Performance of a distributed regression analysis software and workflow

Qoua L Her, Yury Vilk, Jessica Young, Zilu Zhang, Jessica Malenfant, Sarah Malek, Darren Toh

Department of Population Medicine

Harvard Medical School & Harvard Pilgrim Health Care Institute

August 24, 2018

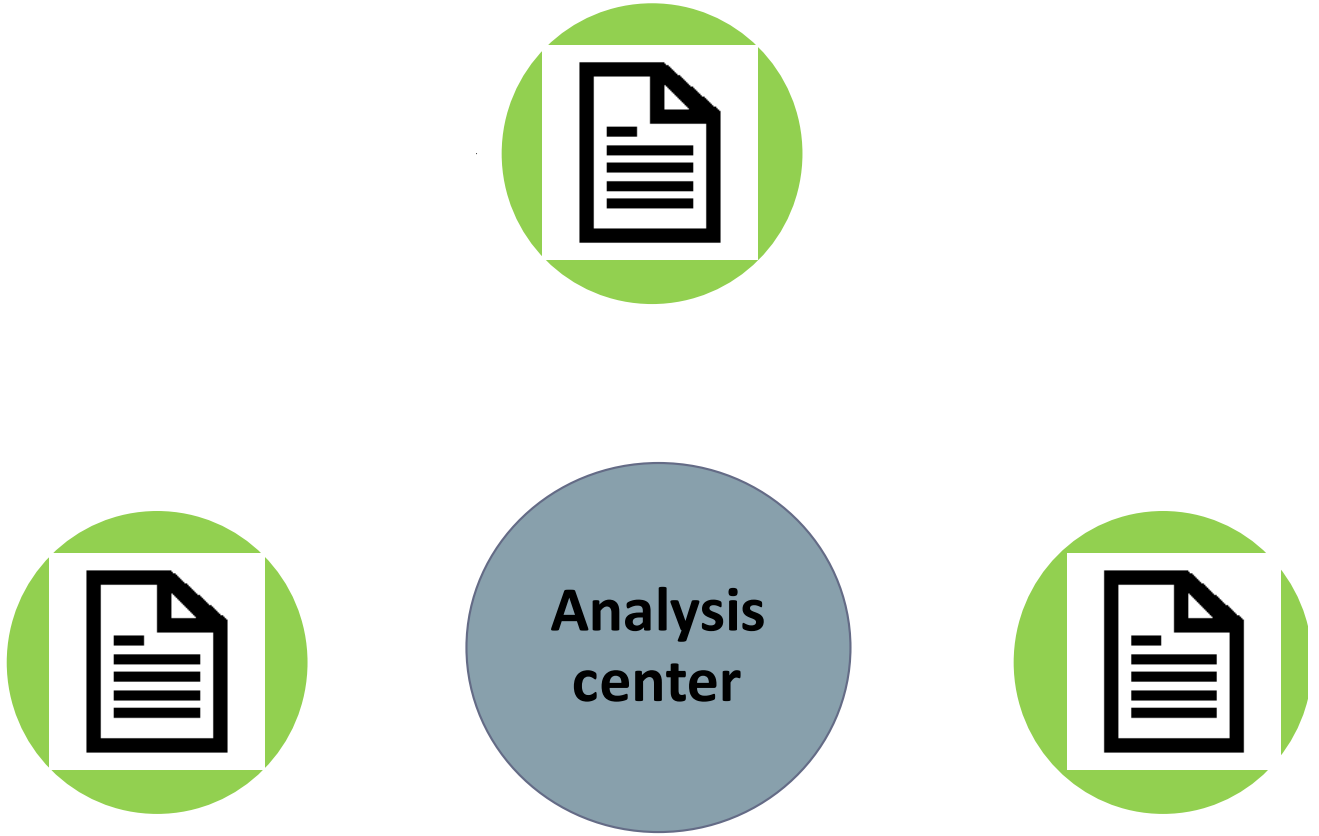
Disclosure

- The authors have no conflicts of interest to disclose
- The project was supported by the Office of the Assistant Secretary for Planning and Evaluation & U.S. Food and Drug Administration (FDA) through the Department of Health and Human Services (HHS) Contract number (HHSF223200910006I)
- This presentation reflects the views of the authors and not necessarily those of the U.S. Food and Drug Administration

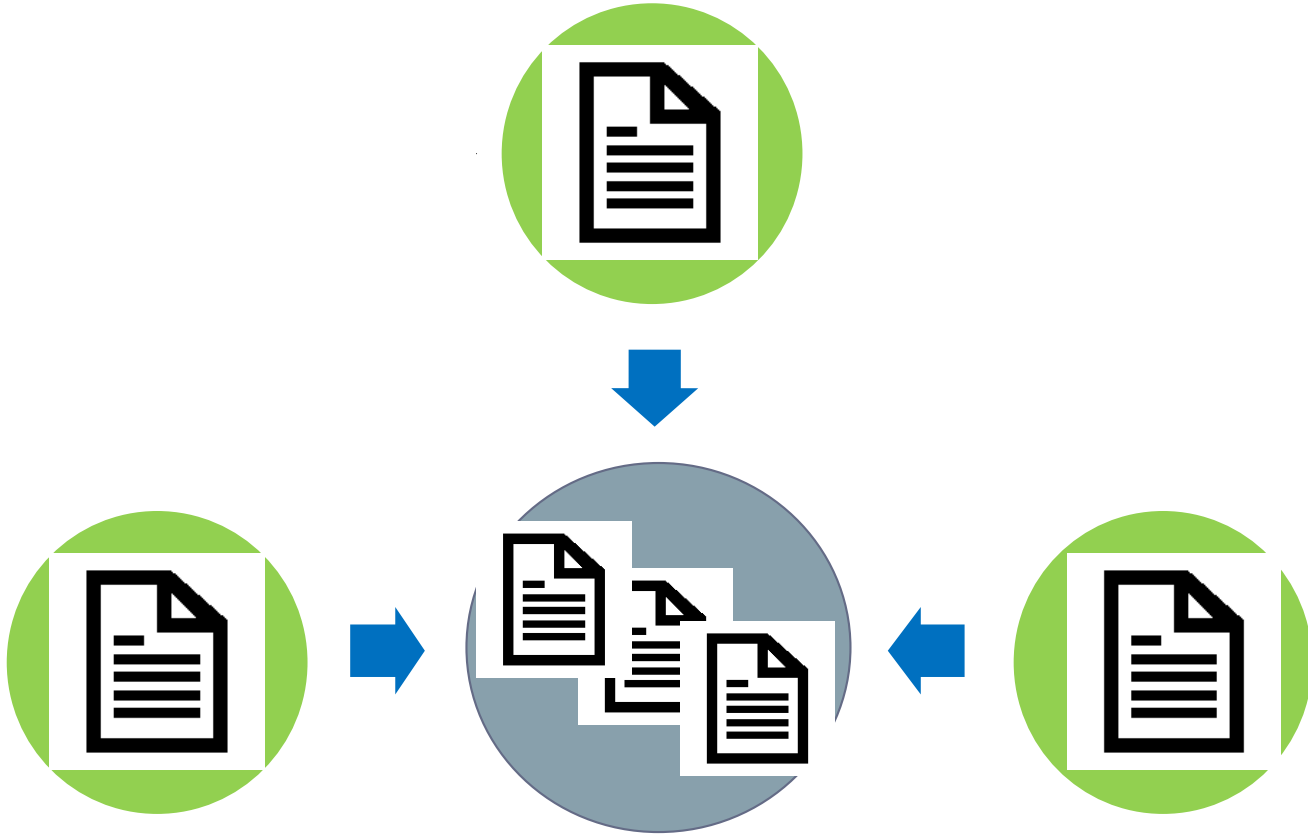
Acknowledgment

- **FDA**
  - Aaron Niman
  - Tyler Coyle
  - Michael Nguyen
- **Data Partners**
  - Kaiser Permanente Colorado
  - Kaiser Permanente Northern California
  - Kaiser Permanente Washington
- **Analysts at Data Partners**
  - Jack Hamilton
  - Ron Johnson
  - David Tabano

Data sharing in conventional multi-center studies



Datasets shared in conventional multi-center studies



Pooling study-specific individual-level datasets

Datasets shared in conventional multi-center studies

PatID	Treatment	Outcome	Age	Sex	Diabetes	CVD	NSAID	…
001	1	0	0	1	0	1	1	…
002	0	0	1	1	0	1	0	…
003	0	0	1	0	0	0	0	…
004	0	0	2	0	1	0	0	…
005	0	1	3	0	0	1	0	…
006	1	1	3	1	0	0	1	…
007	1	0	1	1	1	0	1	…
008	1	0	0	0	1	0	0	…
009	0	1	2	1	0	0	0	…
010	0	0	1	1	0	0	0	…
011	0	0	1	0	0	0	0	…
…	…	…	…	…	…	…	…	…

Datasets shared in conventional multi-center studies

PatID	Treatment	Outcome	Age	Sex	Diabetes	CVD	NSAID	…
001	1	0	0	1	0	1	1	…
002	0	0	1	1	0	1	0	…
003	0							
004	0							
005	0	1	3	0	0	1	0	…
006	1	1	3	1	0	0	1	…
007	1	0	1	1	1	0	1	…
008	1	0	0	0	1	0	0	…
009	0	1	2	1	0	0	0	…
010	0	0	1	1	0	0	0	…
011	0	0	1	0	0	0	0	…
…	…	…	…	…	…	…	…	…

