# CURIE: Policy-based Secure Data Exchange

Z. Berkay Celik
SIIS Laboratory, Department of CSE
The Pennsylvania State University
zbc102@cse.psu.edu

Abbas Acar, Hidayet Aksu
CPS Security Lab, Department of ECE
Florida International University
aacar001,haksu@fiu.edu

Ryan Sheatsley
SIIS Laboratory, Department of CSE
The Pennsylvania State University
rms5643@cse.psu.edu

Patrick McDaniel
SIIS Laboratory, Department of CSE
The Pennsylvania State University
mcdaniel@cse.psu.edu

A. Selcuk Uluagac
CPS Security Lab, Department of ECE
Florida International University
suluagac@fiu.edu

## ABSTRACT

Data sharing among partners—users, companies, organizations—is crucial for the advancement of collaborative machine learning in many domains such as healthcare, finance, and security. Sharing through secure computation and other means allow these partners to perform privacy-preserving computations on their private data in controlled ways. However, in reality, there exist complex relationships among members (partners). Politics, regulations, interest, trust, data demands and needs prevent members from sharing their complete data. Thus, there is a need for a mechanism to meet these conflicting relationships on data sharing. This paper presents CURIE[1], an approach to exchange data among members who have complex relationships. A novel policy language, CPL, that allows members to define the specifications of data exchange requirements is introduced. With CPL, members can easily assert who and what to exchange through their local policies and negotiate a global sharing agreement. The agreement is implemented in a distributed privacy-preserving model that guarantees sharing among members will comply with the policy as negotiated. The use of CURIE is validated through an example healthcare application built on recently introduced secure multi-party computation and differential privacy frameworks, and policy and performance trade-offs are explored.

## CCS CONCEPTS

• **Information systems** → **Data exchange**; • **Security and privacy** → *Economics of security and privacy*.

## KEYWORDS

Collaborative learning; policy language; secure data exchange

## 1 INTRODUCTION

Inter-organizational data sharing is crucial to the advancement of many domains including security, health care, and finance. Previous works have shown the benefit of data sharing within distributed, collaborative, and federated learning [5, 12, 37]. Privacy-preserving machine learning offers data sharing among multiple members while avoiding the risks of disclosing the sensitive data (e.g., healthcare records, personally identifiable information) [14]. For example, secure multiparty computation enables multiple members, each with its training dataset, to collaboratively learn a shared predictive model without revealing their datasets [31]. These approaches solve the privacy concerns of members during model computation, yet do not consider the complex relationships such as regulations, competitive advantage, data sovereignty, and jurisdiction among members on private data sharing. Members want to be able to articulate and enforce their conflicting requirements on data sharing.

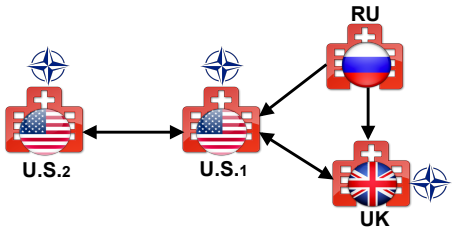


**Figure 1: An illustration of data exchange requirements of countries learning a predictive model on their shared data. Arrows show the data requirements of countries.**

To illustrate such complex data sharing requirements, consider health care organizations that collaborate for a joint prediction model of diagnosis of patients experiencing blood clots (see Figure 1). Members wish to dictate their needs through their legal and political limitations as follows: U.S.$_1$ is able to share its complete data for nation-wide members (U.S.$_2$) [3, 23], yet it is obliged to share the data of patients deployed in NATO countries with NATO members (UK) [17]. However, U.S.$_1$ wishes to acquire all patient data from other countries. UK is able to share and acquire complete data from NATO members, yet it desires to acquire only data of certain race groups from U.S$_1$ to increase its data diversity. RU wishes to share and acquire complete data from all members, yet members limit their data share to Russian citizens who live in their countries. Such complex data sharing requirements also commonly occur today in non-healthcare systems [28, 38]. For instance, National Security Agency has varying restrictions on how human intelligence is shared with other countries; financial companies share data based on trust, and competition among each other.

This paper presents a policy-based data exchange approach, called CURIE, that allows secure data exchange among members

[1]Our paper named after Marie Curie. She is physicist and chemist who conducted pioneering research in health care and won Nobel prize twice.


*CODASPY '19, March 25–27, 2019, Richardson, TX, USA*

ACM ISBN 978-1-4503-6099-9/19/03…$15.00
https://doi.org/10.1145/3292006.3300042

that have such complex relationships. Members specify their requirements on data exchange using a policy language (CPL). The requirements defined with the use of CPL form the local data exchange policies of members. Local policies are defined separately for data sharing and data acquisition policies. This property allows asymmetric relations on data exchange. For example, a member does not necessarily have to acquire the data that the other members dictate to share. By using these two policies, members specify statements of who to share/acquire and what to share/acquire. The statements are defined using *conditional* and *selection* expressions. Selections allow members to filter data and limit the data to be exchanged, whereas conditional expressions allow members to define logical statements. Another advanced property of CPL is predefined *data-dependent conditionals* for calculating the statistical metrics between member's data. For instance, members can define a conditional to compute the intersection size of data columns without disclosing their data. This allows members to define content-dependent conditional data exchange in their policies.

Once members have defined their local policies, they negotiate a sharing agreement. The guarantee provided by Curie is that all data exchanged among members will respect the agreement. The agreement is executed in a multi-party privacy-preserving prediction model enhanced with optional differential privacy guarantees. In this work, we make the following contributions:

- We introduce Curie, an approach for secure data exchange among members that have complex relationships. Curie includes CPL policy language allowing members to define complex specifications of data exchange requirements, negotiate an agreement, and execute agreements in a multi-party predictive model that policies respect the negotiated policy.
- We validate Curie through an example of real healthcare application used to prescribe warfarin dosage. A privacy-preserving joint dose model among medical institutions is compiled with the use of various data exchange policies while protecting the privacy of members' healthcare records.
- We show Curie incurs low overhead and policies are effective at improving the dose accuracy of medical institutions.

We begin in the next section by defining the analysis task and outlining the security and attacker models.

## 2 PROBLEM SCOPE AND ATTACKER MODEL

**Problem Scope.** We introduce Curie Policy Language (CPL) to express data exchange requirements of distributed members. Unlike the programming languages used for writing secure multi-party computation (MPC) [24, 33] and the frameworks designed for privacy-preserving machine learning (ML) [7, 14, 29, 31, 32], CPL is a policy language in a Backus Normal Form (BNF) notation to express the conflicting relationships of members on data sharing. Members can express data exchange requirements using the conditionals, selections, and secure pairwise data-dependent statistics. Curie then enforces the policy agreements in a shared predictive model through an MPC protocol that ensures members comply with the policies as negotiated.

