Re*designing* PopMedNet™ for Distributed Regression Analysis with Vertically Partitioned Data

Qoua L Her

DEPARTMENT OF POPULATION MEDICINE

HARVARD MEDICAL SCHOOL

Harvard Pilgrim Health Care Institute

Privacy-Protecting Methods

Team

Harvard Pilgrim Health Care Institute

- Darren Toh, ScD (Principal Investigator, Epidemiologist)
- Mia Gallagher, MPH (Project Manager)
- Yury Vilk, PhD (Programmer)
- Qoua Her, PharmD, MSc (Informatician)

Pennsylvania State University

- Aleksandra Slavkovic, PhD (Co-Investigator, Statistician)
- Yuji Samizo, PhD (Statistician)
- Thomas Kent, PhD (Statistician)

Disclosure

- This work was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number U01EB023683.
  - The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

- The authors have no relevant conflicts of interest to disclose

Background

Insurance Organization 1

ID	X1	X2	X3	Y
001	1	0	0	0
002	0	0	0	1
…	…	…	…	…
006	1	1	1	0

ID	X1	X2	X3	Y
001	1	0	0	0
002	0	0	0	1
003	0	0	0	1
004	0	0	1	0
005	0	1	1	0
006	1	1	1	0
007	1	0	0	1
008	1	0	0	0
009	0	1	1	1
010	0	0	0	1
011	0	0	1	1
012	0	0	0	0

Horizontally partitioned data

Insurance Organization 2

ID	X1	X2	X3	Y
007	1	0	0	1
008	1	0	0	0
…	…	…	…	…
012	0	0	0	0

- ❑ **The number of patients at each data partner site may be too small to conduct any meaningful analysis**

Karr et al, 2005; Fienberg et al, 2006

Background

Insurance Organization

ID	X1	X2
001	1	0
002	0	0
...	...	...
012	0	0

ID	X1	X2	X3	Y
001	1	0	0	0
002	0	0	0	1
003	0	0	0	1
004	0	0	1	0
005	0	1	1	0
006	1	1	1	0
007	1	0	0	1
008	1	0	0	0
009	0	1	1	1
010	0	0	0	1
011	0	0	1	1
012	0	0	0	0

Vertically partitioned data

Hospital

ID	X3	Y
001	0	0
002	0	1
...	...	...
012	0	0

- ❑ **Important variables (outcome or confounders) may exist in another data sources**
  - ❑ **Lab data**

Karr et al, 2005; Fienberg et al, 2006

Background

- Data owners may be **unwilling** or **unable** to share their individual-level data
  - Patient privacy
  - Disclosing propriety or sensitive institutional information
  - Even if sharing individual-level data is possible, methods that are equally valid and precise that shares less granular data (summary-level information) should be preferred

- Privacy protecting analytical methods may alleviate these concerns
  - Meta-analysis
  - Confounder summary score-based methods
  - Encryption-based methods
  - **Distributed regression analysis**
    - Suite of methods that performs <u>outcomes regression analysis</u> without the need to share any individual-level data
    - Requires sharing only highly <u>summarized information</u> (intermediate statistics)

Toh et al, 2013

Background

ID	X1	X2	X3	Y
001	1	0	0	22
002	0	0	0	25
003	0	1	0	17
004	1	1	0	33
…	…	…	…	…
1005	1	1	1	57

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$$

$$X_1 = (X_1^T X_1)$$

$$y_1 = (X_1^T y_1)$$

	Intercept	X1	X2	X3
Intercept	172	765.579	2024.91	639.423
X1	765.579	18389.53	13708.48	1477.328
X2	2024.91	13708.48	31694.4	5591.684
X3	639.423	1477.328	5591.684	3253.901

$$\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (X_1^T X_1)^{-1} * (X_1^T y_1)$$

	y
Intercept	3689.2
X1	11114.16
X2	39097.86
X3	14486.67

Data partner **Intermediate Statistics** **Analysis Center**

Karr et al, 2005; Fienberg et al, 2006; Lu et al, 2015

Background

$$\boldsymbol{X_1} = (\boldsymbol{X_1^T}\, \boldsymbol{X_1})$$

$$\boldsymbol{y_1} = (\boldsymbol{X_1^T}\, \boldsymbol{y_1})$$

$$\boldsymbol{X_2} = (\boldsymbol{X_2^T}\, \boldsymbol{X_2})$$

$$\boldsymbol{y_2} = (\boldsymbol{X_2^T}\, \boldsymbol{y_2})$$

$$\boldsymbol{X_3} = (\boldsymbol{X_3^T}\, \boldsymbol{X_3})$$

$$\boldsymbol{y_3} = (\boldsymbol{X_3^T}\, \boldsymbol{y_3})$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \vdots \\ \vdots \\ \hat{\boldsymbol{\beta}}_p \end{bmatrix} = \sum (\boldsymbol{X_k^T}\, \boldsymbol{X_k})^{-1} * \sum (\boldsymbol{X_k^T}\, \boldsymbol{y_k})$$

