Journal of Computer-Aided Molecular Design (2005) 19: 739–747
DOI 10.1007/s10822-005-9011-5

Secure analysis of distributed chemical databases without data integration

Alan F. Karr[a,*], Jun Feng[a], Xiaodong Lin[a], Ashish P. Sanil[a], S. Stanley Young[a] & Jerome P. Reiter[b]
[a]*National Institute of Statistical Sciences Research, Triangle Park, NC 27709-4006, USA;* [b]*Duke University, Durham, NC 27708, USA;* [c]*University of Cincinnati, Cincinnati, OH USA;* [d]*Bristol-Myers Squibb, Princeton, NJ USA*

Received 17 May 2005; accepted 28 July 2005

Key words: chemical database, distributed data, regression model, secure multi-party computation

Summary

We present a method for performing statistically valid linear regressions on the union of distributed chemical databases that preserves confidentiality of those databases. The method employs *secure multi-party computation* to share local sufficient statistics necessary to compute least squares estimators of regression coefficients, error variances and other quantities of interest. We illustrate our method with an example containing four companies' rather different databases.

Introduction

Many scientific investigations require statistical analyses that 'integrate' data stored in multiple, distributed databases. For example, a regression analysis on integrated chemical databases to identify molecular features influencing biological activity would be more insightful than individual analyses. At the same time, the barriers to actually integrating the databases are numerous. In the setting of this paper,[1] the proprietary nature of the data is the principal impediment to integration. Scale is another barrier: despite advances in networking technology, the only way to move a terabyte of data from point A today to point B tomorrow may be FedEx.

The good news is that for many analyses it is not necessary to move or share individual data records. Instead, using techniques from computer science known generically as *secure multiparty computation* [1, 2] the participating organizations – we term them 'companies' – can share summaries of the data anonymously, but in a way that the analysis can be performed in a statistically valid manner.

In this paper, we illustrate linear regression on 'horizontally partitioned' data, in which each company's database contains the same chemical descriptors for its own set of molecules. The need for protecting descriptor values is apparent: given the method of descriptor calculation and descriptor values, it is easy to guess structures through similarity searching over a large database. The basis of the method is one particular protocol for secure multi-party computation – that of secure summation, which is discussed in section 'Secure summation.'

Problem formulation

We assume that there are $K > 2$ companies, each with the same numerical descriptors on its own n_j compounds – p predictors X^j (in the example in section 'Example,' molecular descriptions) and a response y^j (in the example, water solubility), and

*To whom correspondence should be addressed. E-mail: karr@niss.org

that the companies wish to fit the usual linear model

$$y = X\beta + \epsilon, \tag{1}$$

to the 'global' data

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y^1 \\ \vdots \\ y^K \end{bmatrix}$$

Each X^j is $n_j \times p$. Horizontal data partitioning for $K = 3$ companies is illustrated in section 'Secure regression'.

We embed the constant term of the regression in the first predictor: $X_1^j \equiv 1$for all$j$. To illustrate the subtleties of analysis of distributed data, the alternative strategy of centering the predictors and response at mean values does not work, at least not directly. The means in this case are the global means, which are not available, but could be calculated with another round of secure computation.

Under the condition that $\mathrm{Cov}(\epsilon) = \sigma^2 I$, the least squares estimator for β is

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{2}$$

In section 'Secure regression' we show how $\hat{\beta}$ can be computed without integrating the companies' databases.

Several assumptions about company behavior are necessary. First, the companies agree to cooperate to perform the regression, and none of them is specifically interested in breaking the confidentiality of the others' data. Second, each company reports accurately the results of computations on its own data, and follows the agreed-on computational protocols, such as secure summation, properly. Finally, there is no collusion among companies. (Otherwise, for example, if there were three companies participating, two could collude to learn about the data of the third. The protocol in section 'Secure regression' still prevents them from learning specifics of individual molecules.) We summarize these assumptions by saying that the companies are *semi-honest*.

In addition, the formulation assumes explicitly that the columns of the combined data matrix X in Equation 1 be comparable across companies. In principle, this means that the companies all use the same predictors calculated in the same way. To the extent that predictors are calculated differently by different companies, the analysis may be compromised. However, the degree of compromise is neither increased nor decreased by our approach as compared to literal integration of the data.