We integrate Curie into 24 medical institutions. Without deployment of Curie, institutions compute warfarin dosage of a patient using a model computed on their local patient records. Curie allows institutions to construct various consortia wherein each member defines a data exchange policy for other members via CPL. This

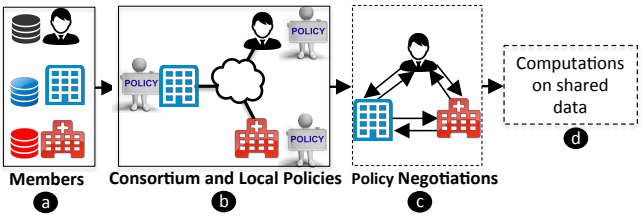


**Figure 2: Curie data exchange process in a collaborative learning setting. The dashed boxes show data remains confidential.**

enables institutions to acquire the patient records based on regulations as well as the records that they need to improve the accuracy of their dose predictions. Curie implements a privacy-preserving dose model through homomorphic encryption (HE) to enforce the policy agreements of the members. We note that a centralized party in HE cannot provide a privacy-preserving model on negotiated data [39]. However, Curie implements a novel protocol that allows institutions to perform local computations by aggregating the intermediate results of the dose model. Additionally, Curie implements an optional differential private (DP) mechanism that allows institutions to perform differentially-private (DP) secure dose model. DP guarantees that no information leaks on the targeted individual (i.e., patient) with high confidence from the released dose model.

**Threat Model.** We consider a semi-honest adversary model. That is, members in a consortium runs the protocol exactly as specified, yet they try to learn the dataset inputs of the other members as much as possible from their views of the protocol. Additionally, we consider non-adaptive adversary wherein members cannot modify inputs of their dataset once the protocol on shared data is initiated.

## 3 ORGANIZATIONAL DATA EXCHANGE

Depicted in Figure 2, Curie includes two independent parts: policy management and multiparty secure computation.

**Policy Management.** We define a *consortium* that is a group made up of two or more members–individuals, companies or governments (ⓐ). Members of a consortium aim to compute a predictive model $m$ over their confidential data in a secure manner. For instance, data may be curated from medical history of patients or financial reports of companies with the objective of building an ML model. Moreover, each member wants to enforce a set of local constraints toward other consortium members to control their requirements on how and with whom they share their confidential data. These constraints define a member's interest, trust, regulations and data demands, and also impacts the accuracy of a model $m$. Thus, there is a need for connecting data needs of members to the privacy-preserving models. In Curie, each member of a consortium defines a *local policy* (ⓑ). The local policy of a member dictates the requirements of data exchange as follows:

(1) The member wishes to specify with whom to share and acquire data (*partnership requirement*).
(2) The member wishes to define what data to share and acquire (*sharing and acquisition requirement*).

In this, the member wishes to refine its sharing and acquisition requirements to express the following:

(1) The member wishes to dictate a set of conditions to restrict data sharing and select which data to be acquired (*conditional selective share and acquisition*); and

(2) The member wishes to dictate conditionals based on the other member's data (*data-dependent conditionals*).

The policy of members need not be-nor are likely to be-symmetric. Local policy is defined with requirements for sharing and acquisition that is tailored to each partner member in the consortium–thus allowing each pairwise sharing to be unique. Here, the local policies are used to negotiate pairwise sharing within the consortium. To illustrate how members negotiate an agreement, consider the consortium of three members in Figure 3.

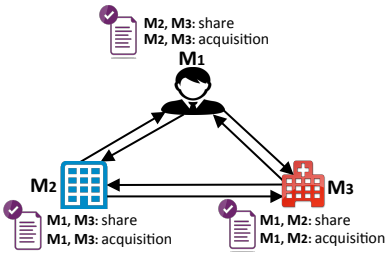


**Figure 3: An example consortium of three members.**

Each member initiates pairwise policy negotiations with other members to reconcile contradictions between acquisition and share policies (ⓒ). A member starts the negotiation by sending a request message including the acquisition policy defined for a member. When a member receives the acquisition policy, it reconciles the received acquisition policy with its share policy specified for that member. Three negotiation outcomes are possible: the acquisition policy is entirely satisfied, partially satisfied with the intersection of acquisition and share policies or is an empty set. A member completes its negotiations after all of its acquisition policies for interested parties are negotiated.

**Computations on Negotiated Data.** Once members negotiate their policies (ⓓ), Curie provides a multiparty data exchange device using secure multi-party computation techniques enhanced with (optional) differential privacy guarantees. This device ensures data and individual privacy. The guarantee provided by Curie is that all computations among members will respect their policies.

To ensure data privacy, Curie includes cryptographic primitives such as Homomorphic Encryption (HE) and garbled circuits from the secure multi-party computation literature that allows members to perform computations on negotiated data with no disclosed data from any single member. At the end of the secure computation, all of the parties obtain a final predictive model based on their policy negotiations. To ensure the privacy of the individuals in the dataset, which the final model is computed on, Curie integrates Differential Privacy (DP). DP protects against an attacker who tries to extract a particular individual's data in the dataset from the final computed model at the end of the secure computation protocol.

## 4 CURIE POLICY DESCRIPTION LANGUAGE

We now illustrate the format and semantics of the Curie Policy Language (CPL). A BNF description of CPL is presented in Appendix A. Turning to the example consortium in Figure 3 established with three members, each member defines its requirements for other members on a dataset having the columns of age, race, genotype, and weight (see Table 1). The criteria defined by members are used throughout to construct their local policies.

| | |
|---|---|
| | **Consortia member: $M_1$** |
| $M_2$– | desires to acquire complete data of users who are older than 25 |
| $M_2$– | shares its complete data |
| $M_3$– | desires to acquire Asian users such that the Jaccard similarity of its age column and $M_3$'s age column is greater than 0.3 |
| $M_3$– | shares its complete data |
| | **Consortia member: $M_2$** |
| $M_1$– | desires to acquire complete data |
| $M_1$– | limits its share to EU and NATO citizen users if $M_1$ is both NATO and EU member and located in North America. Otherwise, it shares only White users |
| $M_3$– | desires to acquire complete data if $M_3$ is a NATO member |
| $M_3$– | shares its complete data |
| | **Consortia member: $M_3$** |
| $M_1$– | desires to acquire complete data of users having genotype 'A/A' |
| $M_1$– | share complete data if intersection size of its and $M_1$'s genotype column is less than 10. Otherwise, it shares data of users that weigh more than 100 pounds |
| $M_2$– | desires to acquire complete data |
| $M_2$– | shares complete data if $M_2$ is EU member and its data size is greater than 1K |

**Table 1: An example of member's data exchange requirements.**

**Share and Acquisition Clauses.** Curie policies are collections of clauses. The collection of clauses for partners defines the local policy of a member. The clauses allow each member to dictate a member specific policy for each other member. Clauses have the following structure:

⟨clause tag⟩ : ⟨members⟩ : ⟨conditionals⟩ :: ⟨selections⟩;

Clause tags are reference names for policy entries. *Share* and *acquire* are two reserved tags. Those clauses are comprised of three parts. The first part, *members*, defines a list of members with whom to share and acquire. This can be a single member or a comma-separated list of members. An empty member entry matches all members. The second part, *conditionals*, is a list of conditions controlling when this clause will be executed. A condition is a Boolean function which expresses whether the share or acquire is allowed or not. For instance, a member may define a condition where the data size is greater than a specific value. Only if all conditions listed in conditionals are true, then this clause is executed. Last part, *selections*, states what to share or acquire. It can be a list of filters on a member's data. For instance, a member may define a filter on a column of a dataset to limit acquisition to a subset of the dataset. More complex selections can be assigned using member defined sub-clauses. A sub-clause has the following structure:

⟨tag⟩ : ⟨conditionals⟩ :: ⟨selections⟩;

where *tag* is the name of sub-clause; conditionals is, as explained above, a list of conditions stating whether this clause will be executed; selections is a list of filters or a reference to a new sub-clause. Complex data selection can be addressed with nested sub-clauses.