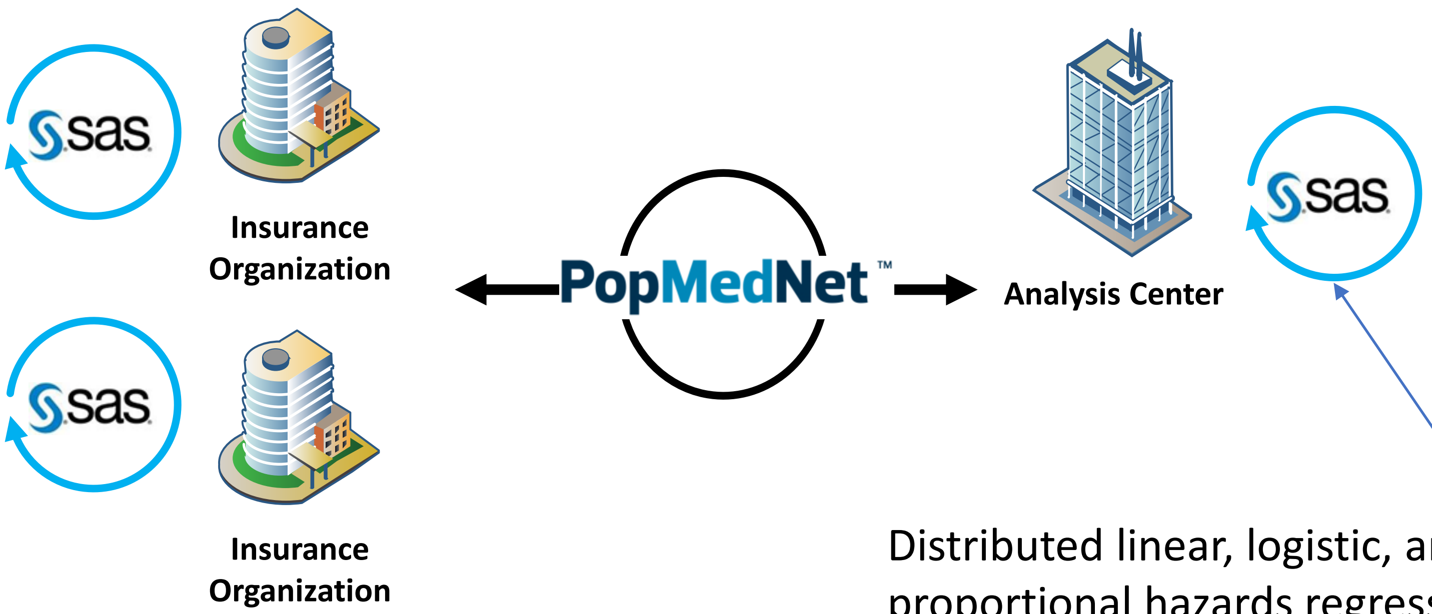
Distributed regression analysis with horizontally partitioned data

Data partners

Analysis Center

Karr et al, 2005; Fienberg et al, 2006; Lu et al, 2015

Background



Distributed linear, logistic, and Cox and stratified Cox proportional hazards regression analysis

PopMedNet powers networks such as:

Sentinel | pcornet | NIH Collaboratory Health Care Systems Research Collaboratory | CRN | ESP | CHORDS

Her et al, 2018; https://www.popmednet.org/

Background

In theory, we should be able to perform vertical distributed regression analysis with PopMedNet^{TM}

2-party workflow

Her et al, 2018; https://www.popmednet.org/

Background

- Distributed regression analysis with vertically partitioned data is **more complex** than with horizontally partitioned data

We want to compute

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} \quad (\boldsymbol{X}^T\boldsymbol{y})$$

…when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are distributed vertically

Data Partner 1

ID	Y
001	0
002	1
003	1
004	1
…	…
1005	0

Data Partner 2

ID	X1	X2	X3
001	1	0	0
002	0	0	0
003	0	1	0
004	1	1	0
…	…	…	…
1005	1	1	1

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T\,\boldsymbol{X}_1 & \boldsymbol{X}_1^T\,\boldsymbol{X}_2 \\ (\boldsymbol{X}_1^T\,\boldsymbol{X}_2)^T & \boldsymbol{X}_2^T\,\boldsymbol{X}_2 \end{bmatrix} \qquad \boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} \boldsymbol{X}_1^T\,\boldsymbol{Y} \\ \boldsymbol{X}_2^T\,\boldsymbol{Y} \end{bmatrix}$$

Karr et al, 2007; Fienberg et al, 2009

Background

We want to compute

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} \quad (\boldsymbol{X}^T\boldsymbol{y})$$

…when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are distributed vertically

Data Partner 1

ID	Y
001	0
002	1
003	1
004	1
…	…
1005	0

Data Partner 2

ID	X1	X2	X3
001	1	0	0
002	0	0	0
003	0	1	0
004	1	1	0
…	…	…	…
1005	1	1	1

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \boxed{\boldsymbol{X}_1^T\boldsymbol{X}_1} & \boldsymbol{X}_1^T\boldsymbol{X}_2 \\ (\boldsymbol{X}_1^T\boldsymbol{X}_2)^T & \boxed{\boldsymbol{X}_2^T\boldsymbol{X}_2} \end{bmatrix}$$

$$\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} \boxed{\boldsymbol{X}_1^T\boldsymbol{Y}} \\ \boldsymbol{X}_2^T\boldsymbol{Y} \end{bmatrix}$$

We can compute these components of the intermediate statistics at each data partner (e.g., horizontally partitioned data).