Secure summation

The simplest secure multi-party computation, and the only one needed for secure regression, is to sum values v_j held by the companies. Let v denote the sum. The secure summation protocol described below computes v in a way that no company j can learn more than the minimum possible about the other companies values – the sum $v_{(-j)} = \Sigma_{l \neq j} v_l = v - v_j$. The secure summation protocol, which is depicted graphically in Figure 1, is straightforward in principle, although a 'production quality' implementation presents challenges. Number the companies 1,..., K. Company 1 generates a very large random integer R, adds R to its value v_1, and sends the sum to company 2. Since R is random, company 2 learns effectively nothing about v_1. Company 2 adds its value v_2 to $R + v_1$, sends the result to company 3, and so on. Finally, company 1 receives $R + v_1 + \ldots + V_K = R + v$ from company K, subtracts R, and shares the result v with the other companies. Here cooperation matters: company 1 is obliged to share v with the other companies.

Figure 1 contains an extra layer of protection. Suppose that v is known to lie in the range $[0, m)$, where m is a very large number, say 2^{100}, known to all the companies. Then R can be chosen randomly from $\{0,\ldots, m-1\}$ and all computations performed modulo m.

Here is a simple application: the companies have molecular weight data and wish to compute the global average weight of their molecules. Let n_j be the number of records in company j's database and W_j be the sum of their molecular weights. The quantity to be computed is $\bar{W} = \Sigma_i W_j / \Sigma_j n_j$. The numerator $\Sigma_j W_j$ can be computed using secure summation on the W_j's, and whose denominator $\Sigma_j n_j$ can be computed using secure summation on the n_j's. Note that no company can learn weights of any other company's individual molecules.

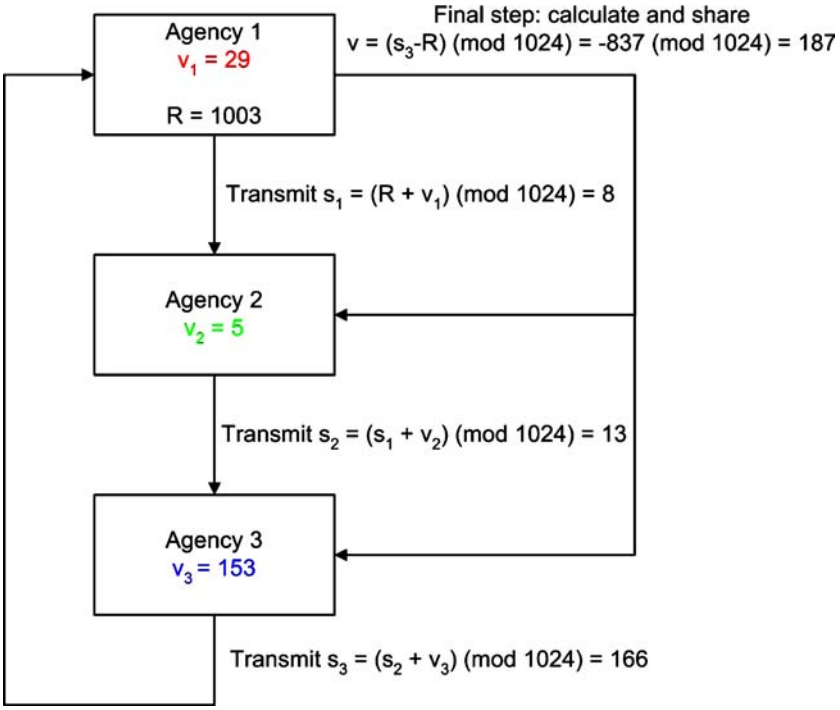

Figure 1. Values computed at each company during secure computation of a sum initiated by company 1. Here $v_1 = 29$, $v_2 = 5$, $v_3 = 152$ and $v = 187$. All arithmetic is modulo $m = 1024$.

Secure regresion

In this section, we show how to perform secure regression for horizontally data.

To compare $\hat{\beta}$ using Equation 2, it is necessary to compute X^TX and X^Ty. Because of the horizontal partioning of the data, these are additive over the companies:

$$X^TX = \sum_{j=1}^{K} (X^j)^T X^j$$

Therefore, company j simply computes its own $(X^j)^T X^j$, which has dimensions $p \times p$, where p is the number of predictors, and these are combined entrywise using secure summation. The protocol is illustrated with $K = 3$ in Figure 2. Of course, because of symmetry, only $\binom{p}{2} + p$ secure summations are needed. Similarly, X^Ty can be computed by secure, entry-wise summation of the $(X^j)^T y^j$.