CPL allows members to define multiple clauses. For instance, a member may share a distinct subset of data for different conditions. CPL evaluates multiple clauses in a top-down order. When conditionals of a clause evaluate to false, it moves to the next clause until a clause is matched or it reaches end of the policy file.

**Conditionals and Selections.** We present the use of conditionals and selections through policies with examples. Their format and semantics are detailed. Consider an example of two members, $M_1$ and $M_2$, within a consortium. They define their local policies as:

@$M_1$ acquire : $M_2$ : :: $s_1$ ;
share : $M_2$ : :: ;
@$M_2$ acquire : $M_1$ : :: ;
share : $M_1$ : $c_1$, $c_2$ :: fine-select ;
fine-select : $c_3$ :: $s_2$ ;
fine-select : :: $s_3$ ;

where $c_1$, $c_2$ and $c_3$ are conditionals, $s_1$, $s_2$ and $s_3$ are selections and fine-select is a tag defined by $M_2$.

The acquire clause of $M_1$ states that data is requested from $M_2$ after it applies $s_1$ selection (e.g., age > 25) to its data. In contrast, its share clause allows complete share of its data if $M_2$ requests. On the other hand, the acquisition clause of $M_1$ dictates requesting complete data from $M_2$. However, $M_2$ allows data sharing if the acquisition clause issued by $M_1$ holds $c_1 \wedge c_2$ conditions (e.g., is both NATO and EU member). Then, $M_2$ delegates selection to member-defined fine-select sub-clauses. fine-select states that if the request satisfies the $c_3$ condition (located in North America) then the request is met with the data that is selected by the $s_2$ selection (e.g., limits share of its data to NATO and EU member country citizens). Otherwise, it shares data that is specified by selection $s_3$ (White users).

CPL supports selections through filters. A filter contains zero or more operations over data inputs describing the share and acquisition criteria to be enforced. Operations are defined as keywords or symbols such as <, >, =, *in*, *like*, and so on. Selections and filters are defined in CPL as follows:

⟨selections⟩ ::= <filters> | <tag>
<filters> ::= <filter> [',' <filters>]
<filter> ::= <var> <operation> <value> | ''

Selections are executed when conditionals evaluated to be true. Conditionals can be consortium and dataset-specific. For instance, a member may require other members to be in a particular country or to be in an alliance such as NATO and to have their dataset size greater than a particular value. Such conditionals do not require any data exchange between members to be evaluated. However, members may want to incorporate a relation between their data and other member's data into their policies as detailed next.

**Data-dependent Conditionals.** A member's decision on whether to share or to acquire data can depend on other member's data. Simply put, one example of a data-dependent conditional among two members could be whether the intersection size of the two sets (e.g., a specific column of a dataset) is not too high. Considering such knowledge, a member can make a conditional decision about share or acquisition of that data. For instance, consider a list of private IP addresses used for blacklisting the domains. If a member knows that the intersection size is close to zero, then the member may dictate an acquire clause to request complete features from that member based on IP addresses [18].

CPL defines an evaluate keyword for data-dependent conditionals through functions on data. Data-dependent conditionals take the following form:

⟨conditionals⟩ ::= <var>'='<value> [',' <conditionals>]
| 'evaluate' '(' <data_ref> ',' <alg_arg> ',' <thshold_arg> ')' [',' <conditionals>] | ''

A member that uses the data-dependent conditionals defines a reference data (data_ref) required for a such computation, an algorithm (alg_arg) and a threshold (thshold_arg) that is compared with the output of the computation. CPL includes four algorithms for data-dependent conditionals (see Table 2). To be brief, intersection size measures the size of the overlap between two sets; Jaccard index is a statistic measure of similarity between sets; Pearson correlation is a statistical measure of how much two sets are linearly dependent; and Cosine similarity is a measure of similarity between two vectors. Each algorithm is based on a different assumption about the underlying reference data. However, central to all of them is to privately (without leaking any sensitive data) measure a relation between two members' data to offer an effective data exchange. We note that these algorithms are found to be effective in capturing input relations in datasets [18, 19].

| Pairwise alg. | Output | Private protocol | Proof |
|---|---|---|---|
| Intersection size | $\lvert\mathcal{D}_i \cap \mathcal{D}_j\rvert$ | Intersection cardinality | [11] |
| Jaccard index | $(\lvert\mathcal{D}_i \cap D_j\rvert)/(\lvert\mathcal{D}_i \cup \mathcal{D}_j\rvert)$ | Jaccard similarity | [6] |
| Pearson correlation | $(COV(\mathcal{D}_i, \mathcal{D}_j))/(\sigma_{\mathcal{D}_i}\sigma_{\mathcal{D}_j})$ | Garbled circuits | [25] |
| Cosine similarity | $(\mathcal{D}_i\mathcal{D}_j)/(\lVert\mathcal{D}_i\rVert\lVert\mathcal{D}_j\rVert)$ | Garbled circuits | [25] |

**Table 2: CPL data-dependent conditional algorithms. Two members of a consortium use the conditionals to compute the pairwise statistics. The members then use the output of the algorithm to determine whether to acquire or share data from another party. ($\mathcal{D}_i$ and $\mathcal{D}_j$ are the inputs of a dataset, and $\sigma$ is std. deviation).**

Data-dependent conditionals are implemented through private protocols (as defined in Table 2). These protocols are implemented with the cryptographic tools of garbled circuits and private functions. Protocols preserve the confidentiality of data. That is, each member gets the output indicated in Table 2 without revealing their sensitive data in plain text. After the private protocol terminates, the output of the algorithm is compared with a threshold value set by the requester. If the output is below the threshold value, the conditional is evaluated to true. Turning to above example $M_3$ joins the consortium. $M_1$ and $M_2$ extend their local policies for $M_3$:

```
@M1 acquire : M3 : evaluate(local data, 'Jaccard', 0.3) :: race=Asian;
      share : M3 : :: ;
@M2 acquire : M3 : M3 in $NATO :: ;
      share : M3 : :: ;
@M3 acquire : M1 : :: Genotype = 'A/A' ;
      share : M1 : evaluate(local data,'intersection size', 10) :: ;
      share : M1 : :: weight>150 ;
      acquire : M2 : :: ;
      share : M2 : M2 in $EU, size(data)> 1K :: ;
```

The acquire clause of $M_1$ defines a data-dependent conditional for $M_3$. It defines a Jaccard measure on its local data through evaluate keyword and sets its threshold value equal to 0.3. $M_3$ agrees to share its local data with $M_1$ if intersection size of its local data is less then 10. Otherwise, it consults the next share clause defined for $M_1$ which states that an individual's weight greater than 150 pounds will be shared. All other share and acquire clauses are trivial. Members agree to share and acquire complete data based on data size (data size > 1K), alliance membership (e.g., NATO or EU member) and inputs (e.g., genotype).

Putting pieces together, CPL allows members independently define a data exchange policy with share and acquire clauses. The policies are dictated through conditionals and selections. This allows members to dictate policies in complex and asymmetric relationships. Defined in Section 3, CPL provides members to dictate partnership, share, acquisition, and data-dependent conditionals.