Karr et al, 2007; Fienberg et al, 2009

Background

We want to compute

$$(X^T X)^{-1} \quad (X^T y)$$

…when X and Y are distributed vertically

Data Partner 1

ID	Y
001	0
002	1
003	1
004	1
…	…
1005	0

Data Partner 2

ID	X1	X2	X3
001	1	0	0
002	0	0	0
003	0	1	0
004	1	1	0
…	…	…	…
1005	1	1	1

$$X^T X = \begin{bmatrix} X_1^T X_1 & \boxed{X_1^T X_2} \\ \boxed{(X_1^T X_2)^T} & X_2^T X_2 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} X_1^T Y \\ \boxed{X_2^T Y} \end{bmatrix}$$

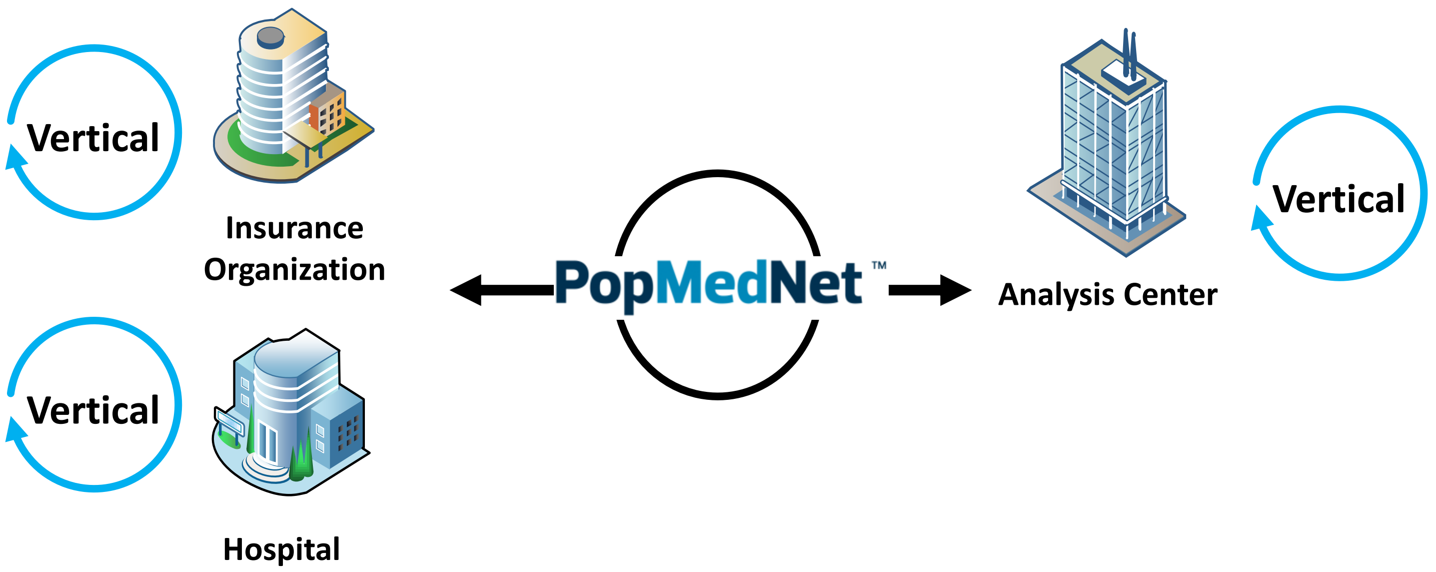
*These components cannot be computed at the data partners and require algorithms that are **computational demanding** and **multiple exchanges** of data files that are of **large sizes**.*

Karr et al, 2007; Fienberg et al, 2009

Background

Inclusion of an analysis center will decrease the computational demand of the vertical distributed regression analysis algorithm and can enhance privacy protection (sharing more granular and less information)