Each row represents an individual

Datasets shared in conventional multi-center studies

PatID	Treatment	Outcome	Age	Sex	Diabetes	CVD	NSAID	…
001	1	0	0	1	0	1	1	…
002	0	0	1	1	0	1	0	…
003	0							
004	0							
005	0	1	3	0	0	1	0	…
006	1	1	3	1	0	0	1	…
007	1	0	1	1	1	0	1	…
008	1	0	0	0	1	0	0	…
009	0	1	2	1	0	0	0	…
010	0	0	1	1	0	0	0	…
011	0	0	1	0	0	0	0	…
…	…	…	…	…	…	…	…	…

Each column represents a variable

Standard approach: pooling individual-level datasets

Data Partner 1

PatID	Exposure	Outcome	Time	X1	X2	X3	X4	X5	…
001	1	0	312	0	M	0	1	1	…
002	1	0	40	1	M	0	2	0	…
003	1	0	365	1	F	0	2	0	…
004	1	0	200	2	F	1	1	0	…
005	0	1	2	3	F	0	3	0	…
006	0	1	15	3	M	0	1	1	…
007	0	0	4	1	M	1	1	1	…
008	0	0	145	0	F	1	3	0	…
009	…	…	…	…	…	…	…	…	…

Data Partner 2

PatID	Exposure	Outcome	Time	X1	X2	X3	X4	X5	…
001	0	1	35	1	F	1	3	0	…
002	0	1	213	2	M	1	1	1	…
003	0	1	453	2	M	0	4	1	…
004	0	0	58	3	M	0	3	1	…
005	1	0	31	3	M	0	3	0	…
006	1	0	56	1	F	1	2	0	…
007	1	0	123	1	F	1	1	1	…
008	1	0	546	0	M	0	3	0	…
009	…	…	…	…	…	…	…	…	…

PatID	Exposure	Outcome	Time	X1	X2	X3	X4	X5	…
001	1	0	312	0	M	0	1	1	…
002	1	0	40	1	M	0	2	0	…
003	1	0	365	1	F	0	2	0	…
004	1	0	200	2	F	1	1	0	…
005	0	1	2	3	F	0	3	0	…
006	0	1	15	3	M	0	1	1	…
007	0	0	4	1	M	1	1	1	…
008	0	0	145	0	F	1	3	0	…
009	…	…	…	…	…	…	…	…	…
001	0	1	35	1	F	1	3	0	…
002	0	1	213	2	M	1	1	1	…
003	0	1	453	2	M	0	4	1	…
004	0	0	58	3	M	0	3	1	…
005	1	0	31	3	M	0	3	0	…
006	1	0	56	1	F	1	2	0	…
007	1	0	123	1	F	1	1	1	…
008	1	0	546	0	M	0	3	0	…
009	…	…	…	…	…	…	…	…	…

Not always possible to pool individual-level datasets

Data Partner 1

PatID	Exposure	Outcome	Time	X1	X2	X3	X4	X5	…
001	1	0	312	0	M	0	1	1	…
002	1	0	40	1	M	0	2	0	
003	1	0	365	1	F	0	2	0	
004	1	0	200	2	F	1	1		
005	0	1	2	3	F	0	3	0	
006	0	1	15	3	M	0	1	1	
007	0	0	4	1	M	1	1	1	…
008	0	0	145	0	F	1	3	0	…
009	…	…	…	…	…	…	…	…	…

Data Partner 2

PatID	Exposure	Outcome	Time	X1	X2	X3	X4	X5	…
001	0	1	35	1	F	1	3	0	…
002	0	1	213	2	M	1	1	1	
003	0	1	453	2	M	0	4	1	
004	0	0	58	3	M	0	3		
005	1	0	31	3	M	0	3	0	
006	1	0	56	1	F	1	2	0	
007	1	0	123	1	F	1	1	1	…
008	1	0	546	0	M	0	3	0	…
009	…	…	…	…	…	…	…	…	…

PatID	Exposure	Outcome	Time	X1	X2	X3	X4	X5	…
001	1	0			M	0	1	1	…
002	1	0			M	0	2	0	…
003	1				F	0	2	0	…
004	1				F	1	1	0	…
005				3	F	0	3	0	…
006				3	M	0	1	1	…
			4	1	M	1	1	1	…
			145	0	F	1	3	0	…
		…	…	…	…	…	…	…	…
		1	35	1	F	1	3	0	…
			213	2	M	1	1	1	…
			453	2	M	0	4	1	…
004				3	M	0	3	1	…
005				3	M	0	3	0	…
006	1				F	1	2	0	…
007	1				F	1	1	1	…
008	1	0			M	0	3	0	…
009	…	…			…	…	…	…	…

Distributed regression

- Regression analysis with data stored at different sites
- Transfer of summary or intermediate statistics only
- Follows the same computation process as conventional individual-level regression analysis
- Results identical to pooled individual-level analysis
- Linear, logistic, Poisson, and Cox model

Karr et al, *J Comput Graph Stat*, 2005;14:263-279

Distributed regression

ID	E	X1	X2	Y
A001	0	13.89	3.42	28.70
A002	1	18.10	1.29	27.90
A003	0	6.41	4.86	33.10
A004	1	16.30	1.45	17.20
A005	1	17.57	2.51	21.70
…	…	…	…	…
A100	0	5.78	2.53	23.76

Analyst inputs patient-level dataset into statistical software

Type	Name	Intercept	E	X1	X2	Y
SSCP	Intercept	100.0	52.0	1157.1	405.9	2235.5
SSCP	E	52.0	52.0	813.2	138.1	1060.9
SSCP	X1	1157.1	813.2	17751.3	3458.7	23815.8
SSCP	X2	405.9	138.1	3458.7	2240.8	9572.3
SSCP	Y	2235.5	1060.9	23815.8	9572.3	56911.9
MEAN		1.0	0.5	11.6	4.1	22.4
STD		0.0	0.5	6.6	2.5	8.4
N		100	100	100	100	100

Statistical software produces intermediate statistics as part of computing process

Variable	Parameter estimate	Standard error
Intercept	25.4540	3.7959
E	-0.4323	1.7865
X1	-0.5643	0.1432
X2	-0.6564	0.4532

Statistical software produces final results

Distributed regression

ID	E	X1	X2	Y
A001	0	13.89	3.42	28.70
A002	1	18.10	1.29	27.90
A003	0	6.41	4.86	33.10
A004	1	16.30	1.45	17.20
A005	1	17.57	2.51	21.70
…	…	…	…	…
A100	0	5.78	2.53	23.76

"Regular" regression shares this

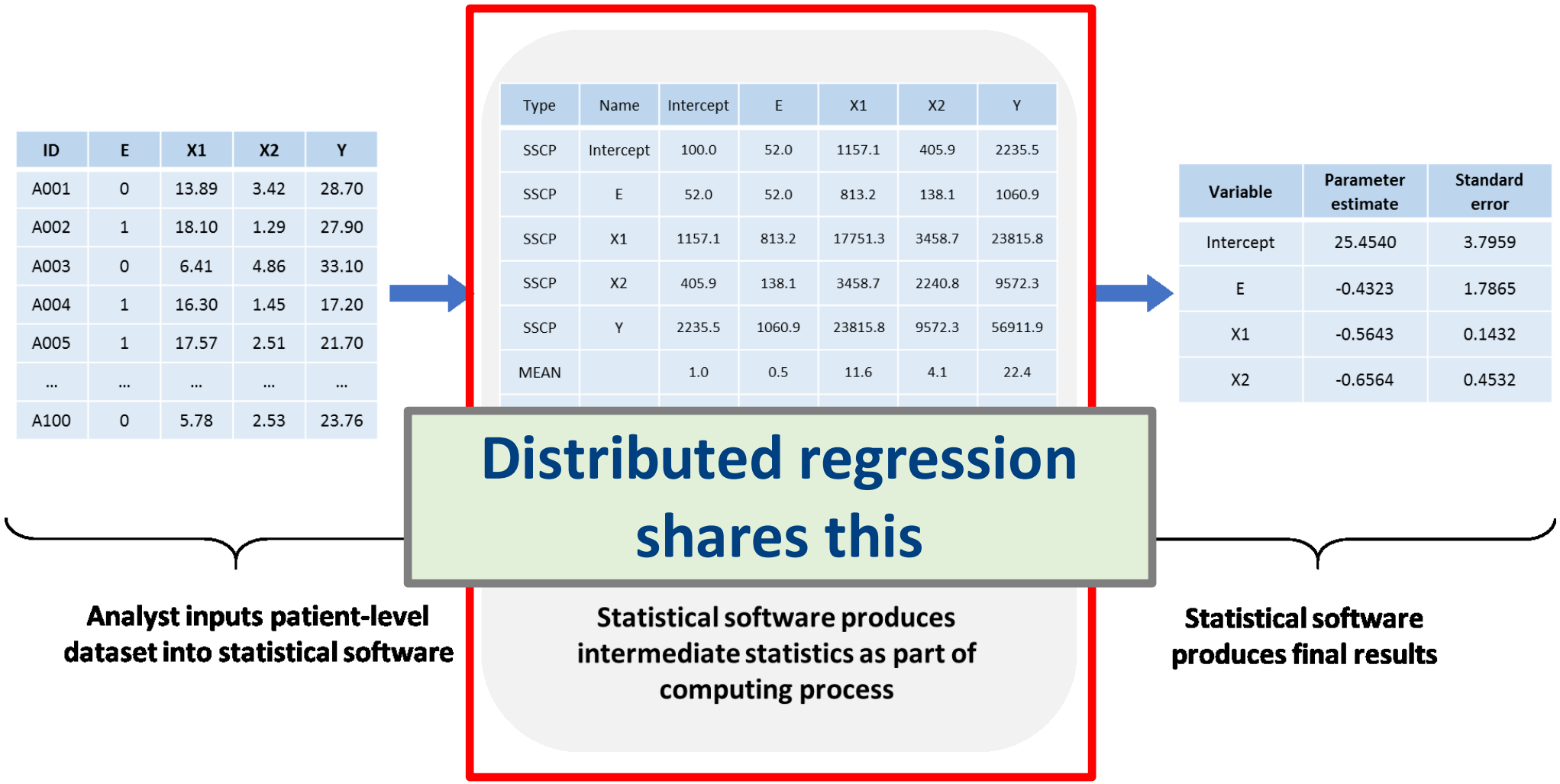
Type	Name	Intercept	E	X1	X2	Y
SSCP	Intercept	100.0	52.0	1157.1	405.9	2235.5
SSCP	E	52.0	52.0	813.2	138.1	1060.9
SSCP	X1	1157.1	813.2	17751.3	3458.7	23815.8
SSCP	X2	405.9	138.1	3458.7	2240.8	9572.3
SSCP	Y	2235.5	1060.9	23815.8	9572.3	56911.9
MEAN		1.0	0.5	11.6	4.1	22.4
STD		0.0	0.5	6.6	2.5	8.4
N		100	100	100	100	100