**Policy Negotiation and Conflicts.** Data exchange between members is governed by matching share and acquire clauses in each member's respective policies. Both share and acquire clauses state conditions and selections on the data exchanged. Consider two example local policies with a share clause $@m_2$ ($share : m_1 : c_1 :: s_1$) and matching acquire clause $@m_1$ ($acquire : m_2 : c_2 : s_2$). Curie's negotiation algorithm respects both autonomy of the data owner and the needs of the requester. It conservatively negotiates share

| Policy ID | Consortium Name | Policy Definition | Acquisition Policy | Share Policy |
|---|---|---|---|---|
| P.1 | Single Source | Each member uses its local patient dataset to learn warfarin dose model. | ✗ | ✗ |
| P.2 | Nation-wide | Members in the same country establish a consortium based on state and country laws. | ✓ | ✓ |
| P.3 | Regional | Members in the same continent establish a consortium. | ✓ | ✓ |
| P.4 | NATO-EU | NATO and EU members establish a consortium independently based on their mutual agreements. | ✓ | ✓ |
| P.5 | Global | Members exchange their complete data to build the warfarin dose model. | ✓ | ✓ |

**Table 3: Consortia constructed among members. Acquisition and share policies of members for each consortium are studied in Section 6.**

and acquire clauses such that it will return the *intersection* of respective data sets in resulting policy assignment. The resolved policy in this example is $share : m_1 : c_1 \wedge c_2 :: s_1 \wedge s_2$ which states that the data exchange from $m_2$ to $m_1$ is subject to both $c_1$ and $c_2$ conditionals and resulting sharing has $s_1$ and $s_2$ selections on $m_2$'s data. This authoritative negotiation makes sure no member's data is shared beyond its explicit intent, regardless how the other members' policies are defined. This is because negotiation fulfilling the criteria for each clause is based on the union of logical expressions defined in two policies. Each member runs the negotiation algorithm for members found in their member list. After all members terminate their negotiations, the negotiated policy is enforced in computations.

## 5 DEPLOYMENT OF CURIE

To validate Curie in a real application, we integrated Curie into 24 medical institutions. Each institution wants to compute a warfarin dose model on the distributed dataset without disclosing the patient health-care records. Without deployment of Curie, institutions compute warfarin dosage of a patient using a model computed on their local patient data. Curie first enables institutions to negotiate their data exchange requirements through CPL. In this, Curie allows members to construct various consortia wherein each member defines a data exchange policy for other members. The next step is to compute a privacy-preserving dose model such that each party does not learn any information about the patient's records of other medical institutions and respects the policy negotiated. Curie implements a secure dose protocol through homomorphic encryption (HE) to enforce the policy agreements of the members. We next present the deployment of Curie to institutions (Section 5.1) and integration of policy agreements in warfarin dose model (Section 5.2).

### 5.1 Deployment Setup

**Warfarin-** known as the brand name Coumadin is a widely prescribed (over 20 million times each year in the United States) anticoagulant medication. It is mainly used to treat (or prevent) blood clots (thrombosis) in veins or arteries. Taking high-dose warfarin causes thin blood which may result in intracranial and extracranial bleeding. Taking low doses causes thick blood which may result in embolism and stroke. Current clinical practices suggest a *fixed* initial dose of 5 or 10 mg/day. Patients regularly have a blood test to check how long it takes for blood to clot (international normalized ratio (INR)). Based on the INR, subsequent doses are adjusted to maintain the patient's INR at the desired level. Therefore, it is important to predict the proper warfarin dose for the patients.

**Consortium Members.** 24 medical institutions from nine countries and four continents individually collected the largest patient data for predicting *personalized* warfarin dose (see Appendix D for details of members involved in the study). Members collect 68

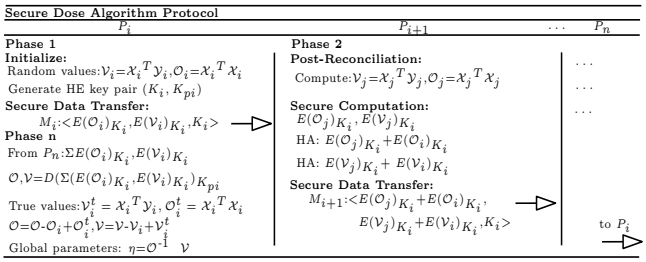


**Figure 4: Secure dose algorithm protocol: Member ($P_i$) starts the protocol, the procedures and message flow among members are highlighted in boldface. At the final phase, $P_i$ is able to compute the dose model coefficients from the negotiated data.**

inputs from patients' genotypic, demographic, background information, yet a long study concluded that eight inputs are sufficient for proper prescriptions [26].

**Warfarin Dose Prediction Model.** To determine the proper personalized warfarin dosage, a long line of work concluded with an algorithm of an ordinary linear regression model [26]. The model is a function $f : \mathcal{X} \to \mathcal{Y}$ that aim at predicting targets of warfarin dose $y \in \mathcal{Y}$ given a set of patient inputs $x \in \mathcal{X}$. We represent the patient dataset of each member $\mathcal{D}_i = \{(x_i, y_i)\}_{i=1}^n$, and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. The loss function penalizes deviations between true dose and predictions. Learning is then searching for a dose model $f$ minimizing the average loss:

$$\mathcal{L}(\mathcal{D}, f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i). \quad (1)$$

The dose model reduces to minimizing the average loss $\mathcal{L}(\mathcal{D}, f)$ with respect to the parameters of the model $f$. The model is linear, i.e., $f(x) = \alpha^\top x + \beta$, and the loss function is the squared loss $\ell(f(x), y) = (f(x) - y)^2$. The dose model gives as well or better results than other more complex numerical methods and outperforms fixed-dose approach[2] [26]. We re-implemented the algorithm in Python by direct translation from the authors' implementation and found that the accuracy of our implementation has no statistically significant difference.

**Consortia and Member Policies.** We define consortia among medical institutions that they state partnerships for data exchange. Table 3 summarizes the consortia. The consortia are defined based on statute and regulations between members, as well as regional, and national partnerships are studied based on their countries [3, 17, 23, 34]. For example, NATO allied medical support doctrine allows strategic relationships that are otherwise not obtainable by non-NATO members. Each member in a consortium exchanges data with

[2]The model has been released online http://www.warfarindosing.org to help doctors and other clinicians for predicting ideal dose of warfarin.

other members based on its CPL policy. Various acquisition and share policies of CPL are studied via conditionals and selections in Section 6. We note that policy construction is a subjective enterprise. Depending on the nature and constraints of a given environment, any number of policies are appropriate. Such is the promise of policy defined behavior; alternate interpretations leading to other application requirements can be addressed through CPL.

## 5.2 Privacy-preserving Dose Prediction Model

The computation of *local dose* model of a medical institution is straightforward: a member calculates the dose model through Equation 2 with the use of patient data collected locally. To implement a privacy-preserving dose model among consortia members of medical institutions, we define the dose prediction formula stated in Equation 1 in a matrix form by minimizing with maximum likelihood estimation:

$$\beta = (\mathcal{X}^\intercal \mathcal{X})^{-1} \mathcal{X}^\intercal \mathcal{Y}, \tag{2}$$

where $\mathcal{X}$ is the input matrix, $\mathcal{Y}$ is the dose matrix, and $\beta$ is the coefficients of the dose model.