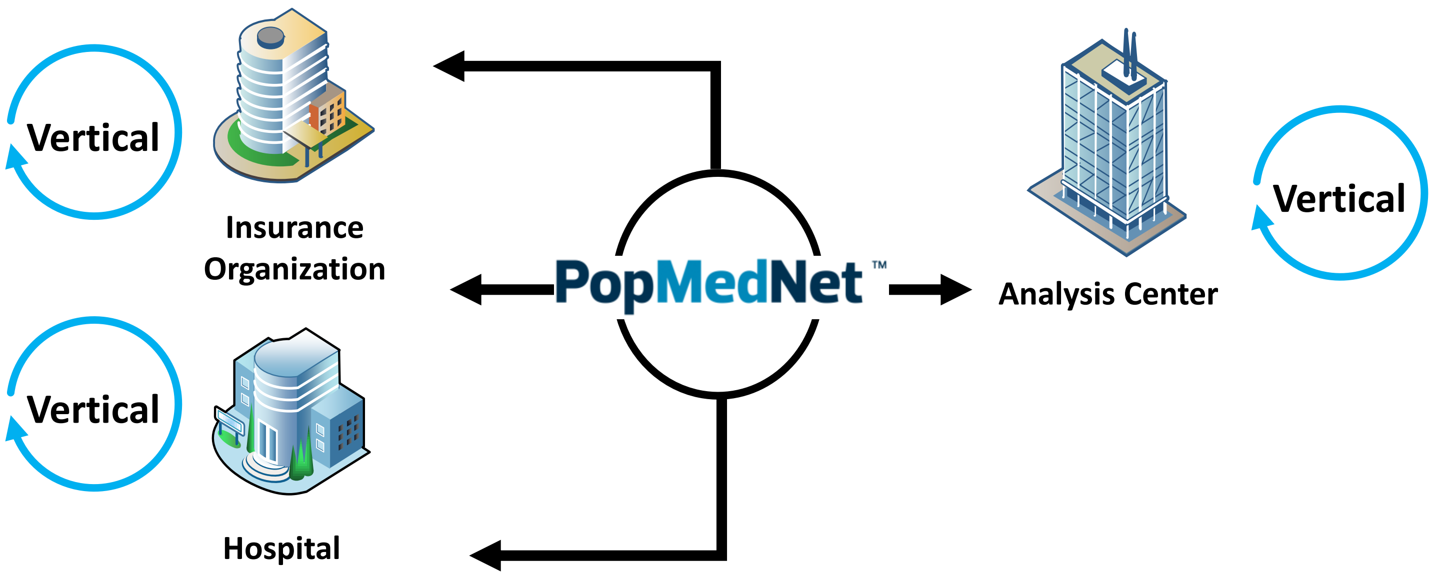
Background



3-party

Inclusion of an analysis center will decrease the computational demand of the vertical distributed regression analysis algorithm and can enhance privacy protection (sharing more granular and less information)

Background



K-party

Inclusion of an analysis center will decrease the computational demand of the vertical distributed regression analysis algorithm and can enhance privacy protection (sharing more granular and less information)

Further decrease computational demand and increase privacy protection

Objective

- Explore the feasibility of using PopMedNet™ to organize and facilitate distributed regression analysis with vertically partitioned data

- Develop a practical R-based application to perform distributed regression analysis with vertically partitioned data

    - Integrate the R-based application with PopMedNet™, and evaluate the application's precision compared to the regression analysis with the pooled individual-level data and operational performance

Methods

- We tested the integration with simulated data (n = 5,760 to 1,023,504 and 48 covariates) in a test environment comprised of **two data partners** and **an analysis center**

Regression Model Type	Outcome Variable (within one-year post surgery)	Variables (exposure and confounders)
Linear	Change in body mass index	Bariatric surgery exposure, age at surgery, sex, race/ethnicity, combined Charlson-Elixhauser comorbidity score, number of ambulatory visits, number of other ambulatory visits, number of inpatient stays, number of non-acute institutional stays, number of emergency department visits, BMI prior to bariatric surgery, and number of days between last weight and height measurement and bariatric surgery
Logistic	Weight loss ≥ 20%	
Cox	Time to weight loss ≥ 20%	

Methods

Data Partner 1

ID	X1	X2	X3	X4	…	X45
001	1	0	22	3	…	0
002	0	0	35	2	…	1
003	1	1	45	8	…	0
004	1	0	28	10	…	1
…	…	…	…	…	…	…
5760	1	1	21	7	…	0

Covariates

Linear regression
Logistic regression

Data Partner 2

ID	Y
001	1
002	0
003	1
004	1
…	…
5760	0

Outcomes

Methods

Data Partner 1

Data Partner 2

ID	X1	X2	X3	X4	…	X44
001	1	0	22	3	…	50
002	0	0	35	2	…	65
003	1	1	45	8	…	21
004	1	0	28	10	…	33
…	…	…	…	…	…	…
5760	1	1	21	7	…	17

Covariates

Cox proportional hazards regression

ID	Y	T	X45
001	1	220	0
002	0	198	1
003	1	200	0
004	1	222	1
…	…	…	…
5760	0	201	0

Outcomes

Contains no covariates

Methods

Test Environment Hardware Description

Site	Operating System	Processor	Random Access Memory
Analysis Center	Windows 7 Professional	Intel(R) Xeon(R) E5-2609 0 @ 2.40GHz, 2400 Mhz, 4 Cores, 4 Logical Processors	16 GB
Data Partner 1	Windows 7 Professional	Intel(R) Xeon(R) E5-2637 v4 @ 3.50GHz, 3501 Mhz, 4 Cores, 8 Logical Processors	32 GB
Data Partner 2	Windows 7 Professional	Intel(R) Xeon(R) E5-2637 v4 @ 3.50GHz, 3491 Mhz, 4 Cores, 8 Logical Processors	32 GB

Results

- We developed a R-based application to compute the **off-diagonal components** of the intermediate statistics using a *secure matrix multiplication algorithm*

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T\,\boldsymbol{X}_1 & \boldsymbol{X}_1^T\,\boldsymbol{X}_2 \\ \left(\boldsymbol{X}_1^T\,\boldsymbol{X}_2\right)^T & \boldsymbol{X}_2^T\,\boldsymbol{X}_2 \end{bmatrix}$$

$$\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} \boldsymbol{X}_1^T\,\boldsymbol{Y} \\ \boldsymbol{X}_2^T\,\boldsymbol{Y} \end{bmatrix}$$

The secure matrix multiplication algorithm requires the data partners to share <u>components</u> of the <u>off-diagonal components</u>.