Finally, each company can calculate $\hat{\beta}$ from the shared values of X^TX and X^Ty using Equation 2. Note that no company learns any other company's $(X^j)^T X^j$ or $(X^j)^T y^j$, but only the sum of these over all the other companies. It does learn this sum, exactly; however, technology under development at the National Institute of Statistical Sciences (NISS) (Karr et al. 2005b, in preperation), removes this incentive to 'cheat.'

Model diagnostics are used by statisticians to assess the applicability of the linear model in Equation 1. The simplest diagonostic is coefficient of determination R^2, which measures the over-all 'fit' of the model. More sophisticated diagonostics, which are typically based on the residuals – differences between actual data values and predictions

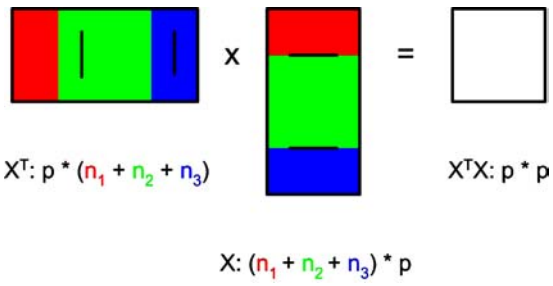

Figure 2. Pictorial representation of the secure regression protocol. The dimensions of various matrices are shown. As in the example in section 'example', two (in this case) of the companies cannot even perform the regression because there are more predictors than records in their databases.

from the model, can reveal other forms of model mis-specification. In order for secure regression to be truly useful, therefore, diagnostics need to be available as well. One approach is to use diagnostics that can be computed using secure summation from corresponding local statistics. A second approach uses 'secure data integration' [3, 4] to share synthetic residuals [5].

Among diagnostics computable by secure summation are R^2 itself, the least square estimate $S^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/(n - p)$ of the error variance σ^2, correlations between predictors and residuals, and the so called hat matrix $H = X(X^T X)^{-1}X^T$. The latter can be used to identify X-outliers.

For diagnosing some types of assumption violations, only *patterns* in relationships among the residuals and predictors suggestive of model mis-specification are needed, rather than exact values of the residuals and predictors. (Sharing exact residuals, of course, is tantamount to sharing the data themselves.) An approach to secure sharing of residuals is outlined in [3], but such diagnostics have not yet been implement.

Example

We illustrate the secure regression protocol for horizontally partitioned data using a data set containing water solubility of 1318 organic compounds as a function of an intercept and 90 X log P atom types [6]. Details of the descriptor are in Ref. 7; their names appear in Tables 2–4 in the appendix. Figure 3 shows visualizations of several of the compounds produced using NISS' PowerMV software [8].

To simulate distributed data, the database was split, using the clustering algorithm in JMP [9], into four subsets corresponding to companies 1,...,4 and containing 499, 572, 16 and 231 compounds, respectively. The effect of the clustering is that there are several descriptors for which only one company has data.

Table 1 summarizes some characteristics of the global regression – for all four companies – and the four companies' individual regressions. Tables 2, 3, 4, 5 contain the full sets of estimated coefficients. Company 3, of course, cannot even perform the regression on its own, and so is omitted from Tables 2–5.

Tables 2–5 are not easy to digest. Figure 4 contains scatterplots of the regression coefficients for companies 1, 2 and 4 (y-axis) against those for the global regression (x-axis). Although there are clear relationship between the regression for companies 1, 2 and 4 global regression, there are also substantial differences. In particular, each company receives global coefficient for descriptors not present in its own data, whose y-values in Figure 4 are zero.