Curie allows members to collaboratively learn a dose model without disclosing their patient records and guarantees data sharing complies with the policy as negotiated. As shown in Equation 3, each member translates its negotiated data into neutral input matrices [41]. Particularly, patient samples to be exchanged by each member are computed as an input matrix $\mathcal{X}_0, \ldots, \mathcal{X}_n$ and dose matrix $\mathcal{Y}_0, \ldots, \mathcal{Y}_n$. The transformation defines each member's *local statistics* $\mathcal{O}_i = \mathcal{X}^\intercal \mathcal{X}$ and $\mathcal{V}_i = \mathcal{X}^\intercal \mathcal{Y}$. Local statistics is the output of the negotiation of each member in a consortium. The aggregation of the local statistics corresponds to a *negotiated dataset* which is the exact amount that a member negotiates to obtain from other members in a consortium. Curie constructs the dose algorithm of the negotiated dataset as a concatenation of members' local statistics as follows:

$$\begin{aligned} \mathcal{X}^\intercal \mathcal{X} &= \left[\mathcal{X}_1^\intercal | \ldots | \mathcal{X}_n^\intercal\right] \left[\mathcal{X}_1 | \ldots | \mathcal{X}_n\right]^\intercal = \sum_{i=1}^{n} \mathcal{X}_i^\intercal \mathcal{X}_i = \sum_{i=1}^{n} \mathcal{V}_i = \mathcal{V} \\ \mathcal{X}^\intercal \mathcal{Y} &= \left[\mathcal{X}_1^\intercal | \ldots | \mathcal{X}_n^\intercal\right] \left[\mathcal{Y}_1 | \ldots | \mathcal{X}_n\right]^\intercal = \sum_{i=1}^{n} \mathcal{X}_i^\intercal \mathcal{Y}_i = \sum_{i=1}^{n} \mathcal{O}_i = \mathcal{O} \end{aligned} \tag{3}$$

In Equation 3, a member computes model coefficients using the sum of other members local statistics. The local statistics includes $m \times m$ constant matrices where $m$ is the number inputs (independent of number of dataset size). Using this observation, a party computes the coefficients of the negotiated dataset:

$$\eta^{(negotiated)} = (\mathcal{X}^\intercal \mathcal{X})^{-1} \mathcal{X}^\intercal \mathcal{Y} = \mathcal{O}^{-1} \mathcal{V} \tag{4}$$

In Equation 4, while the accuracy objective of the dose model is guaranteed using the coefficients obtained from the sum of local statistics, the exchange of clear statistics among parties may leak information about members' data. A member can infer knowledge about the distribution of each input of other members from matrices of $\mathcal{O}_i$ and $\mathcal{V}_i$ [14]. Furthermore, an adversary may sniff data traffic to control and modify exchanged messages. To solve these problems, we use homomorphic encryption (HE) that allows computation on ciphertexts [2]. HE allows members to perform the computation of joint of function without requiring additional communication complexity other than the data exchange. We note that HE itself cannot preserve the confidentiality of data from multiple parties in centralized settings [40]. However, Curie implements a distributed privacy-preserving multi-party dose model, as shown in Figure 4.

To illustrate, we consider an example session of $n$ members authorized for data exchange in a consortium. In this example, a ring topology is used for secure group communication (i.e., $P_i$ talks to $P_{i+1}$, and similarly $P_n$ talks to $P_i$). $P_1$ initially generates a pair of encryption keys using the homomorphic cryptosystem and broadcasts the public key to the members in its member list. $P_1$ then generates random $\mathcal{V}_i$, $\mathcal{O}_i$ and encrypts them $E(\mathcal{O}_i)_{K_i}$ and $E(\mathcal{V}_i)_{K_i}$ using its public key $K_i$. It starts the session by sending them to the next member in the ring. When next member receives the encrypted message, it adds its local $\mathcal{V}_i$ and $\mathcal{O}_i$ matrices through homomorphic addition to the output of its policy reconciliation for $P_1$ and passes to the next member. Remaining members take the similar steps. Secure computation executes one round per member in which the computation for the particular member visits other members. This allows Curie to enforce HE on shared data of a particular member in each round uses and does not suffer insecurities associated with centralized HE constructions [40].

At the final stage of the protocol, $P_1$ receives the sum statistics of $\mathcal{O}_i$ and $\mathcal{V}_i$ from $P_n$. $P_1$ decrypts the sum of the statistics using its private key and then subtracts the initial random values of $\mathcal{V}_i$, $\mathcal{O}_i$ and adds its true values used for computation of the local dose model coefficients. The final result $\mathcal{O}$ and $\mathcal{V}$ is the coefficients of the dose model that respects $P_1$'s policy negotiations. Other consortium members similarly start the protocol and compute the coefficients. We present the security analysis of the dose protocol in Appendix C, and show its differentially-private extension in Appendix B.

# 6 EVALUATION

This section details the operation of the Curie through policies. We show how flexible data exchange policies are implemented and operated. We focus on the following questions:

(1) What are the performance trade-offs in configuring CPL?
(2) Can members reliably use Curie to integrate various policies?
(3) Do members improve the accuracy of dose predictions with the use of CPL?

The answers to the first two questions are addressed in Section 6.1, and the last question is answered in Section 6.2. As detailed throughout, Curie allows 50 members to compute the privacy-preserving model using 5K data samples with 40 inputs in less than a minute. We also show how an algorithm with flexible data exchange policies can improve–often substantially–the accuracy of the warfarin dose model accuracy.

**Experimental Setup.** The experiments were performed on a cluster of machines with 32 GB of maximum memory and 16-core Intel Xeon CPU at 1.90 GHz, where we use one core to get a lower bound estimate. Each member is simulated in a server that stores its data. Secure computation protocols of Curie are implemented using the open-source HElib library [4]. We set the security parameter of HElib as 128 bits. Multiplication level is optimized per member to increase the number of allowed homomorphic operations without decryption failure and to reduce the computation time.

We validate the accuracy of dose model in various consortia defined in Table 3 with members defining different data exchange policies. The dataset used in our experiments contains 5700 patient records from 21 members. Dose model accuracy of each member

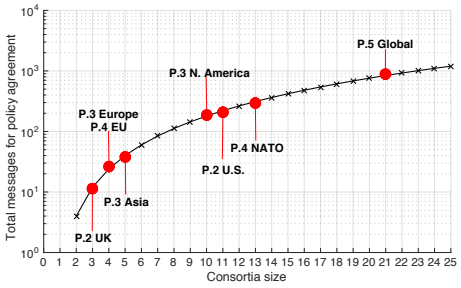


**Figure 5: CPL negotiation cost - Costs associated with a number of varying members in a consortium. Each member defines asymmetric share and acquisition policy for other members. The number of members in warfarin consortia is marked with red circles.**

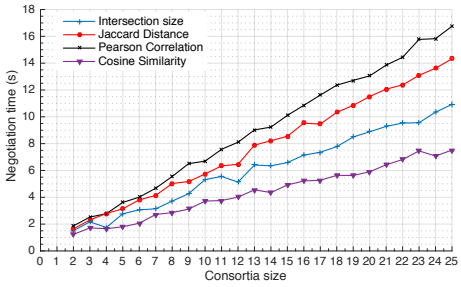


**Figure 6: CPL selections and data-dependent conditional costs - Costs associated with varying members and algorithms. All consortia members agree on policy including a different data-dependent conditional and selections over one input of having 200 samples.**