Statistical software produces intermediate statistics as part of computing process

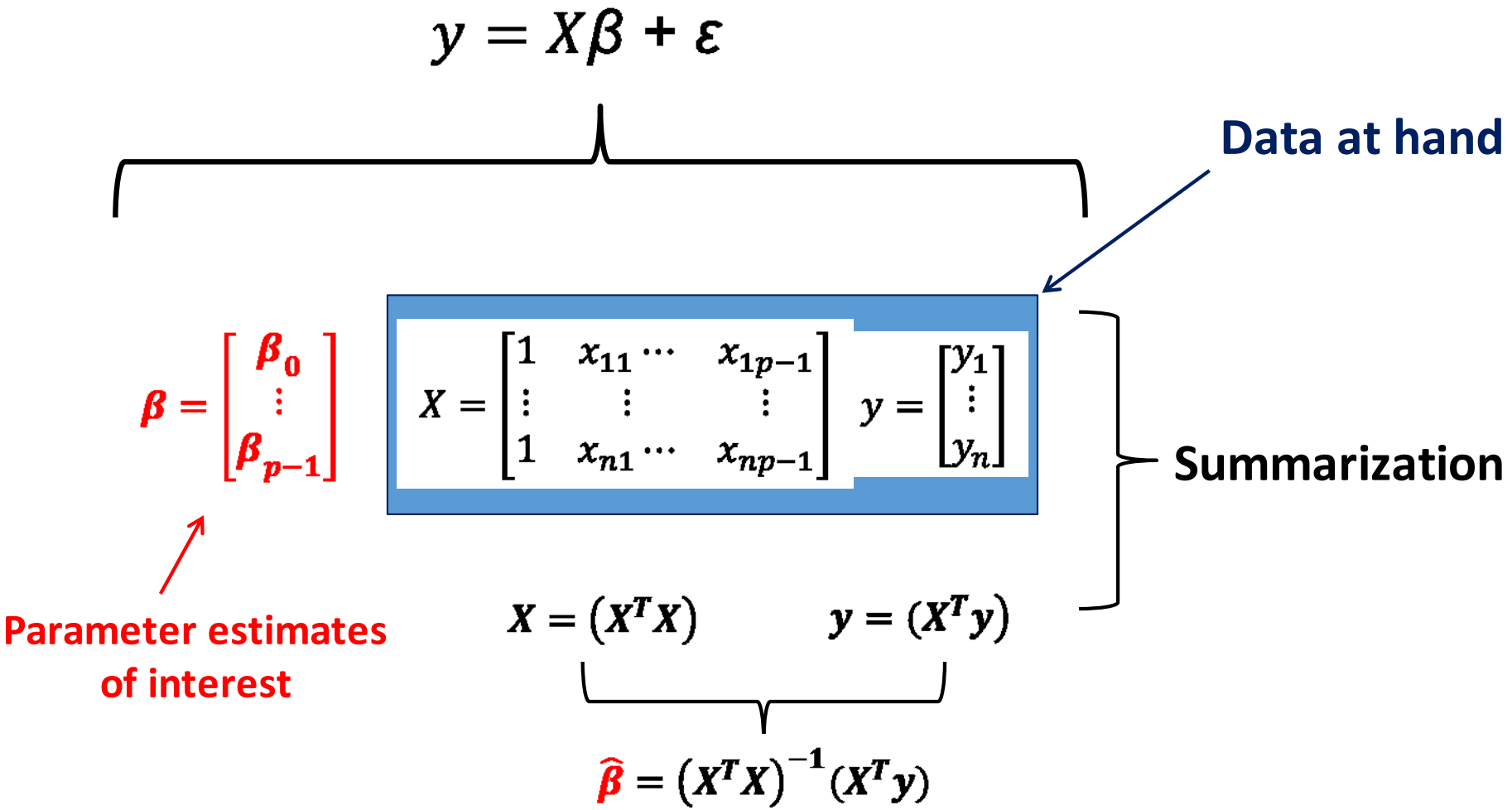
Variable	Parameter estimate	Standard error
Intercept	25.4540	3.7959
E	-0.4323	1.7865
X1	-0.5643	0.1432
X2	-0.6564	0.4532

Statistical software produces final results

Distributed regression



Distributed linear regression



Karr et al, *J Comput Graph Stat*, 2005;14:263-279

Distributed linear regression

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

$X = (X^T X)$ $\qquad$ $y = (X^T y)$

Karr et al, *J Comput Graph Stat*, 2005;14:263-279

Distributed linear regression

Agency j	n_j	$(X^j)^T X^j$	$(X^j)^T y^j$
1	172	$\begin{bmatrix} 172.00 & 49.03 & 1581.19 & 781.52 \\ 49.03 & 40.42 & 556.29 & 180.95 \\ 1581.19 & 556.29 & 23448.60 & 5631.35 \\ 781.52 & 180.95 & 5631.35 & 4186.07 \end{bmatrix}$	$\begin{bmatrix} 4057.90 \\ 909.24 \\ 32227.19 \\ 18996.12 \end{bmatrix}$
2	182	$\begin{bmatrix} 182.00 & 94.47 & 1563.50 & 746.1 \\ 94.47 & 160.90 & 1433.20 & 231.87 \\ 1563.50 & 1433.20 & 18970.98 & 5224.19 \\ 746.12 & 231.87 & 5224.19 & 3882.02 \end{bmatrix}$	$\begin{bmatrix} 4691.10 \\ 2299.13 \\ 37949.83 \\ 19193.18 \end{bmatrix}$
3	152	$\begin{bmatrix} 152.00 & 1684.95 & 2490.52 & 392.64 \\ 1684.95 & 43769.02 & 30489.61 & 3053.46 \\ 2490.52 & 30489.61 & 44106.05 & 5365.14 \\ 392.64 & 3053.46 & 5365.14 & 1458.68 \end{bmatrix}$	$\begin{bmatrix} 2652.60 \\ 22478.73 \\ 41387.06 \\ 7524.57 \end{bmatrix}$

Karr et al, *Chance*, 2004;17:27-30

Distributed linear regression

Agency j	n_j	$(X^j)^T X^j$	$(X^j)^T y^j$
1	172	$\begin{bmatrix} 172.00 & 49.03 & 1581.19 & 781.52 \\ 49.03 & 40.42 & 556.29 & 180.95 \\ 1581.19 & 556.29 & 23448.60 & 5631.35 \\ 781.52 & 180.95 & 5631.35 & 4186.07 \end{bmatrix}$	$\begin{bmatrix} 4057.90 \\ 909.24 \\ 32227.19 \\ 18996.12 \end{bmatrix}$
2	182	$\begin{bmatrix} 182.00 & 94.47 & 1563.50 & 746.1 \\ 94.47 & 160.90 & 1433.20 & 231.87 \\ 1563.50 & 1433.20 & 18970.98 & 5224.19 \\ 746.12 & 231.87 & 5224.19 & 3882.02 \end{bmatrix}$	$\begin{bmatrix} 4691.10 \\ 2299.13 \\ 37949.83 \\ 19193.18 \end{bmatrix}$
3	152	$\begin{bmatrix} 152.00 & 1684.95 & 2490.52 & 392.64 \\ 1684.95 & 43769.02 & 30489.61 & 3053.46 \\ 2490.52 & 30489.61 & 44106.05 & 5365.14 \\ 392.64 & 3053.46 & 5365.14 & 1458.68 \end{bmatrix}$	$\begin{bmatrix} 2652.60 \\ 22478.73 \\ 41387.06 \\ 7524.57 \end{bmatrix}$

$$X^T X = (X^1)^T X^1 + (X^2)^T X^2 + (X^3)^T X^3 =$$

$$\begin{bmatrix} 506.00 & 1828.44 & 5635.21 & 1920.29 \\ 1828.44 & 43970.34 & 32479.10 & 3466.28 \\ 5635.21 & 32479.10 & 86525.63 & 16220.67 \\ 1920.29 & 3466.28 & 16220.67 & 9526.77 \end{bmatrix}$$

and

$$X^T y = (X^1)^T y^1 + (X^2)^T y^2 + (X^3)^T y^3 = \begin{bmatrix} 11401.60 \\ 25687.10 \\ 111564.08 \\ 45713.87 \end{bmatrix}$$