Not surprisingly, the extent to which the one-company regression resemble the global regression

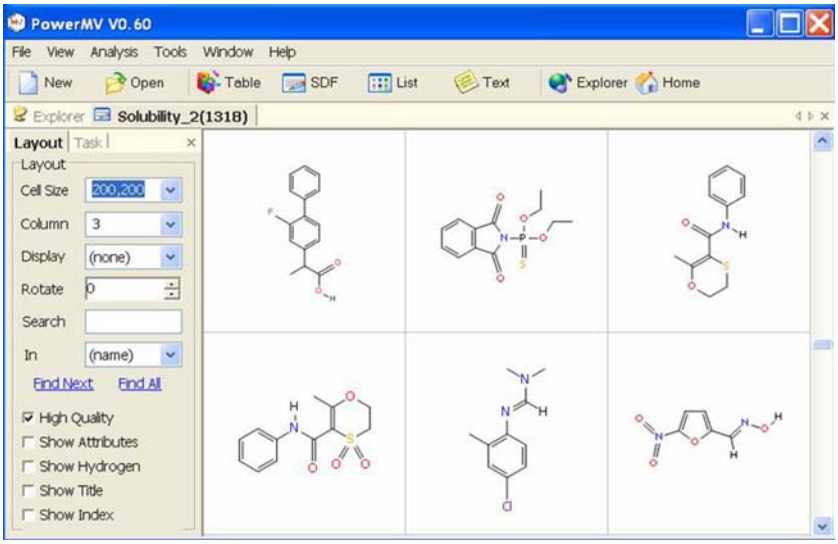

Figure 3. Representative compounds in the database analyzed in Ref. 6, visualized using PowerMV [8].

Table 1. Summary statistics for the global (4-company) regression and the individual regressions for companies 1, 2 and 4. Company 3 does not possess sufficient data to perform the regression on its own.

Regression	R2	RMS error
Global	0.876	0.717
Company 1	0.806	0.647
Company 2	0.869	0.688
Company 3	n/a	N/a
Company 4	0.936	0.573

is a function of the relative sizes of their databases. Thus, as shown Figure 4, the company 2 regression is the closest to the global regression. Other than company 3, company 4 has the smallest database, and Figure 4 confirms that among the companies 1, 2 and 4 its regression differs most from the global regression.

Given that company 3 only has 16 data points, it is natural to ask whether it is in the interest of companies 1, 2 and 4 to include it in the process. Figure 5 shows a scatterplot of the coefficient in

Table 2. Full regression results for the intercept and molecular descriptors 1–31.

Descriptor		Regression coefficients				
Number	Description (?)	Global (4 companies)	Company 1	Company 2	Company 4	Companies 1, 2, 4
Intercept:		0.2694	1.0262	0.4066	0.2346	0.2237
Sp^3 carbon in						
1	CH_3R ($\pi = 0$)	−0.4193	−0.4889	−0.4993	−0.2678	−0.4419
2	CH_3R ($\pi = 1$)	−0.1270	−0.2896	−0.1920	−0.1370	−0.1922
3	CH_3X	0.2347	−0.2134	0.6251	−0.6557	0.2399
4	CH_2R_2 ($\pi = 0$)	0.4226	0.1654	1.6889	0.2607	0.3375
5	CH_2R_2 ($\pi = 1$)	0.2331	0.3032	0.4665	1.0651	0.3462
6	CH_2R_2 ($\pi = 2$)	0.6835	0.5401	0.9802	1.2322	0.5450
7	$CH_2R_nX_{2-n}$ ($\pi = 0$)	−0.2079	−0.3397	0.3102	0.2660	−0.1544
8	$CH_2R_nX_{2-n}$ ($\pi = 1$)	−0.0095	−0.4834	−0.0877	0.2149	−0.1518
9	$CH_2R_nX_{2-n}$ ($\pi = 2$)	0.0874	−0.0609	0.4871	−1.1743	0.1438
10	CHR_3 ($\pi = 0$)	0.7274	0.6610	1.8204	−0.0830	0.7576
11	CHR_3 ($\pi = 1$)	0.8486	−0.1267	1.1435	0	0.8036
12	CHR_3 ($\pi \geq 2$)	0.0168	−0.0941	0.2511	−0.2050	0.0555
13	CHR_nX_{3-n} ($\pi = 0$)	0.4532	0.1758	0.6223	1.0085	0.2146
14	CHR_2X_{3-n} ($\pi = 1$)	−0.4669	−0.1711	−0.5066	−2.8964	−0.4245
15	CHR_nX_{3-n} ($\pi \geq 2$)	−0.6945	−0.7253	−1.1392	1.2612	−0.7324
16	CR_4 ($\pi = 0$)	−0.2262	−0.1396	−0.2817	−0.3110	−0.2739
17	CR_4 ($\pi = 1$)	−0.2666	0.6324	−0.4612	0.3388	−0.2632
18	CR_4 ($\pi \geq 2$)	0.2566	0.7716	0.9946	0	0.4585
19	CR_nX_{4-n} ($\pi = 0$)	−0.8035	−0.6940	−1.4628	0.0532	−0.8384
20	CR_nX_{4-n} ($\pi > 0$)	−0.5069	−0.4057	−0.6639	−0.7056	−0.4804
sp^2 carbon in						
21	$A{=}CH_2$	−0.3555	−0.2123	−0.7735	−0.8327	−0.4067
22	$A{=}CHR$ ($\pi = 0$)	−0.2350	−0.2870	−0.4190	−0.2811	−0.2693
23	$A{=}CHR$ ($\pi = 1$)	0.3014	0.0549	0.3973	0.4920	0.2839
24	$A{=}CHX$ ($\pi = 0$)	0	0	0	0	0
25	$A{=}CHX$ ($\pi = 1$)	0	0	0	0	0
26	$A{=}CR_2$ ($\pi = 0$)	−0.2688	−0.4086	−0.1087	−0.3213	−0.2662
27	$A{=}CR_2$ ($\pi > 0$)	−0.0030	−0.1326	−0.1991	0.6575	-0.0232
28	$A{=}CRX$ ($\pi = 0$)	−0.6614	−0.6164	−0.8087	−0.5878	−0.6649
29	$A{=}CRX$ ($\pi > 0$)	−0.0421	−0.1081	−0.7344	−0.0593	−0.0515
30	$A{=}CX_2$ ($\pi = 0$)	−0.3793	−0.3315	−0.5330	0.1145	−0.3775
31	$A{=}CX_2$ ($\pi > 0$)	−0.8619	−0.4366	−1.2388	0.3363	−0.8814