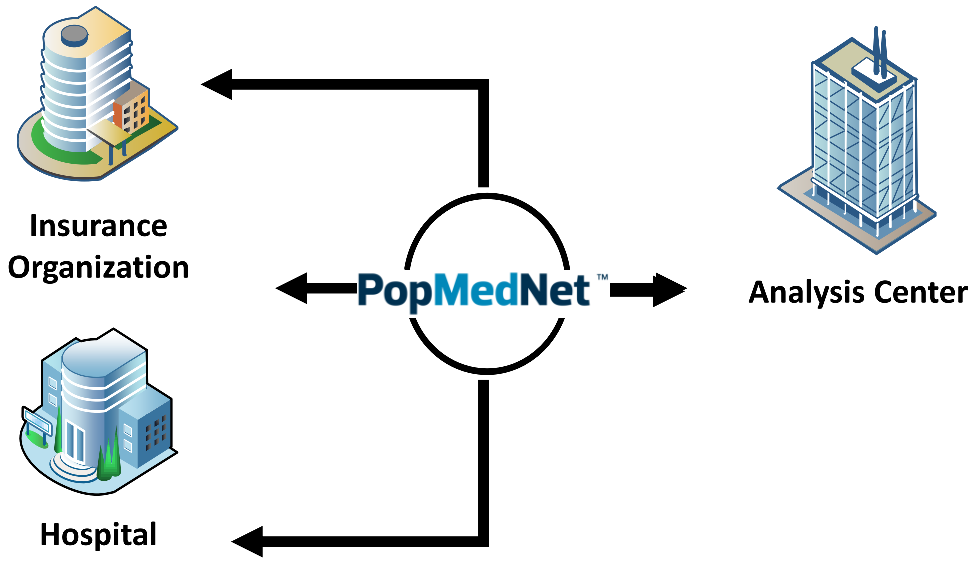
Karr et al, 2007; Fienberg et al, 2009

Results

- We enhanced PopMedNet™ to support workflow that organizes and facilitates distributed regression analysis with vertically partitioned data

  - *Concurrency of file upload and download*

  - *Enhanced trust model that supports the transfer of files between data partners and different workflows*



Distributed Linear Regression vs. Pooled Individual-level Linear Regression

Covariates	Distributed Regression Analysis		Pooled Individual-Level Analysis		Difference in Parameters	Difference in Std Errors
	Parameter	Std Error	Parameter	Std Error		
Intercept	-31.996104	0.010663	-31.9961	0.010663	**-5.06E-09**	**-2.46E-11**
Exposure	-4.998626	0.001978	-4.998626	0.001978	**3.17E-10**	**-4.60E-12**
Age	0.200061	0.000099	0.200061	0.000099	**1.31E-11**	**-2.27E-13**
Pre-Index Body Mass Index (BMI)	0.000005	0.000108	0.000005	0.000108	**1.16E-12**	**-2.51E-13**
Combined Comorbidity Score	0.299788	0.000537	0.299788	0.000537	**2.11E-10**	**-1.25E-12**
No. Ambulatory Visits	0.999908	0.000149	0.999908	0.000149	**-3.01E-11**	**-3.47E-13**

**N = 1,023,504, number of variables = 48*

K-party workflow

Results of 42 variables are not shown, all regression parameter estimates and standard errors are precise to the results obtained from the pooled individual-level data analysis (<10^{-10})

Distributed Logistic Regression vs. Pooled Individual-level Logistic Regression

Covariates	Distributed Regression Analysis		Pooled Individual-Level Analysis		Difference in Parameters	Difference in Std Errors
	Parameter	Std Error	Parameter	Std Error		
Intercept	-6.11833	0.034394	-6.11833	0.034394	-1.81E-11	5.59E-09
Exposure	0.006777	0.00567	0.006777	0.00567	3.83E-13	5.41E-10
Age	-0.000269	0.000283	-0.000269	0.000283	2.89E-13	2.71E-11
Pre-Index Body Mass Index (BMI)	0.165749	0.000508	0.165749	0.000508	-2.72E-13	1.51E-10
Combined Comorbidity Score	0.004295	0.00154	0.004295	0.00154	7.34E-14	1.48E-10
No. Ambulatory Visits	0.000589	0.000427	0.000589	0.000427	-2.56E-14	4.08E-11

**N = 1,023,504, number of variables = 48*

K-party workflow

Results of 42 variables are not shown, all regression parameter estimates and standard errors are precise to the results obtained from the pooled individual-level data analysis (<10^{-9})

Distributed Cox Proportional Hazards Regression vs. Pooled Individual-level Cox Proportional Hazards Regression

Covariates	Distributed Regression Analysis		Pooled Individual-Level Analysis		Difference in Parameters	Difference in Std Errors
	Parameter	Std Error	Parameter	Std Error		
Exposure	0.002599	0.002176	0.002599	0.002176	**3.92E-13**	**2.38E-13**
Age	-0.000156	0.000109	-0.000156	0.000109	**-1.24E-14**	**1.51E-13**
Pre-Index Body Mass Index (BMI)	0.055755	0.000113	0.055755	0.000113	**-9.15E-12**	**1.73E-13**
Combined Comorbidity Score	0.001032	0.00059	0.001032	0.00059	**-9.98E-14**	**1.95E-12**
No. Ambulatory Visits	0.000073	0.000164	0.000073	0.000164	**-1.39E-14**	**2.98E-13**

**N = 1,023,504, number of variables = 48*

K-party workflow

Only one data partner contributes variable data to the regression analysis

Results of 43 variables are not shown, all regression parameter estimates and standard errors are precise to the results obtained from the pooled individual-level data analysis ($<10^{-12}$)

Distributed Cox Proportional Hazards Regression vs. Pooled Individual-level Cox Proportional Hazards Regression