Regression	$\hat{\beta}$ CONST	$\hat{\beta}$ CRIME	$\hat{\beta}$ IND	$\hat{\beta}$ DIST
Global	35.505	-0.273	-0.730	-1.016
Agency 1	39.362	-8.792	-0.720	-1.462
Agency 2	35.611	2.587	-0.896	-0.849
Agency 3	34.028	-0.241	-0.708	-0.893

Karr et al, *Chance*, 2004;17:27-30

Distributed logistic regression

1st Iteration:

$\hat{b}_{r=1} = [\,0 \quad 0 \quad 0 \quad 0\,]$

All data computers use this coefficient vector for iteration 1

Data Computer 1:

$$I(\hat{b}_{j=1,r=1}) = \begin{bmatrix} 500 & -11.61089 & 0 & 294.75 \\ -11.61089 & 7972.37088 & 0 & -25.55092 \\ 0 & 0 & 0 & 0 \\ 294.75 & -25.55092 & 0 & 387.75 \end{bmatrix}$$

$$S(\hat{b}_{j=1,r=1}) = \begin{bmatrix} -38 & 203.1316 & 0 & 87.5 \end{bmatrix}$$

$D_{j=1,r=1} = 2772.589$

Data Computer 2:

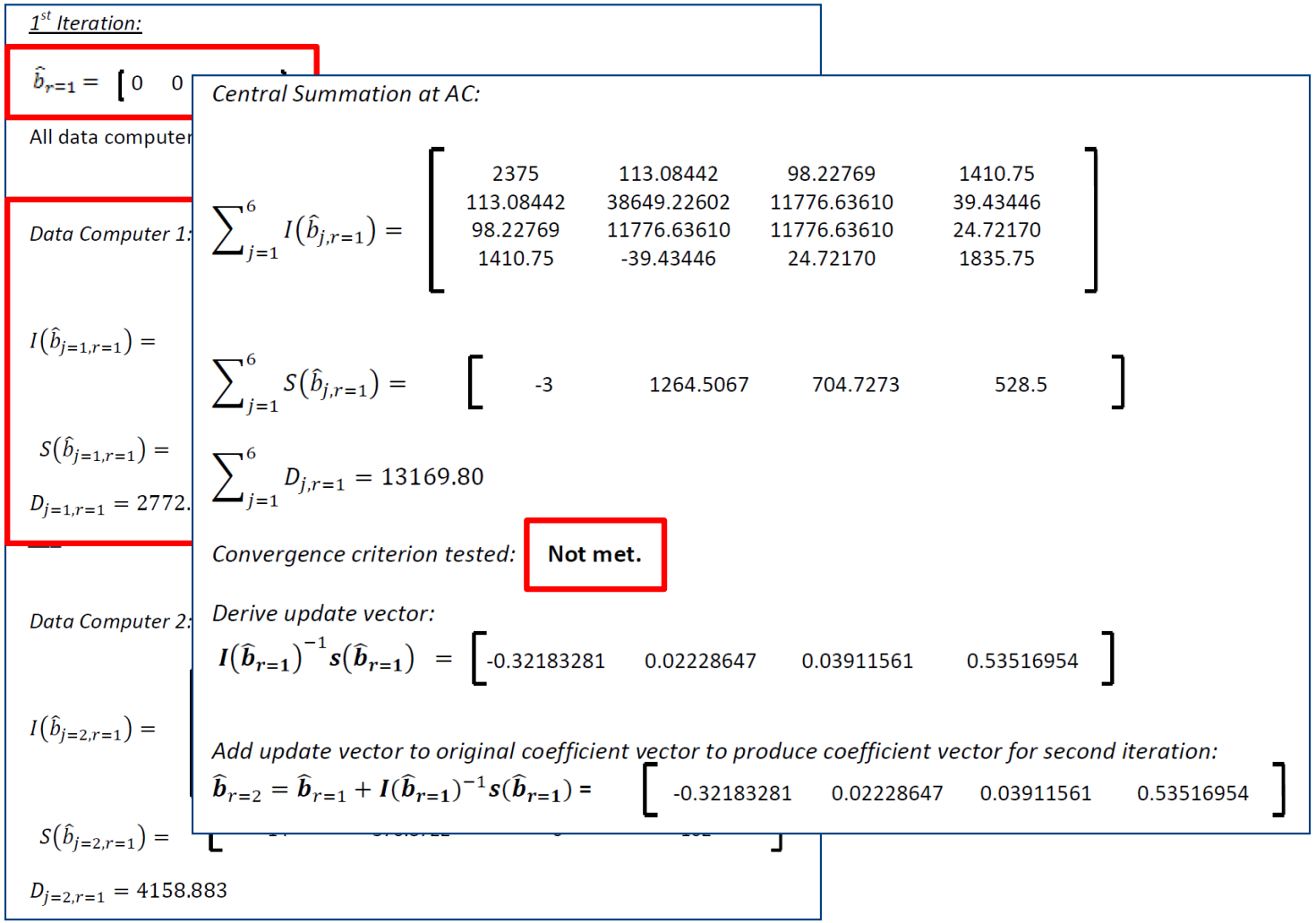
$$I(\hat{b}_{j=2,r=1}) = \begin{bmatrix} 750 & -8.417491 & 0 & 443 \\ -8.417491 & 12492.689094 & 0 & -11.777043 \\ 0 & 0 & 0 & 0 \\ 443 & -11.777043 & 0 & 578.5 \end{bmatrix}$$

$$S(\hat{b}_{j=2,r=1}) = \begin{bmatrix} -14 & 370.8722 & 0 & 162 \end{bmatrix}$$

$D_{j=2,r=1} = 4158.883$

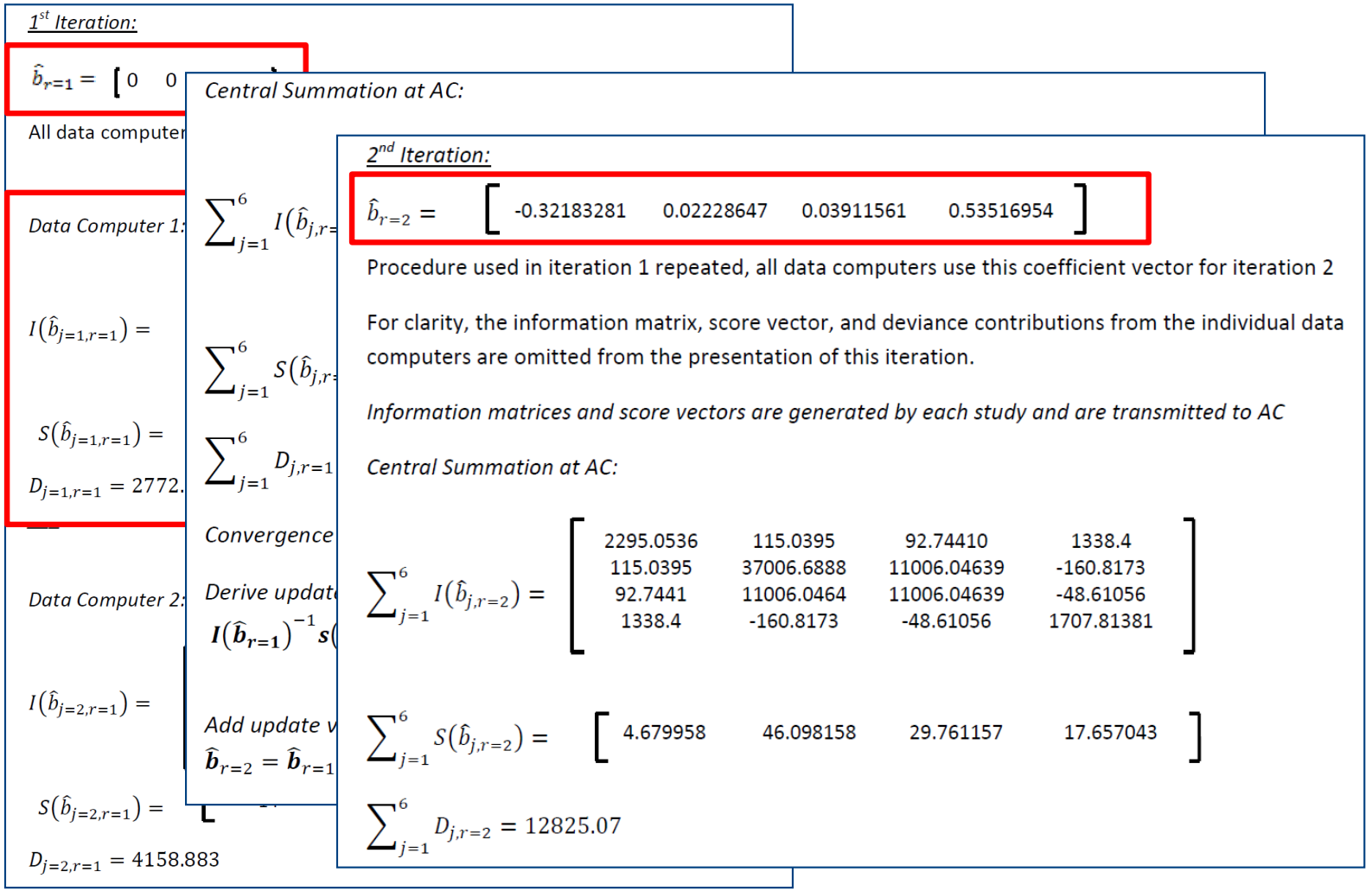
Wolfson et al, *Int J Epidemiol*, 2010;39:1372-1382