Table 3. Full regression results for molecular descriptors 32–58.

Descriptor		Regression coefficients				
Number	Description (?)	Global (4 companies)	Company 1	Company 2	Company 4	Companies 1, 2, 4
Aromatic carbon in						
32	C ···C(H) ···C	−0.5220	−0.6317	−1.4855	0	−0.2949
33	A ···C(H) ···N	−0.7922	−0.9268	−0.0887	−0.3530	−0.8256
34	C ···C(R) ···C	−0.4392	−0.4194	−0.4821	−0.2054	−0.4477
35	C ···C(X) ···C	−0.9804	0	0	−0.2073	−0.9252
36	A ···C(R) ···N	−0.4573	−0.3379	−0.2752	0.3896	−0.3406
37	A ···C(X) ···N	0.0701	−0.0286	0.1256	0.4078	0.0967
sp^2 carbon in						
38	R ≡ CH	0.7334	0.7909	1.2111	0	0.7416
39	A ≡ C–A	0.0458	0.2248	−0.2964	0.5783	−0.0289
40	A =C =A	−0.6505	−0.5747	−0.8719	−0.2133	−0.6514
sp^2 nitrogen in						
41	R–NH_2 (π = 0)	0.0541	0.1540	0.2312	−0.9457	0.0510
42	R–NH2 (π = 1)	−0.1610	−0.0699	−0.6751	0.0890	−0.1422
43	X–NH_2	0.2374	0.0103	0	0	0.2814
44	R–NH–R (π = 0)	−0.2139	0.1575	−0.7090	−0.9228	−0.3354
45	R–NH–R (π > 0)	0.0498	−0.0986	0.4588	−0.2283	0.0743
46	R–NH–R (ring)[c]	−1.1025	−1.1852	−1.0727	−1.6556	−1.1105
47	A–NH–X	0.0415	0.1070	0.1562	−0.5722	0.1589
48	A–NH–X (ring)	−0.7717	0.2356	−0.5285	−1.4208	−0.7800
49	NR_3 (π = 0)	0.3698	0.7583	−1.5967	−1.2539	0.3777
50	NR_3 (π > 0)	−0.2057	−0.1241	−0.1845	0	−0.2054
51	NR_3 (ring)	−0.6170	−0.2850	−0.7948	−0.6501	−0.6125
52	NR_nX_{3-n}	0.0769	−0.6290	−0.4152	0	0.0998
53	NR_nX_{3-n} (ring)	−0.7652	−0.4574	−0.7552	−0.8866	−0.7470
sp^2 Amide nitrogen in						
54	–NH_2	0.2749	−1.2397	−1.4679	0.6689	0.3076
55	–NHR	1.2690	0	0	1.5712	1.2838
56	–NHX	0.1852	0.1885	0.8887	−0.1770	0.2094
57	–NR2	−0.0898	−0.6792	−0.2987	0	−0.1443
58	–NRX	−0.3816	−1.0241	0.1612	0	−0.3032