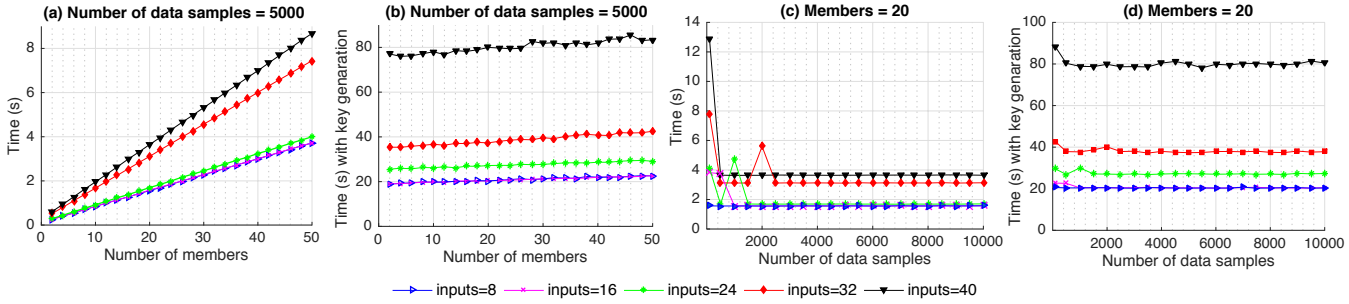


**Figure 7: CPL performance on privacy-preserving and differential private protocol - All members define an asymmetric share and acquisition policy through selections and conditionals. The agreements of CPL policies between consortia members are studied with the different number of consortia members, data samples, and input size. (Std. dev. of ten runs is ± 3.6 and ± 0.3 sec. with and without homomorphic key generation.)**

is validated with Mean Absolute Percentage Error (MAPE). MAPE measures the percentage of how far predicted dosages are away from true dosage. Lower values indicate better quality of treatment.

## 6.1 Performance Evaluation

We present the costs associated with various Curie mechanisms. We illustrate the cost of the CPL in policy negotiations, in the use of data-dependent conditionals, and in the dose algorithm.

*6.1.1 CPL Benchmarks.* Our first set of experiments characterize the policy construction and negotiation costs. Various consortia and policies are instrumented to analyze the overhead of the number of messages and time required to compute the CPL selections and data-dependent conditionals. All the costs not specific to the policies are excluded in measurements (e.g., network latency). The benchmark results are summarized in Figure 5 and 6 and discussed below.

Figure 5 shows the number of messages for policy construction required for different consortia size. The number of members in warfarin study is also labeled. For instance, NATO consortium has 13 members; ten members from U.S. and three from UK. The experiments illustrate the upper bound results wherein each member defines a different share and acquisition policy for other members (i.e., asymmetric relations). In this, each member sends acquisition policy request to consortium members. After a member gets the acquisition request, it reconciles with its share policy and output of negotiation message is returned. The number of messages associated with varying number of selections and conditionals dictated by the members does not require any additional messages. For instance, the acquisition request of a member includes arguments when conditionals are defined (e.g., reference data and a threshold value for data-dependent conditionals such as pairwise Jaccard distance), and the result is returned with the negotiation output message. However, the use of the selections and data-dependent conditionals brings additional processing cost as detailed next.

Figure 6 shows the costs associated with the use of CPL selection and data-dependent conditionals. All the members dictate data-dependent conditionals and selections on a single input. The members input size for the data-dependent conditional computations is set to 200 real values. This is the average number of inputs found in members' dataset. Since selections and conditionals reconcile contradictions between acquisition and share policies, they do not require any additional computation overhead and yield a processing time of milliseconds. However, the time associated with varying data-dependent conditionals depend on the protocol of associated secure pairwise algorithm. In our experiments, cosine similarity and intersection size exhibited shorter computation time than Pearson correlation and Jaccard distance. Overall, we found that 25 members compute the metrics less than 18 seconds. Note that the results serve as an upper bound that all members define a set of selections and a data-dependent conditional on one input.

*6.1.2 Dose Model Benchmarks.* Our second series of experiments characterize the impact of CPL on the average time of computing privacy-preserving dose model with varying number of members and dataset sizes. Though the warfarin study includes eight inputs,

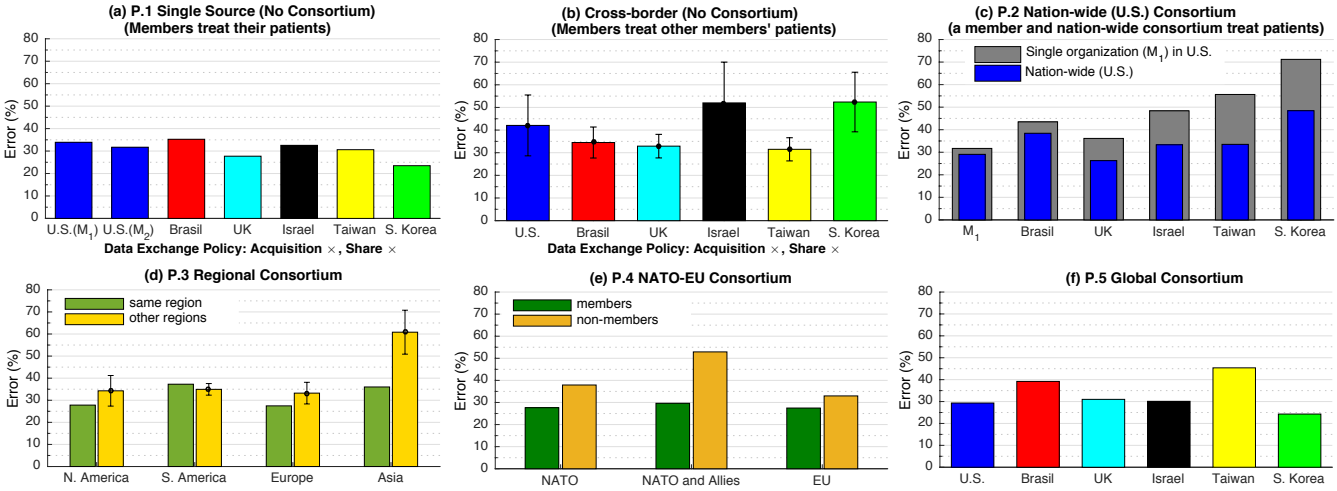


**Figure 8: The implication of policies on model accuracy - errors are validated in various consortia through data exchange policies. Figure 6(c-f): The local acquisition policies of members comply with the sharing policy within a consortium (i.e., members acquire complete data of the consortia members. Std. devs. of errors are within %5, if not illustrated).**

evaluations are repeated with the input size of 8, 16, 24, 32, and 40 through various dataset sample sizes for completeness. The input and sample size together represents the total dataset shared for a member as a result of the policy agreements. Our experiments show that 80% of computation overhead is attributed to HE key generation. The cost of the differential privacy takes microseconds, as the members can calculate the (optional) differential private algorithm model at the end of the secure dose protocol. Computations are instrumented to classify the overheads incurred by key generation, encryption, decryption, and evaluation. We next present the costs with and without key generation to study the impact of the number of members and data size.

Figure 7 (a-b) presents the computation cost with varying number of members. Each member's dataset includes 5000 data samples which acquired as a result of the policy negotiations. Figure 7 (a) presents the cost of the total computation time excluding HE key generation. There is a linear increase in time with the growing number of members. This is the fundamental cost of encryption and evaluation operations dominated by matrix encryption and addition. To profile the generation of key cost, in Figure 7 (b), we conducted similar experiments. Each input size cost increases because of the key generation overhead. The increase is quadratic as a number of slots (plaintext elements) are set to square of input size not to lose any data during input conversion. It is important to note that the cost is independent of the member size because a member generates the key only once in a computation of a consortium. We note that the time overhead of key generation is not a limiting factor as members may generate keys before a consortium is established.