Distributed logistic regression



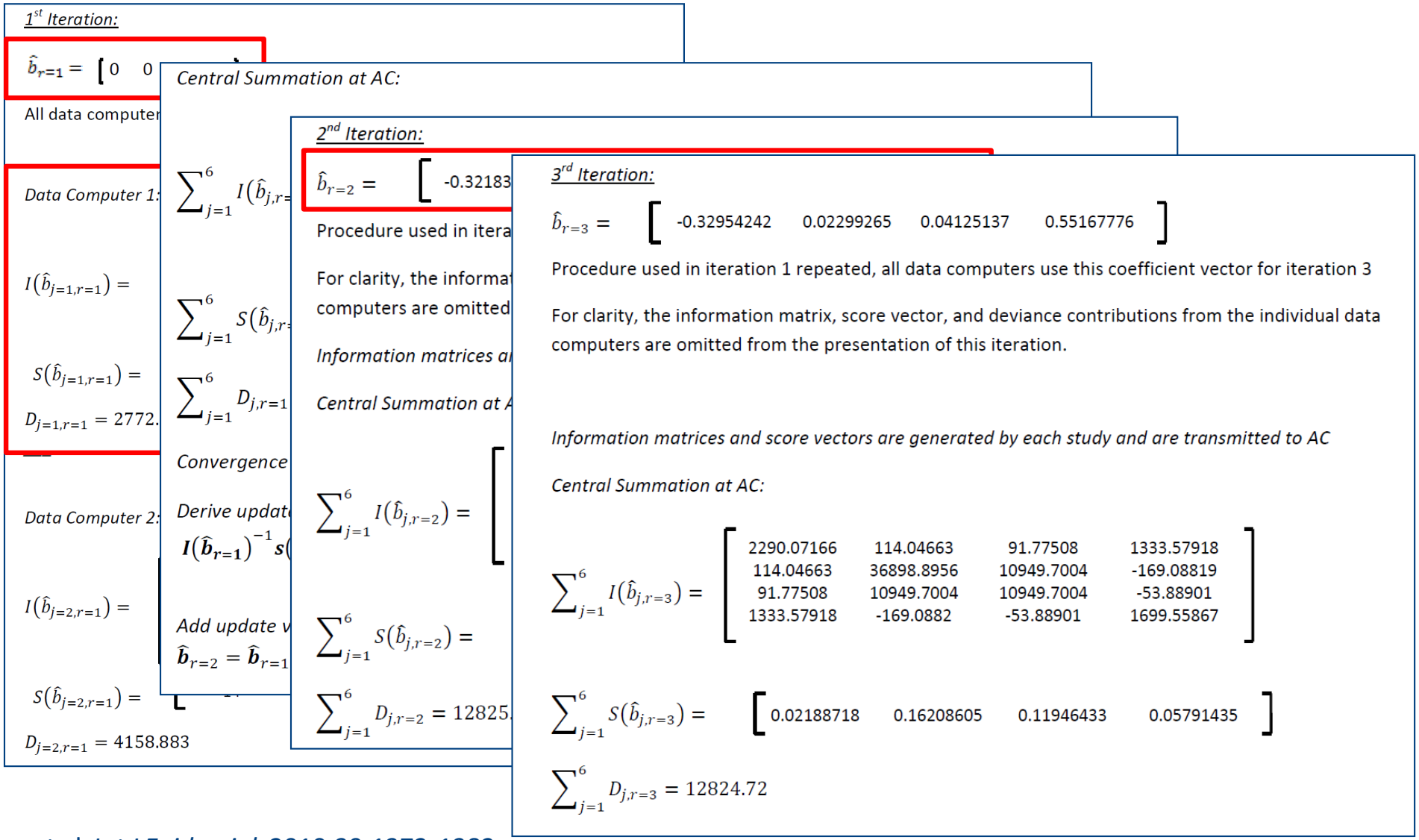
Wolfson et al, *Int J Epidemiol*, 2010;39:1372-1382

Distributed logistic regression



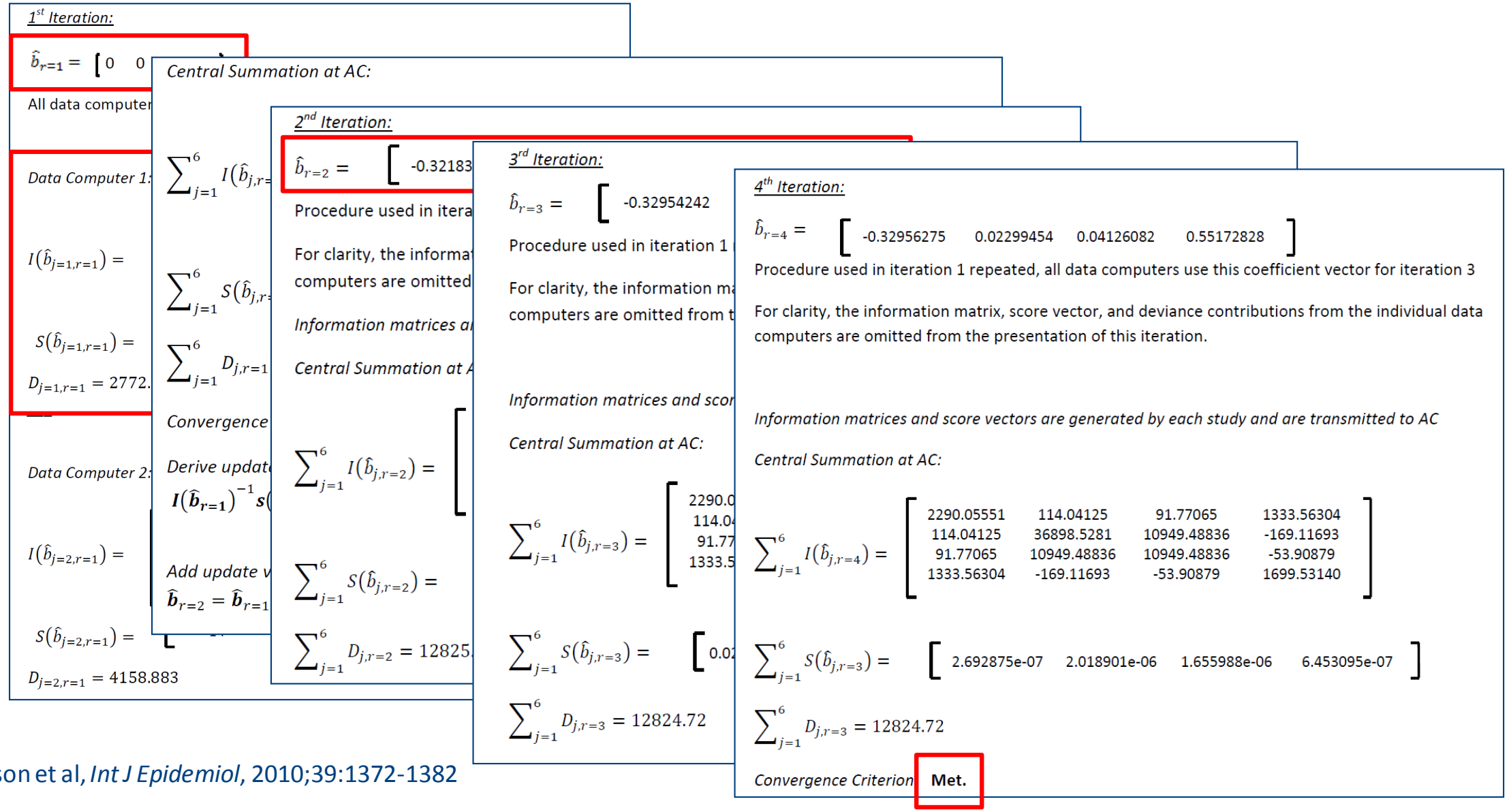
Wolfson et al, *Int J Epidemiol*, 2010;39:1372-1382

Distributed logistic regression



Wolfson et al, *Int J Epidemiol*, 2010;39:1372-1382

Distributed logistic regression



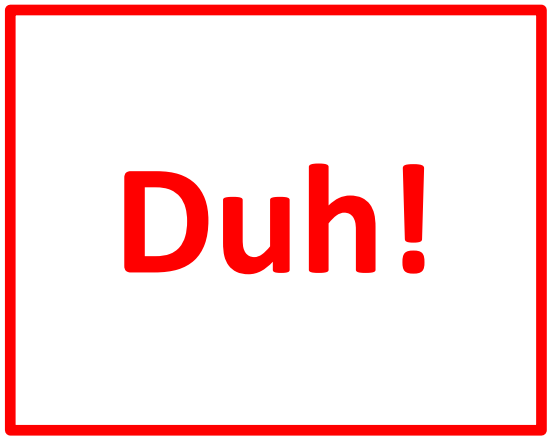
Wolfson et al, *Int J Epidemiol*, 2010;39:1372-1382

Improving practicality of distributed regression

Sentinel

- Two parallel development activities
- **Analytic code** to perform distributed regression analysis
- **Communication code** to enable automatable file transfers among physically separated computers

Checking in



Everyone knows this. What are you doing this silly study?