the regression involving only companies 1, 2, and 4 (y-axis) against those of the global regression (x-axis). While there are minor differences, it is arguable that the participation of company 3 does not change the regression significantly. However, as Table 4 shows, only company 3 has data for descriptor 69, so without company 3, company 1, 2, and 4 would learn nothing about its effect on solubility.

Discussion

In this paper, we have presented a framework for secure linear regression in a cooperative environment. The analysis requires only summaries of the detailed molecular structure information from the companies; there is no sharing of the information about individual molecules. The companies have a strong incentive to participate: they learn more they can know individually about which molecular features contribute to biological activity or physical characteristics without revealing structural details of their own molecules.

A huge number of variations is possible. For example, in order to give the companies flexibility, it may be important to give them the option of withdrawing from the computation when their

Table 4. Full regression results for molecular descriptors 59–80.

Descriptor		Regression coefficients				
Number	Description (?)	Global (4 companies)	Company 1	Company 2	Company 4	Companies 1, 2, 4
sp^2 nitrogen in						
59	C =N–R (π = 0)	0	0	0	0	0
60	C =N–R (π = 1)	−0.9730	0	0	−0.6759	−0.9543
61	C =N–X (π = 0)	0	0	0	0	0
62	C =N–X (π = 1)	−0.1338	0.9293	1.6523	0	−0.0417
63	N =N–R	−1.7453	−1.2866	−2.1384	−2.1590	−1.8283
64	N =N–X	0.2233	0.3427	0.5024	−0.5343	0.2674
65	A–NO	0.8069	0.6865	0	0	0.8786
66	A–NO2	−0.0076	0.3031	0.2646	0.0472	0.3239
Aromatic nitrogen in						
67	A ···N ···A[d]	0.3187	0.1887	0.4713	0.5372	0.3331
Sp nitrogen in						
68	–C ≡ N	0.1069	0.0797	0.0553	0.7148	0.1152
Sp^3 oxygen in						
69	R–OH (π = 0)	0.7056	0	0	0	0.6787
70	R–OH (π = 1)	−0.5039	−0.1300	−0.8018	0	−0.4235
71	X–OH	−0.5311	−0.3763	−0.5469	−0.9567	−0.4903
72	R–O–R (π = 0)	−0.2815	0	−0.3568	0	−0.3980
73	R–O–R (π > 0)	0.3934	0.2139	0.4556	0.7688	0.4391
74	R–O–X	−1.0560	−0.8892	−1.1234	0	−1.0395
sp^2 oxygen in						
75	A =0	−0.3861	−0.2067	−0.2683	0.2039	−0.3140
Sp^3 sulphur in						
76	A–SH	−0.5518	−0.0413	−0.5404	−1.1424	−0.5094
77	A–S–A	1.9727	2.0870	1.5921	0	1.8749
Sp^2 sulphur in						
78	A =S	0.4549	0.4867	0.3994	0.5200	0.3947
Sulfoxide sulfur in						
79	A–SO–A	0.5612	−0.0811	0.3510	0	0.3859
Sulfone sulfur in						
80	A–SO_2–A	0.4882	0	1.2170	−1.3501	0.2167

perceived risk becomes too great. To illustrate, company j may wish to withdraw if its sample size n_j is too large relative to the global sample size n. This is the classical p-rule in the statistical disclosure limitation literature [10]. But, n can be computed using secure summation, and so companies may then 'opt out' according to whatever criteria they wish to employ. It is even possible to allow the opting out itself to be anonymous. The concept of partially trusted third parties (Karr et al. 2005b, in preparation) shows promise in removing incentives for companies not to be semi-honest.

There are other approaches to this problem for lower risk situations. For example, the NISS has developed techniques for secure data integration [3, 4] that build the integrated database in such a way that no company can determine the source of any data elements other than its own, at least under the assumption that the data values themselves do not reveal the source of records. Any

Table 5. Full regression results for molecular descriptors 81–90.