In Figure 7 (c-d), we show the costs associated with different data samples. The number of members in a consortium is set to 20. Similar to the previous experiments, the key generation dominates the computation costs. Our experiments also reported no relationship between the cost and number of samples. That is, even though the size of the data samples increases, the overhead is amortized over the operations on the local statistics of the computations (which is the square matrix of the input size in the warfarin dataset); thus the time of computing dose algorithm converges to the number of dataset inputs. This explains the similar trends observed in plots.

## 6.2 Effectiveness of Policies

We validate the performance of privacy-preserving dose model quantitatively and qualitatively. For the warfarin study, these are translated to the following questions: How do policies impact the accuracy of members' warfarin dose prediction? (Section 6.2.1), and Does policies help to prevent the adverse impacts of dose errors on patient health? (Section 6.2.2).

*6.2.1 Implications of CPL on Model Accuracy.* In our first set of experiments, we validate how well a member prescribe warfarin dose for its local patients and patient's of the consortium members without using CPL. These results are used as a baseline for comparison of varying consortia and data exchange policies throughout. Figure 8 (a) sought to identify the local algorithm errors (**P.1**). The errors significantly differ between countries and for the members of the same country (depicted as $M_1$ and $M_2$ in the U.S.). The low results are due to having homogeneous data; all the inputs in these countries have similar traits. For instance, similar age and ethnicity found in a dataset produce over-fitted computation results for its local patients. These findings are validated with use of local algorithms for treatment of other countries' patients. As illustrated in Figure 8 (b), the dose errors yield significantly high for particular countries' patients. The results indicate that improvements in dose predictions of local patients and members' patients lay in the creation of data exchange policies to increase the patient diversity.

The next experiments measure the impact of CPL in nation-wide (**P.2**), regional (**P.3**), NATO-EU (**P.4**) and global (**P.5**) consortia. Each member creates a local acquisition policy to acquire the complete data of consortia members (i.e., the acquisition policy of a consortium member complies with the share policy of the requested member). We make three major observations. First, varying partnerships yield different dose accuracy. For instance, members of nation-wide consortium get better dose accuracy than their local results. This result is validated through nationwide consortia

| Member | Agreement of policy negotiations |
|---|---|
| U.S. | (Race="Asian")∨(EVALUATE(age))∨(height <160) ∨(weight <65)∨(CYP2C9 IN ( 2*/*2, 2*/*3)∨(Amiodarone="Y")∨(Enzyme="Y") |
| Brasil | (Race="Asian")∨(height <165)∨(CYP2C9 IN (2*/*2, 2*/*3)∨EVALUATE (Amiodarone)∨(Enzyme="Y") |
| UK | (Race≠"White")∨(age BETWEEN 20-29 AND >80)∨(height<165)∨(60<weight <100)∨EVALUATE(CYP2C9)∨(Amiodarone="Y"), (Enzyme="Y") |
| Israel | (Race≠ "White")∨(height <160cm)∨(weight <60)∨(CYP2C9=3*/*3)∨(Amiodarone="Y")∨(Enzyme Inducer ="Y") |
| Taiwan | (Race=All)∨(age BETWEEN 20-29)∨(height >170)∨(weight >65)∨(CYP2C9 IN (1*/*2, 2*/*2, 2*/*3, 3*/*3)∨(VK0RC1="G/G")∨(Amiodarone="Y")∨(Enzyme="Y") |
| S. Korea | (Race=All)∨ (age BETWEEN 20-29)∨(height >165)∨(weight >60)∨(CYP2C9 IN (1*/*2, 2*/*2, 2*/*3, 3*/*3)∨(VK0RC1="G/G")∨(Amiodarone="Y")∨(Enzyme="Y") |

**Table 4: An exploration of CPL policies in the global consortium (illustrated as a plain language): Each member defines asymmetric local policy based on its data diversity. The agreement of share and acquisition policies are depicted as a policy clause in a single row. The agreement result of each member for other members is not presented for brevity.**

and a single member ($M_1$) in United States (see Figure 8 (c)). Second, supporting previous findings, all regional (excluding Asia) and NATO-EU policies decrease the error for both treatment of their patients and the other countries' patients (see Figure 8 (d-e)). However, Asia consortium results in unexpected dose errors for the treatment of other regions' patients. This is because nation-wide, regional, and NATO-EU policies include patient population having different characteristics; thus the data obtained through policy negotiations better generalize to the dosages. In contrast, Asia collaboration lacks large enough White and Black groups. Third, the global consortium results in higher dose errors when evaluated for particular countries such as Brazil and Taiwan (see Figure 8 (f)). To conclude, while CPL is effective in reducing dose error of a member, the results highlight the need for the systematic use of CPL through selections and conditionals to obtain better results.

In these experiments, each member dictates a different acquisition policy based on its racial groups. Members aim at having an ideal patient population uniformity. To do so, each member defines a local acquisition policy and negotiates it with other members. Each member sets its share policy to conditionals of being in the same consortium and data size greater than 200; thus, the policy of each member is asymmetric. Table 4 shows the simplified notation of the policy agreements in the global consortium. For instance, a member having a small number of white patients defines selections to solely acquire that group and a member having large enough patients for all genotypes sets data-dependent conditionals to obtain patient inputs that are not similar in its data samples (e.g., acquires different genotypes). Figure 9 presents a subset of results on dose errors per patient race. The errors of the other races yield similar for each member. The results without CPL conditionals and selections are plotted as a dashed line for comparison. We find that members can improve the dose accuracy with the use of policies. We note that the use of different data-dependent conditionals defined in evaluate does not result in statistically significant accuracy gain.

*6.2.2 Implications of CPL on Patient Health.* We examine the impact of the dose errors found in the previous section to better quantify the effectiveness of policies on patient health.

To identify the adverse effects of warfarin, we use a clinical study to evaluate the clinical relevance of prediction errors [9] and a medical guide to identify the consequences of over- and under-prescriptions [16]. We define errors that are inside and outside of the warfarin safety window, and the under- or over prescriptions. We consider weekly errors for each patient because using weekly values eliminates the errors posed by the initial (daily) dose. The weekly dose is in the safety window if an estimated dose falls within 20% of its corresponding clinically-deduced value [26, 27].

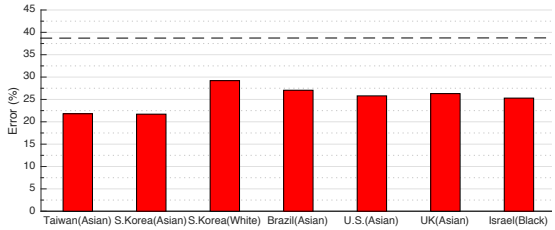


**Figure 9: Dose accuracy of members using CPL policies defined in Table 4. Members construct a model per race after they reconcile the policies. The dashed line is the average error found without the use of conditionals and selections in policies.**

| Consortium | U | SW | O | Selections | Conditionals |
|---|---|---|---|---|---|
| Single Source | 37.7% | 43.4% | 18.8% | ✗ | ✗ |
| Nation-wide | 18.9% | 52.3% | 28.8% | ✓ | ✓ |
| NATO | 19.3% | 51.5% | 29.2% | ✓ | ✓ |
| Regional | 19% | 51.3% | 29.7% | ✓ | ✓ |
| Global | 21.2% | 46.8% | 32% | ✓ | ✓ |

**Table 5: Impact of policies on health-related risks: Results are from a global consortium patients using policy agreement of a member located in the U.S. The member uses the policy defined in Table 4. (U: Under-prescription, SW: Safety Window, O: Over-prescription)**

The deviations falling outside of the safety window is an under- or over prescriptions, and cause health-related risks.