Checking in

Sentinel

Wow!

This is the coolest thing ever!

Checking in

Sentinel

Meh!

Who cares?

Checking in

Duh!

Wow!

Meh!

Pooled patient-level linear regression (from PROC REG)

Distributed linear regression

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	Pr > |t|	Lower 95% CL Parameter	Upper 95% CL Parameter
Intercept	1	31.79302	1.68240	<.0001	28.48757	35.09847
crim	1	-0.23283	0.04755	<.0001	-0.32626	-0.13940
indus	1	-0.51302	0.08165	<.0001	-0.67343	-0.35260
dis	1	-1.05423	0.22632	<.0001	-1.49888	-0.60957
dummy_dp_var2	1	4.62054	0.88611	<.0001	2.87958	6.36150
dummy_dp_var3	1	-1.22053	1.04369	0.2428	-3.27109	0.83003

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	P-Value	Lower 95% CL Parameter	Upper 95% CL Parameter
Intercept	1	31.79302	1.68240	<.0001	28.48757	35.09847
crim	1	-0.23283	0.04755	<.0001	-0.32626	-0.13940
indus	1	-0.51302	0.08165	<.0001	-0.67343	-0.35260
dis	1	-1.05423	0.22632	<.0001	-1.49888	-0.60957
dummy_dp_var2	1	4.62054	0.88611	<.0001	2.87958	6.36150
dummy_dp_var3	1	-1.22053	1.04369	0.2428	-3.27109	0.83003

Pooled patient-level logistic regression (from PROC LOGISTIC)

Parameter Estimates						
Parameter	DF	Parameter Estimate	Standard Error	P-Value	Lower 95% CL Parameter	Upper 95% CL Parameter
Intercept	1	1.68778	0.53174	0.0015	0.64558	2.72998
crim	1	-0.15315	0.04653	0.0010	-0.24435	-0.06195
indus	1	-0.10329	0.02570	<.0001	-0.15366	-0.05292
dis	1	-0.16344	0.07341	0.0260	-0.30732	-0.01956
dummy_dp_var2	1	1.33919	0.27156	<.0001	0.80694	1.87144
dummy_dp_var3	1	0.31595	0.37325	0.3973	-0.41560	1.04750

Distributed logistic regression

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	P-Value	Lower 95% CL Parameter	Upper 95% CL Parameter
Intercept	1	1.68778	0.53174	0.0015033	0.64558	2.72998
crim	1	-0.15315	0.04653	0.0009974	-0.24435	-0.06195
indus	1	-0.10329	0.02570	0.0000583	-0.15366	-0.05292
dis	1	-0.16344	0.07341	0.0259855	-0.30732	-0.01956
dummy_dp_var2	1	1.33919	0.27156	8.1622E-7	0.80694	1.87144
dummy_dp_var3	1	0.31595	0.37325	0.3972768	-0.41560	1.04750

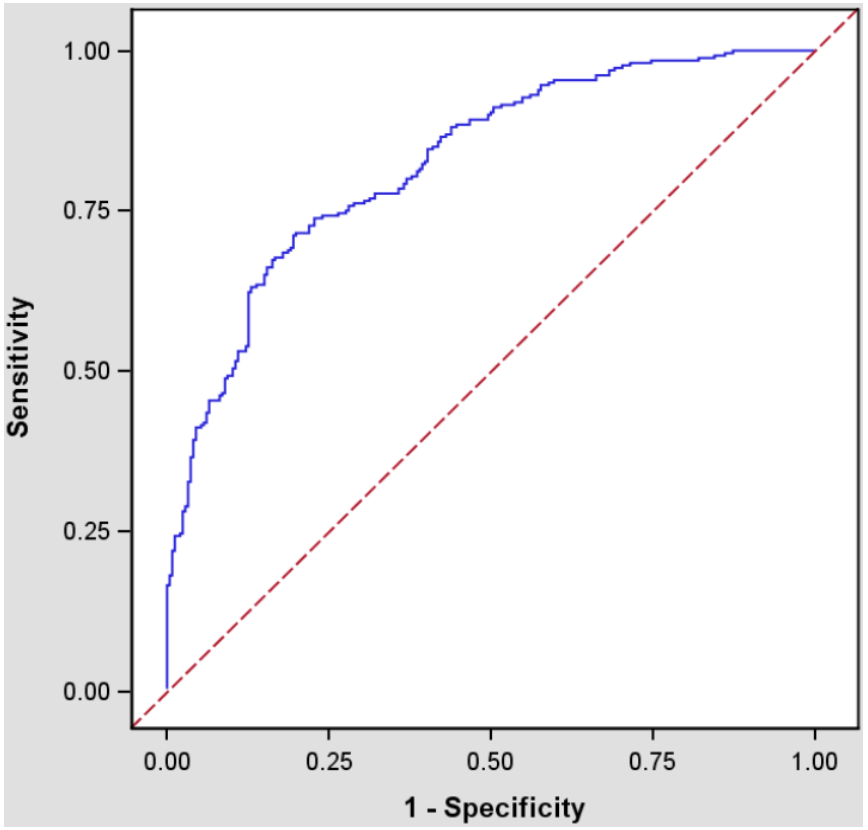
Pooled patient-level Cox PH regression (from PROC PHREG)

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Square	Hazard Ratio	Lower 95% CL Hazard Ratio	Upper 95% CL Hazard Ratio
fin	1	-0.346444	0.190236	3.316518	0.0686	0.707198	0.4870936	1.0267629
age	1	-0.066921	0.020840	10.311876	0.0013	0.935269	0.8978378	0.9742614
prio	1	0.096528	0.027241	12.556144	0.0004	1.101341	1.0440804	1.1617414

Distributed Cox PH regression

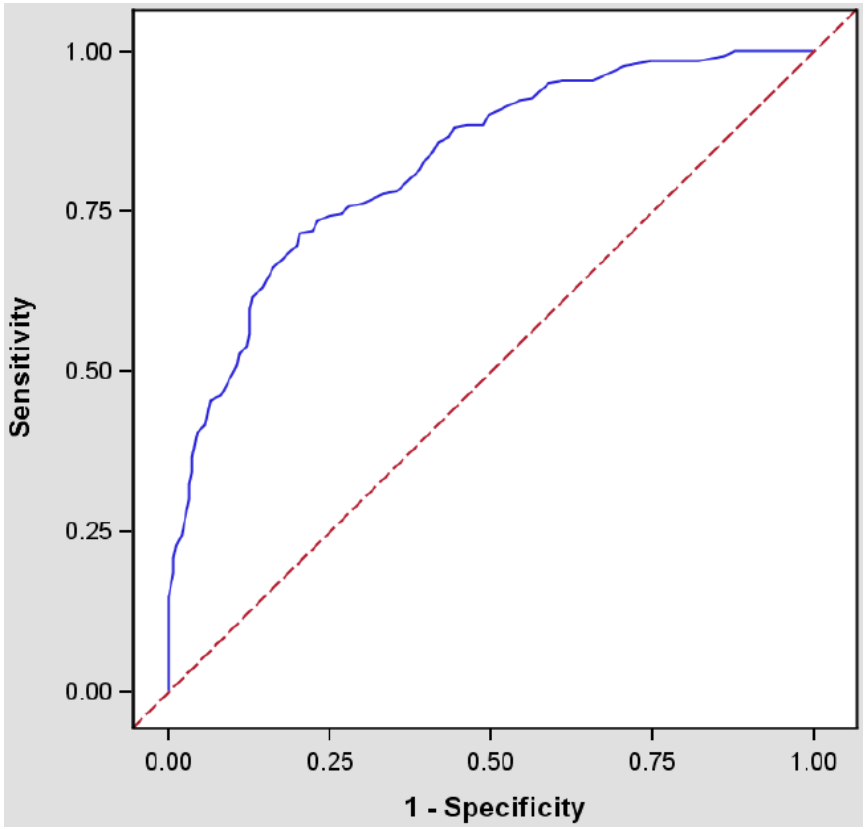
Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Square	Hazard Ratio	Lower 95% CL Hazard Ratio	Upper 95% CL Hazard Ratio
fin	1	-0.346444	0.190236	3.316518	0.0686	0.707198	0.4870936	1.0267629
age	1	-0.066921	0.020840	10.311876	0.0013	0.935269	0.8978378	0.9742614
prio	1	0.096528	0.027241	12.556144	0.0004	1.101341	1.0440804	1.1617414