Descriptor		Regression coefficients				
Number	Description (?)	Global (4 companies)	Company 1	Company 2	Company 4	Companies 1, 2, 4
81	O =PA_3	−0.3088	−0.9665	−0.8287	0	−0.4606
82	S =PA_3	−0.2959	−0.2269	−0.4007	−0.0758	−0.2935
Fluorine in						
83	–F (π = 0)	0.0726	−0.0472	0.0926	−0.3672	0.0850
84	–F (π = 1)	−0.6605	−0.5379	−0.6235	−0.7285	−0.6119
Chlorine in						
85	–Cl (π = 0)	−0.3651	−0.5859	−0.1896	−0.4058	−0.3547
86	–Cl (π = 1)	−0.7165	0	−0.7109	0	−0,6620
Bromine in						
87	–Br (π = 0)	−0.6630	−0.9264	0.2759	−0.4676	−0.6492
88	–Br (π = 1)	−1.2125	0	−1.1810	0	−1.1233
Iodine in						
89	–I (π = 0)	−0.1282	−0.5625	0	−0.1240	−0.1025
90	–I (π = 1)	−0.6353	−0.7882	−0.4176	−0.1545	−0.6167

statistical analysis could then be conducted. Of course, however, the point of secure regression is to obviate the need for even a securely integrated database.

There is also technology available to handle *vertically partitioned* databases containing different sets of attributes for the same compounds. This would arise, for example, if each participating company had its own set of chemical descriptors, or different responses that might predict one another, but for the same molecules as the other companies. The goal, again, is to calculate the least square estimators $\hat{\beta}$ of (2).

Strong assumptions are necessary for vertically partitioned data. First, companies must know that they have data on the same subjects, or that there must be a secure method for determining which subjects are common to all their databases. The second assumption is that companies can link records without error. Operationally, this requires in effect that the databases have a common primary key, such as a CAS number.

When only one company holds the response, techniques similar to those in section 'Secure regression' can be used to calculate the 'off-diagonal' blocks of the full data covariance matrix (Karr et al. 2004a, submitted for publication) [4]. However – and in contrast to the horizontally partitioned case – there is loss of protection: if there are n data points, each company's data are known by the other companies to lie in an $n/2$ dimensional space. When all companies hold the response, or the holder of the response is willing to share it, Powell's method for quadratic optimization problems [11] be applied to solve the least squares problem

$$\hat{\beta} = \arg\min_{\beta}(y - X\beta)^T(y - X\beta).$$

The information loss is much less than for the method based on secure matrix products (Sanil et al. 2004, submitted for publication) [4].

Note

1. As compared to the 'official statistics' setting of Karr et al. [3], Sanil et al. (2004, submitted for publication) and Karr et al. (2004a, submitted for publication) and the homeland security setting of Karr et al. [4], where confidentiality of data subjects is paramount.

Acknowledgements

This research was supported by NSF Grant EIA-0131884 to the National Institute of Statistical Sciences (NISS) and by the HighQ Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily

