussion). Journal of the American Statistical Association, 72, 77–106.

Fairley, W.B., Izenman, A.J., and Bagchi, P. (1990). Inference for Welfare Quality Control Programs. Journal of the American Statistical Association, 85, 874–890.

Fay, R.E. and Thompson, J. (1993). The 1990 Post Enumeration Survey: Statistical Lessons, in Hindsight (with discussion). Proceedings of the U.S. Bureau of the Census Annual Research Conference, 9, 71–96.

Furnival, G.M. and Wilson Jr., R.W. (1974). Regression by Leaps and Bounds. Technometrics, 16, 499–511.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1996a). Bayesian Data Analysis. New York: Chapman & Hall.

Gelman, A., Meng, X.L., and Stern, H. (1996b). Posterior Predictive Assessment of Model Fitness (with discussion). Statistica Sinica, 6, 733–807.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal (with discussion). Statistical Science, 9, 55–93.

Griffin, R. and Vacca, E.A. (1998). Estimation in the Census 2000 Dress Rehearsal. Proceedings of the Survey Research Methods Section of the American Statistical Association, 635–641.

Hansen, M.H. and Tepping, B.J. (1990). Regression Estimates in Welfare Quality Control Programs (with discussion). Journal of the American Statistical Association, 85, 856–873.

Hodges, J.S. (1998). Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics (with discussion). Journal of the Royal Statistical Society, Series B, 60, 497–536.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 12, 68–82.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. Journal of the American Statistical Association, 88, 1047–1060.

Huber, P.J. (1981). Robust Statistics. New York: John Wiley & Sons.

Kramer, F.D. (1990). Statistics and Policy in Welfare Quality Control: A Basis for Understanding and Assessing Competing Views. Journal of the American Statistical Association, 85, 850–855.

Miller, A.J. (1990). Subset Selection in Regression. New York: Chapman & Hall.

Morris, C.N. (1983). Parametric Empirical Bayes Inference: Theory and Applications (with discussion). Journal of the American Statistical Association, 78, 47–65.

Mulry, M.H. (1996). Comment on a Paper by Kadane. Journal of Official Statistics, 12, 101–104.

Raftery, A.E. (1995). Bayesian Model Selection in Social Research (with discussion). In Sociological Methodology, 1995, 78, ed. P.V. Marsden, 111–195, Cambridge, MA: Blackwell Publishers.

Reiter, J.P. (1999). Estimation in the Presence of Constraints that Prohibit Explicit Data Pooling. Ph.D. thesis, Department of Statistics, Harvard University.

Rubin, D.B. (1980). Using Empirical Bayes Techniques in the Law School Validity Studies. Journal of the American Statistical Association, 75, 801–825.

Rubin, D.B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. The Annals of Statistics, 12, 1151–1172.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6, 461–464.

Thompson, J.H. and Fay, R.E. (1998). Census 2000: The Statistical Issues. Proceedings of the Survey Research Methods Section of the American Statistical Association, 101–110.

United States Bureau of the Census (1997). Census 2000 Operational Plan.

United States Bureau of the Census (1999). Census 2000 Operational Plan: Using Traditional Census-Taking Methods.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Zaslavsky, A.M. (1993). Combining Census, Dual-system, and Evaluation Study Data to Estimate Population Shares. Journal of the American Statistical Association, 88, 1092–1105.