Pooled patient-level logistic regression



Area under ROC curve: ***0.8255***

Distributed logistic regression



Area under ROC curve: ***0.8247****

** Slight difference was due to distributed regression requesting less granular info when performing calculation (in order to provide more privacy protection)*

Additional testing

Sentinel

Distributed Regression vs. Pooled Patient-Level Regression – LINEAR

Covariates	Distributed Regression		Pooled Patient-Level		Differences in Parameter Estimates	Differences in Standard Errors
	Parameter Estimates	Standard Errors	Parameter Estimates	Standard Errors		
Intercept	35.50548	1.57690	35.50548	1.57690	-8.38E-13	2.26E-14
Variable 1	-0.27283	0.04401	-0.27283	0.04401	4.44E-16	9.92E-16
Variable 2	-1.01582	0.23259	-1.01582	0.23259	1.09E-13	3.22E-15
Variable 3	-0.73017	0.07229	-0.73017	0.07229	3.54E-14	1.32E-15

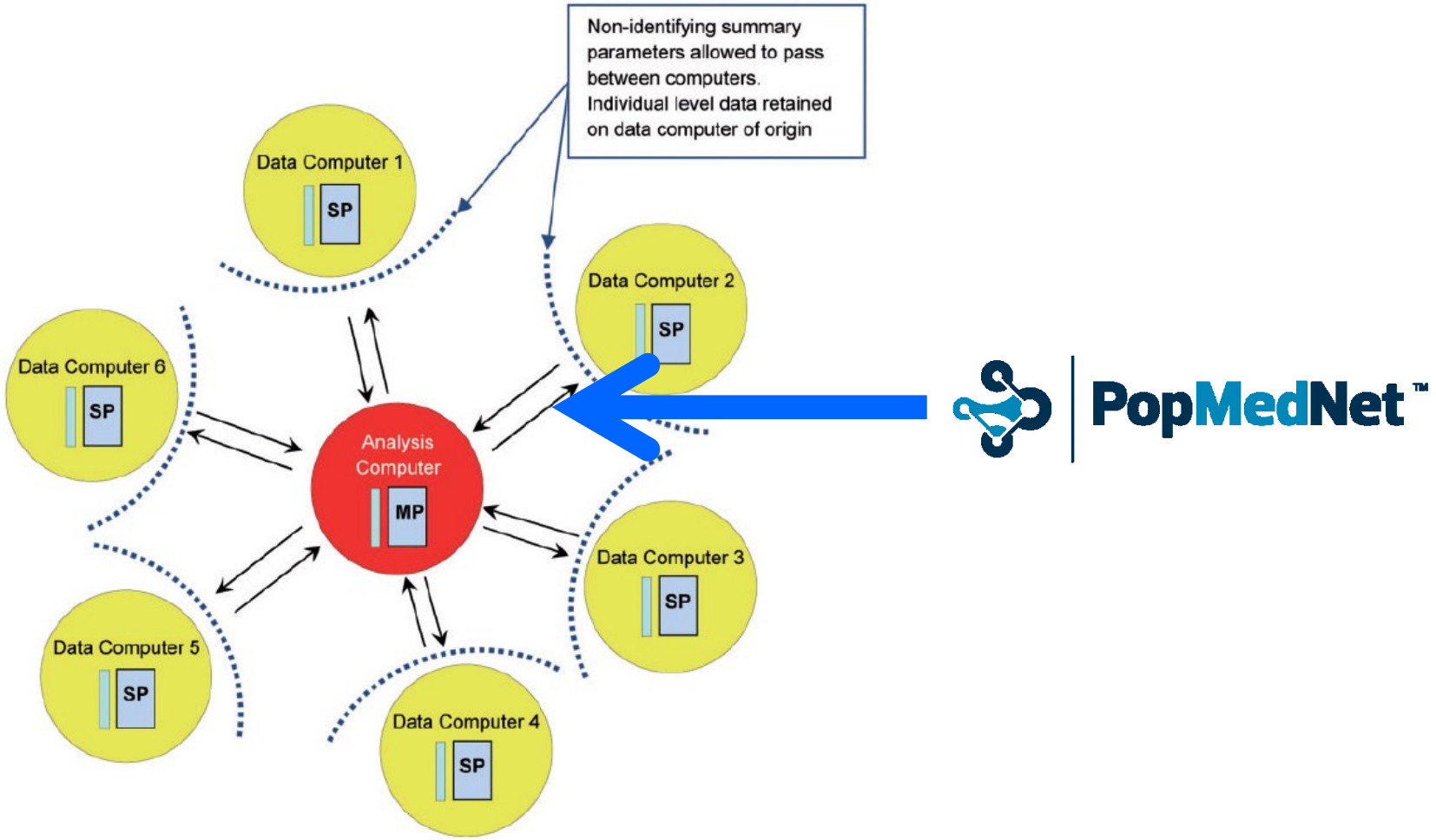
Distributed Regression vs. Pooled Patient-Level Regression – LOGISTIC

Covariates	Distributed Regression		Pooled Patient-Level		Differences in Parameter Estimates	Differences in Standard Errors
	Parameter Estimates	Standard Errors	Parameter Estimates	Standard Errors		
Intercept	2.49660	0.49057	2.49660	0.49060	1.33E-15	9.99E-16
Variable 1	-0.14465	0.03686	-0.14460	0.03690	2.04E-13	-2.97E-14
Variable 2	-0.14105	0.06976	-0.14100	0.06980	1.38E-14	-2.22E-16
Variable 3	-0.13889	0.02376	-0.13890	0.02380	-2.42E-14	-2.19E-16

Distributed Regression vs. Pooled Patient-Level Regression – COX

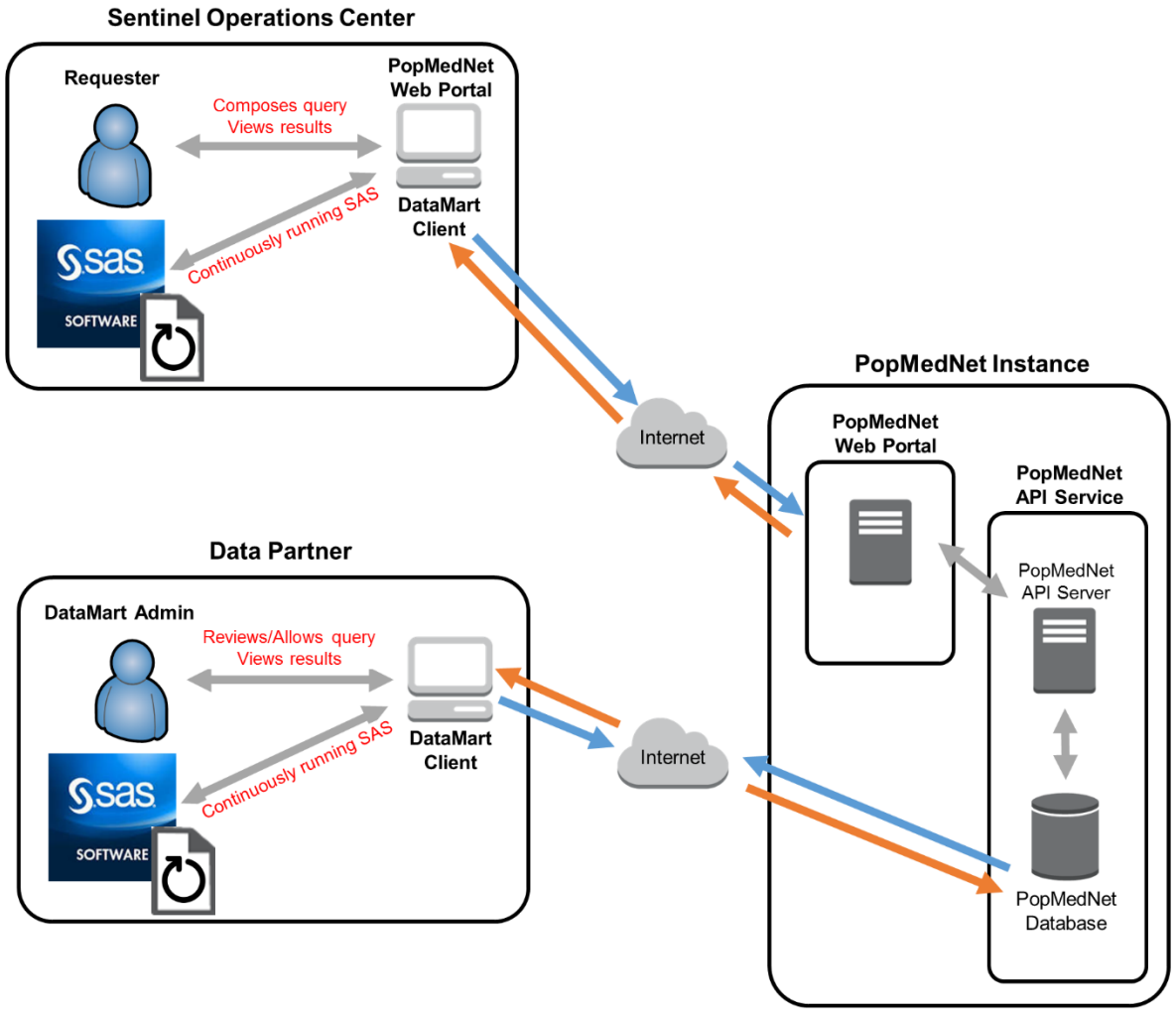
Covariates	Distributed Regression		Pooled Patient-Level		Differences in Parameter Estimates	Differences in Standard Errors
	Parameter Estimates	Standard Errors	Parameter Estimates	Standard Errors		
Variable 1	-0.06692	0.02084	-0.06692	0.02084	-1.39E-16	2.78E-17
Variable 2	-0.34644	0.19024	-0.34644	0.19024	2.22E-16	-2.78E-17
Variable 3	0.09653	0.02724	0.09653	0.02724	-1.80E-16	1.73E-17