Covariates	Distributed Regression Analysis		Pooled Individual-Level Analysis		Difference in Parameters	Difference in Std Errors
	Parameter	Std Error	Parameter	Std Error		
Exposure	0.002599	0.002176	0.002599	0.002176	**4.10E-13**	**9.85E-14**
Age	-0.000156	0.000109	-0.000156	0.000109	**-1.49E-14**	**2.42E-13**
Pre-Index Body Mass Index (BMI)	0.055755	0.000113	0.055755	0.000113	**-9.15E-12**	**3.71E-15**
Combined Comorbidity Score	0.001032	0.00059	0.001032	0.00059	**-1.63E-13**	**4.66E-13**
No. Ambulatory Visits	0.000073	0.000164	0.000073	0.000164	**-1.41E-14**	**-1.67E-14**

**N = 1,023,504, number of variables = 48*

K-party workflow

Both data partners contribute variable data to the regression analysis

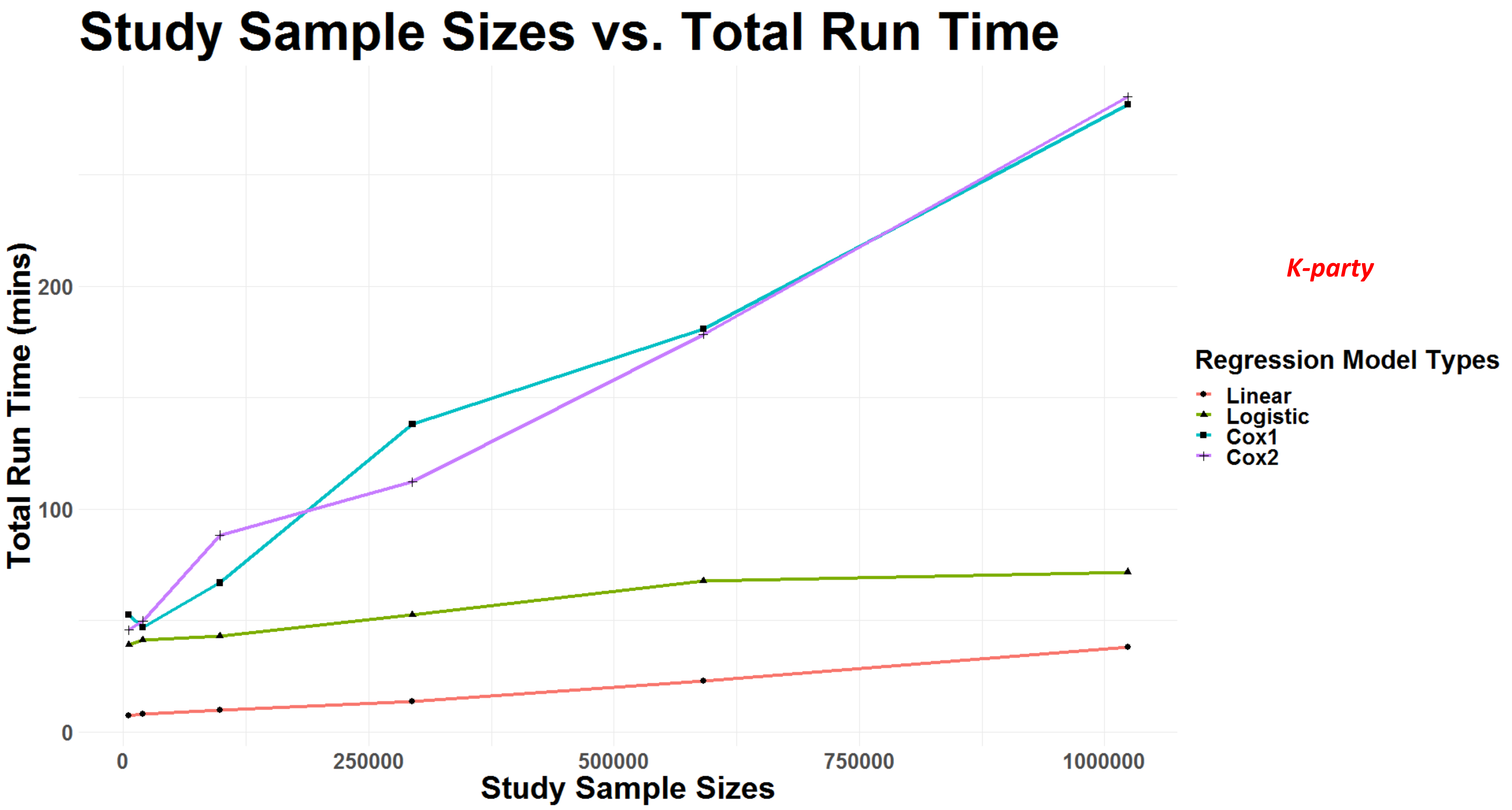
Results of 43 variables are not shown, all regression parameter estimates and standard errors are precise to the results obtained from the pooled individual-level data analysis (<10^{-10})