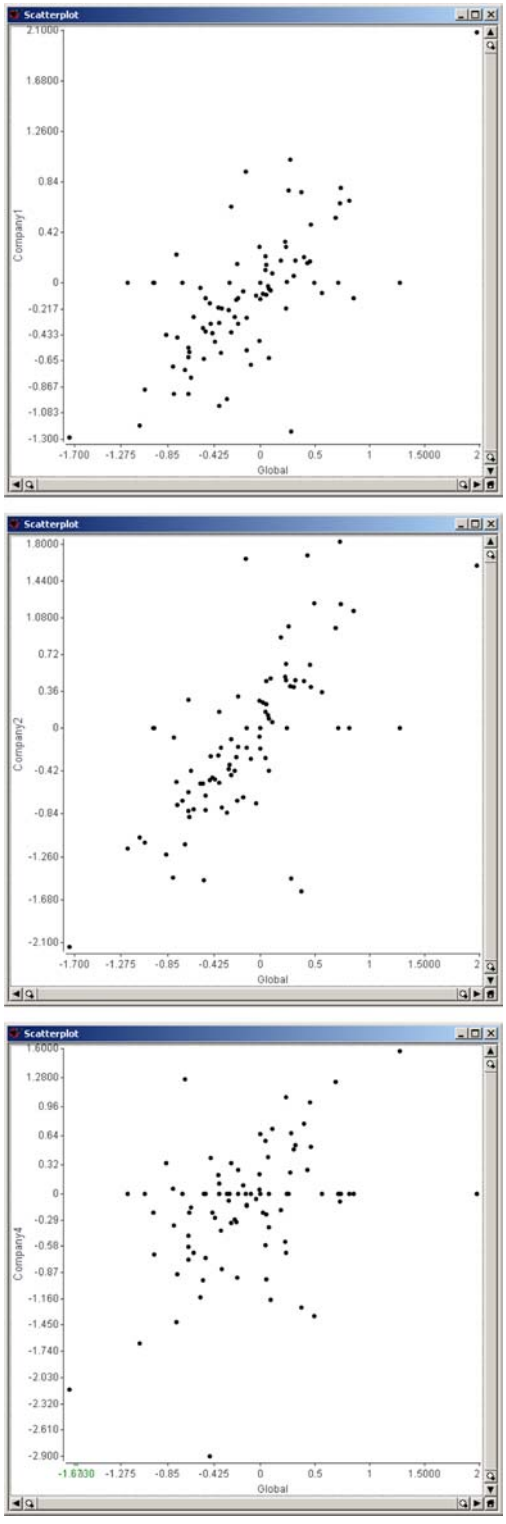

Figure 4. Scatterplots of regression coefficients, including intercept, for companies 1 (top) 2 (center) and 4 (bottom) against those for the global regression. Company coefficients are on the *y*-axes and the global coefficients on the *x*-axes.

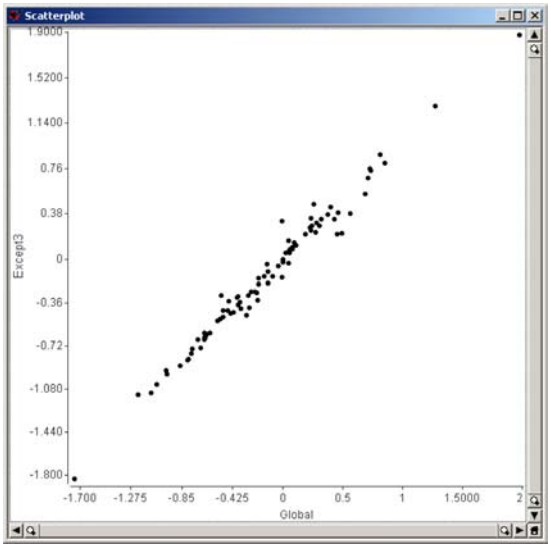

Figure 5. Scatterplot of regression coefficients, including intercept, for the regression involving only companies 1, 2 and 4 against those for the global regression.

reflect the views of the National Science Foundation. The data and structures used in this paper are available at www.niss.org/PowerMV.

References

1. Goldwasser, S., Multi-Party Computations: Past and Present. In Proceedings of the 6th Annual ACM Symposium on Principles of Distributed Computing, ACM Press, New York, 1997, pp. 1–6.
2. Yao, A.C., Protocols for secure computations. In Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, ACM Press, New York, 1982, pp. 160–164.
3. Karr, A.F., Lin, X., Reiter, J.P. and Sanil, A.P., J. Comput. Graph. Stat., (2004b). To appear. Available on-line at www.niss.org/dgii/technicalreports.html.
4. Karr, A.F., Lin, X., Reiter, J.P. and Sanil, A.P., Secure analysis of distributed databases. ASA/SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, 2005a. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.
5. Reiter, J.P., Stat. Comput., 13 (2003) 371.
6. Huuskonen, J., J. Chem. Inf. Comput. Sci., 40 (2000) 773.
7. Wang, R., Gao, Y. and Lai, L., Perspect Drug Discov Design, 19 (2000) 47.
8. Liu, K., Feng, J. and Young, S.S., J. Chem. Inf. Model., 45(2) (2005) 515.
9. SAS Institute, Inc. JMP, the Statistical Discovery Software, 2005. Information available on-line at www.jmp.com.
10. Willenborg, L.C.R.J. and de Waal, T. Elements of Statistical Disclosure Control. Springer-Verlag, New York, 2001.
11. Powell, M.J.D., Comput. J., 7 (1964) 152.