Development of communication code



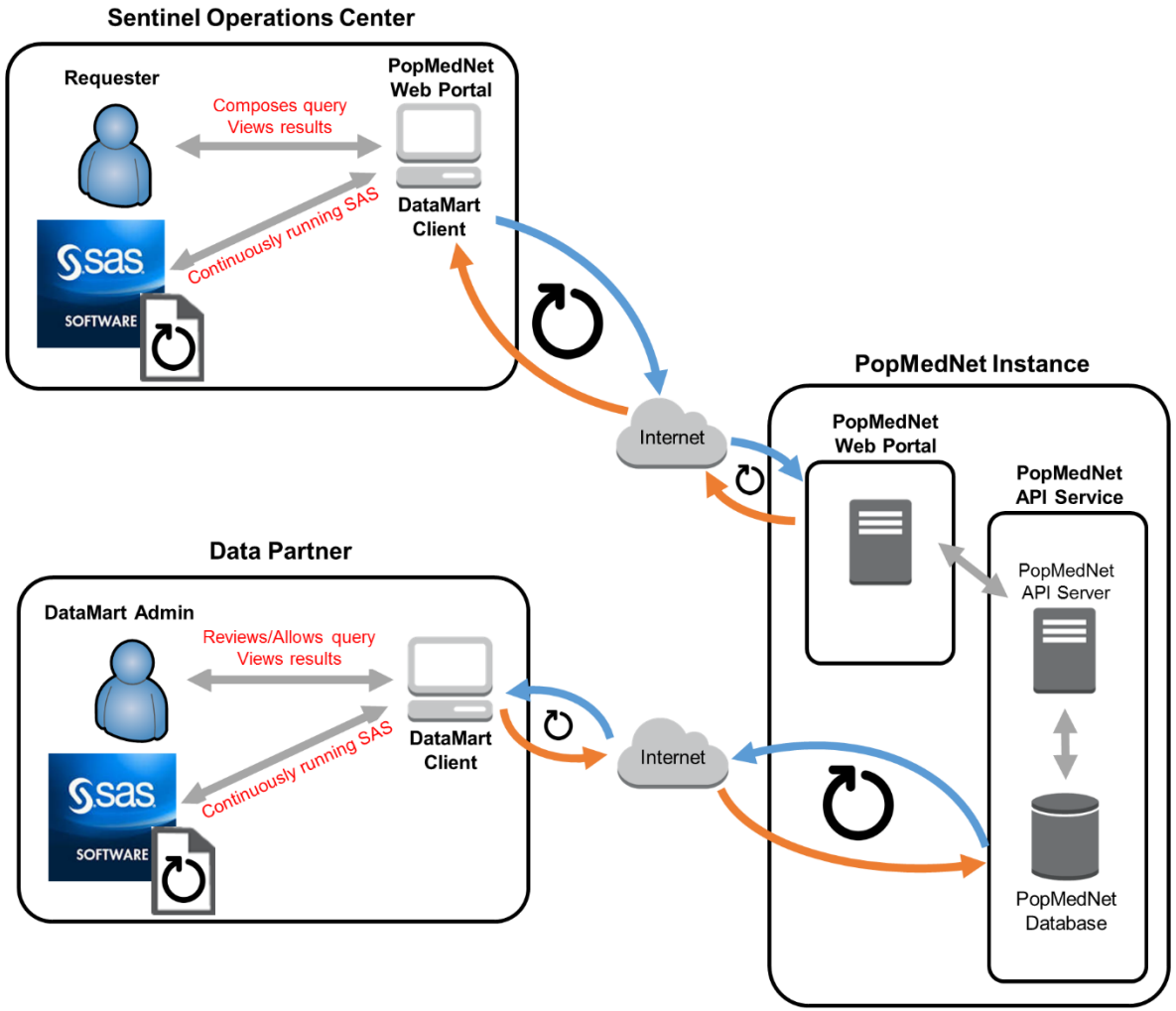
Wolfson et al, *Int J Epidemiol*, 2010;39:1372-1382

PopMedNet query workflow (Old)



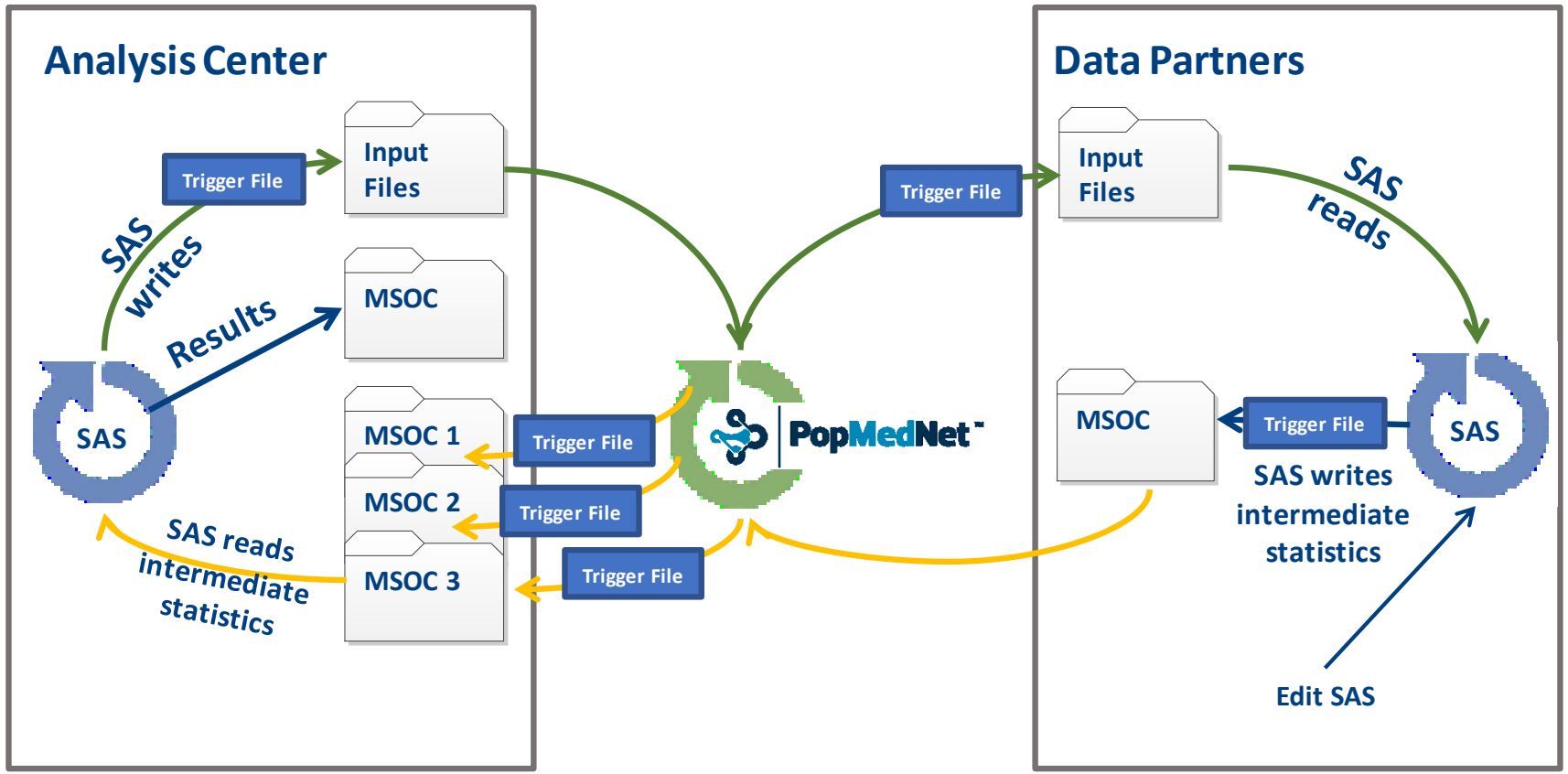
One-time info exchange

PopMedNet query workflow (Enhanced)



Iterative info exchange

Integrating analytic and communication code



Discussion

- We are getting very close to making distributed regression a practical analytic option in real-world distributed data networks
- Effect estimation is relatively straightforward
- Some model diagnostics are trickier
- We also have functional prototypes for distributed linear and logistic regression for vertically partitioned data environments

Checking in

Duh!

Wow!

Meh!

Darren_Toh@harvardpilgrim.org

@darrentoh_epi

https://www.distributedanalysis.org