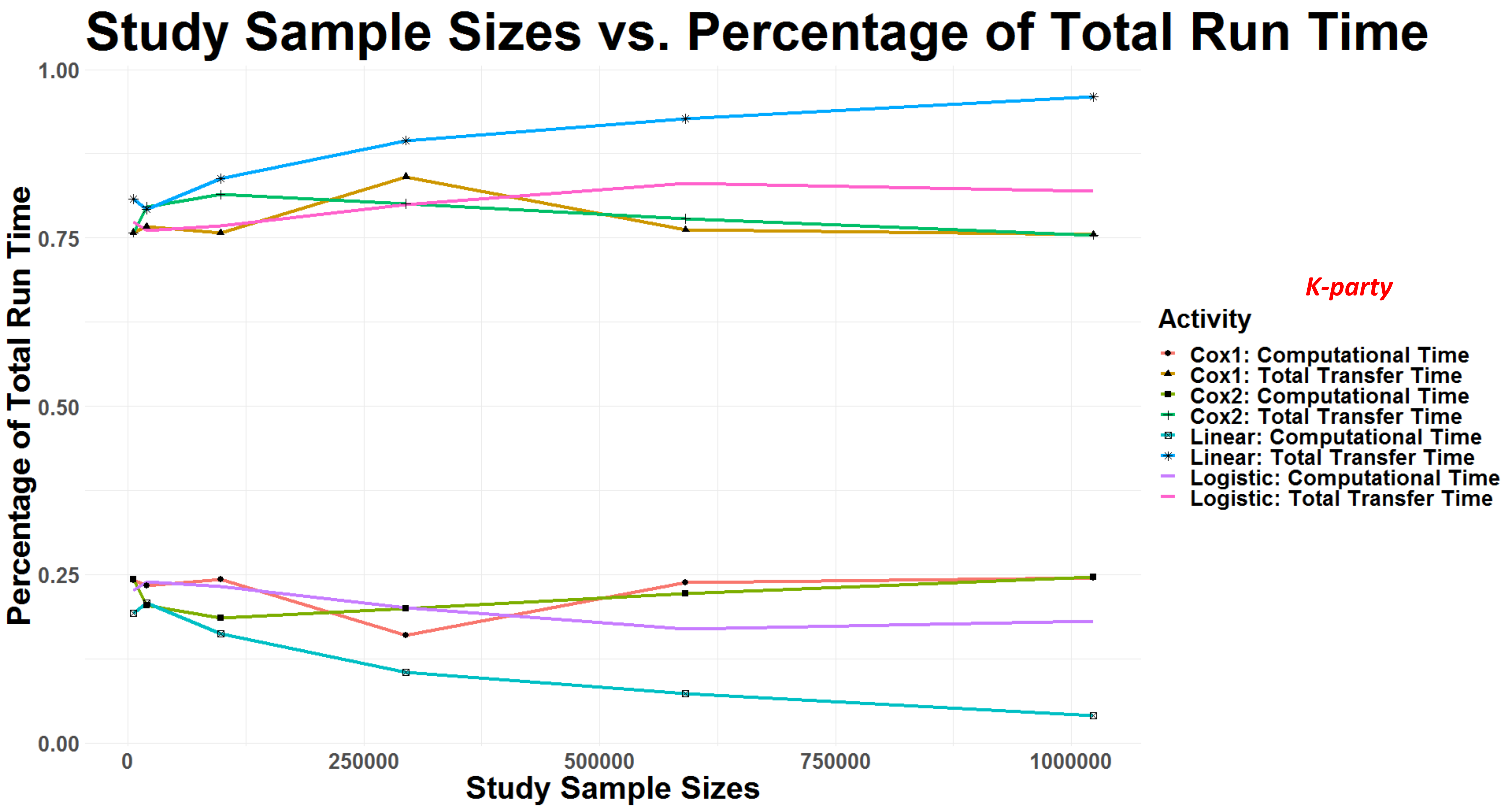
K-party

Operational Performance (N = 5,740) [Mean Time Elapses (Standard Deviation) (minutes)]				
	Linear	**Logistic**	**Cox1**	**Cox2**
Required number of data exchange cycles for model convergence	**2**	**19**	**23**	**23**
Average data exchange cycle time	**2.24 (1.04)**	**1.84 (0.38)**	**1.78 (0.77)**	**1.76 (0.74)**
Analysis Center				
Download Time	0.1 (0.09)	0.06 (0.05)	0.06 (0.04)	0.06 (0.04)
Computational Time	0.14 (0)	0.14 (0.01)	0.14 (0)	0.14 (0)
Upload Time	0.26 (0.07)	0.33 (0.06)	0.31 (0.05)	0.3 (0.05)
File Transfer Time (to Data Partners)	0.18 (0.05)	0.2 (0.03)	0.38 (0.24)	0.37 (0.23)
Data Partners				
Download Time	0.06 (0.02)	0.06 (0.02)	0.08 (0.14)	0.08 (0.15)
Computational Time	0.27 (0.14)	0.36 (0.14)	0.42 (0.2)	0.42 (0.19)
Upload Time	0.33 (0.09)	0.3 (0.04)	0.29 (0.07)	0.32 (0.15)
File Transfer Time (to Analysis Center)	0.37 (0.17)	0.34 (0.14)	0.31 (0.12)	0.32 (0.13)
File Transfer Time (to Data Partner)	0.4 (0.34)	0.41 (0.17)	0.32 (0.18)	0.32 (0.19)
Total Run Time (minutes)	**7.49**	**39.22**	**52.54**	**45.74**
Computational Time	19.2%	22.7%	24.2%	24.2%
File Transfer Process (Download, Upload, and Transfer)	**80.8%**	**77.3%**	**75.8%**	**75.8%**
*analysis excludes initial setup				

K-party

Operational Performance (N = 1,023,504) [Mean Time Elapses (Standard Deviation) (minutes)]	Linear	Logistic	Cox1	Cox2
Required number of data exchange cycles for model convergence	**2**	**22**	**23**	**23**
Average data exchange cycle time	**16.25 (20.8)**	**3.09 (4.79)**	**12.07 (13.4)**	**12.23 (13.53)**
Analysis Center				
Download Time	12.26 (17.26)	1.04 (4.08)	4.83 (7.93)	4.83 (7.86)
Computational Time	0.18 (0.07)	0.17 (0.04)	2.21 (3.88)	2.22 (3.9)
Upload Time	0.26 (0.11)	0.33 (0.06)	0.35 (0.06)	0.34 (0.05)
File Transfer Time (to Data Partners)	0.21 (0.05)	0.18 (0.03)	0.2 (0.08)	0.21 (0.09)
Data Partners				
Download Time	0.05 (0.01)	0.14 (0.16)	1.45 (4.28)	1.53 (4.41)
Computational Time	0.46 (0.22)	0.45 (0.17)	0.98 (1.93)	1.06 (2)
Upload Time	0.8 (1.23)	0.33 (0.32)	0.72 (0.82)	0.68 (0.83)
File Transfer Time (to Analysis Center)	1.15 (1.74)	0.39 (0.35)	1.75 (3.97)	1.77 (3.93)
File Transfer Time (to Data Partner)	0.43 (0.26)	0.37 (0.32)	0.55 (0.33)	0.71 (0.45)
Total Run Time (minutes)	**38.29**	**71.75**	**281.32**	**284.85**
Computational Time	4.1%	18.1%	24.5%	24.6%
File Transfer Process (Download, Upload, and Transfer)	**95.9%**	**81.9%**	**75.5%**	**75.4%**
*analysis excludes initial setup				


Study Sample Sizes vs. Total Run Time
Total Run Time (mins)
200
100
0
0
250000
500000
750000
1000000
Study Sample Sizes
K-party
Regression Model Types
Linear
Logistic
Cox1
Cox2


Study Sample Sizes vs. Percentage of Total Run Time
Percentage of Total Run Time
Study Sample Sizes
1.00
0.75
0.50
0.25
0.00
0
250000
500000
750000
1000000
K-party
Activity
Cox1: Computational Time
Cox1: Total Transfer Time
Cox2: Computational Time
Cox2: Total Transfer Time
Linear: Computational Time
Linear: Total Transfer Time
Logistic: Computational Time
Logistic: Total Transfer Time

Discussion

- We are able to perform distributed regression analysis with vertically partitioned data without the need to share any individual-level data
- Regression parameters and standard errors are precise compared to estimates obtained from regression analysis with the pooled individual-level data
- We made moderate enhancements to PopMedNet[TM] to support the R-based application
- We have a functional prototype and future work is needed to evaluate the R-based application in a real world setting

References

- Her Q, Malenfant J, Malek S, et al. A query workflow design to perform automatable distributed regression analysis in large distributed data networks. *EGEMS (Wash DC).* 2018a;**6**(1):11 doi: http://doi.org/10.5334/egems.209.
- Her Q, Malenfant J, Vilk Y, et al. Utilizing Data from Various Data Partners in a Distributed Manner. 2018c. https://www.sentinelinitiative.org/sentinel/methods/utilizing-data-various-data-partners-distributed-manner. Accessed 01/01/2019.
- Her QL, Vilk Y, Young J, et al. A distributed regression analysis application based on SAS software. Part I: Linear and logistic regression. ArXiv e-prints 2018b. https://ui.adsabs.harvard.edu/#abs/2018arXiv180802387H (accessed April 15, 2019).
- Fienberg SE, Fulp WJ, Slavkovic AB, Wrobel TA. “Secure” log-linear and logistic regression analysis of distributed databases. Privacy in Statistical Databases: Springer, 2006:277-90.
- Fienberg SE, Nardi Y, Slavković AB. Valid Statistical Analysis for Logistic Regression with Multiple Sources. 2009. In: Gal CS, Kantor PB, Lesk ME. Protecting Persons While Protecting the People. Springer. Berlin, Heidelberg.
- Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP. Secure analysis of distributed chemical databases without data integration. *J Comput Aided Mol Des.* 2005;**19**(9-10):739-47 doi: 10.1007/s10822-005-9011-5.
- Karr AF, Fulp WJ, Vera F, Young SS, Lin X, Reiter JP. Secure, Privacy-Preserving Analysis of Distributed Databases. *Technometrics.* 2007;**49**(3):335-45 doi: 10.1198/004017007000000209.
- Lu CL, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc.* 2015;**22**(6):1212-9 doi: 10.1093/jamia/ocv083.
- Toh S, Gagne JJ, Rassen JA, et al. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. *Med Care.* 2013;51(8 Suppl 3):S4-10.

Questions?

Qoua_Her@harvardpilgrim.org

Darren_Toh@harvardpilgrim.org

https://www.distributedanalysis.org/

DEPARTMENT OF POPULATION MEDICINE

HARVARD MEDICAL SCHOOL

Harvard Pilgrim Health Care Institute

PPM Privacy-Protecting Methods