Table 5 presents the percentage of patients falls in safety window, over- and -under prescriptions with varying policies of a member. We find that use of CPL increases the number of patients in the safety window. For instance, a member has 43.4% patient with using its local data (single source model), and the member increases the percentage of patients in a safety window with varying consortia and policies, for instance, it is 52.4% in the nation-wide consortium. We conclude that CPL might be useful in preventing errors that introduce health-related risks.

## 7 LIMITATIONS AND DISCUSSION

One requirement for correctly interpreting the CPL policies is a shared schema for solving the compatibility issues among members. For instance, members may interpret the data columns (e.g., column names and types) differently or may not have the information about consortium members (e.g., membership status of an alliance). CPL implements a shared schema describing column names, their types, and explanations of data fields as well as consortium-specific

information. Members can negotiate the schema similar to the policy negotiations and revise the schema based on the schema of a negotiation initiator.

CPL provides a set of data-dependent statistical functions (e.g., cosine similarity) to compute pairwise statistics among member's local data. However, there might be a need for other functions that help members decide their data exchange policies. For example, data exchange among finance companies may require calculating the similarity between data distributions. Future work will investigate the integration of different data-dependent statistics into CPL.

Lastly, we did not focus much on the reasons of policy impacts on the prediction success of the dose algorithm and its adverse outcomes on patient health over time. While our evaluation results showed that members could express both complex relations and constraints on the data exchange through CPL policies, members require establishing true partnerships to improve the prediction model accuracy. While this explanation matches both our intuition and the experimental results, a further domain-specific formal analysis is needed. We plan to pursue this in future work.

## 8 RELATED WORK

Policy has been used in several contexts as a vehicle for representing configuration of secure groups [30], network management [35], threat mitigation [18], access control [13], and data retrieval systems [15]. These approaches define a schema for their target problem and do not consider the challenges in secure data exchange. In contrast, Curie defines a formal policy language to dictate the data exchange requirements of members and enforces the agreement in collaborative ML settings.

On the other hand, secure computation on sensitive proprietary data has recently attracted attention. Federated learning [20, 37], anonymization [14], multi-site statistical models [10], secure multiparty computation [28], and secure and differentially-private multiparty computation [1] have started to shed light on this issue. Such techniques have been used both for training and classification phases in deep learning [36], clustering [22], and decision trees [8]. To allow programmers to develop such applications, secure computation programming frameworks and languages are designed for general purposes [7, 14, 24, 32, 33]. However, these approaches do not consider complex relationships among members and assume members share their all data or nothing. We view our efforts in this paper to be complementary to much of these works. CPL can be integrated into these frameworks to establish partnerships and manage data exchange policies before a computation starts.

## 9 CONCLUSIONS

We presented Curie which provides a novel policy language called CPL to define the specifications of data exchange requirements securely for use in collaborative learning settings. Members can assert who and what to exchange separately for data sharing and data acquisition policies. This allows members to efficiently dictate their policies in complex and asymmetric relationships through selections, conditionals, and pairwise data-dependent statistics. We validated Curie in an example real-world healthcare application through varying policies of consortia members. A secure multiparty and (optional) differentially-private model is implemented to illustrate the policy/performance trade-offs. Curie allowed 50 different members to efficiently compute a privacy-preserving model using 5K data samples with 40 inputs in less than a minute. We also showed how an algorithm with effective use of data exchange policies could improve the accuracy of the dose prediction model.

Future work will investigate the use of Curie in other collaborative learning settings exploring different statistics for data-dependent conditionals and explore its performance trade-offs by integrating it into other off-the-shelf secure computation frameworks.

## ACKNOWLEDGMENT

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). This work is also partially supported by US National Science Foundation (NSF) under the grant numbers NSF-CNS-1718116 and NSF-CAREER-CNS-1453647. The statements made herein are solely the responsibility of the authors. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## REFERENCES

[1] Abbas Acar et al. 2017. Achieving Secure and Differentially Private Computations in Multiparty Settings. In *IEEE Privacy-Aware Computing (PAC)*.
[2] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. 2017. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *CoRR* abs/1704.03578 (2017). arXiv:1704.03578 http://arxiv.org/abs/1704.03578
[3] American Recovery and Reinvestment Act of 2009. 2017. https://en.wikipedia.org/wiki/American_Recovery_and_Reinvestment_Act_of_2009. [Online; accessed 01-June-2018].
[4] An Implementation of Homomorphic Encryption. 2017. https://github.com/shaih/HElib. [Online; accessed 01-January-2017].
[5] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235* (2018).
[6] Carlo Blundo et al. 2013. EsPRESSo: Efficient Privacy-preserving Evaluation of Sample Set Similarity. In *Data Privacy Management Security*.
[7] Dan Bogdanov et al. 2016. Rmind: a Tool for Cryptographically Secure Statistical Analysis. *IEEE Transactions on Dependable and Secure Computing* (2016).
[8] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. 2015. Machine Learning Classification over Encrypted Data. In *NDSS*.
[9] Z. Berkay Celik, David Lopez-Paz, and Patrick McDaniel. 2016. Patient-Driven Privacy Control through Generalized Distillation. *IEEE Symposium on Privacy-Aware Computing* (2016).
[10] Fida K Dankar. 2015. Privacy Preserving Linear Regression on Distributed Databases. *Transactions on Data Privacy* (2015).
[11] Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. 2012. Fast and Private Computation of Cardinality of Set Intersection and Union. In *Cryptology and Network Security*.
[12] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. 2012. Large scale distributed deep networks. In *NIPS*.
[13] Li Duan, Yang Zhang, Chen, et al. 2016. Automated Policy Combination for Secure Data Sharing in Cross-Organizational Collaborations. *IEEE Access* (2016).
[14] Khaled El Emam et al. 2013. A Secure Dist. Logistic Regression Protocol for the Detection of Rare Adverse Drug Events. *American Medical Informatics* (2013).
[15] Eslam Elnikety et al. 2016. Thoth: Comprehensive Policy Compliance in Data Retrieval Systems. In *USENIX Security*.
[16] U.S. Food and Drug Administration. 2017. Medication guide, Caumadin (warfarin sodium). http://www.fda.gov. [Online; accessed 01-June-2018].
[17] NATO Standard Allied Joint Doctrine for Medical Support. 2017. http://www.nato.int. [Online; accessed 01-June-2018].
[18] Julien Freudiger, Emiliano De Cristofaro, and Alejandro E Brito. 2015. Controlled Data Sharing for Collaborative Predictive Blacklisting. In *DIMVA*.
[19] Roberto Garrido-Pelaz et al. 2016. Shall We Collaborate?: A model to Analyse the Benefits of Information Sharing. In *ACM Workshop on Information Sharing and Collaborative Security*.
[20] Robin C Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially Private Federated Learning: A Client Level Perspective. *arXiv preprint arXiv:1712.07557* (2017).

[21] Oded Goldreich. 2009. *Foundations of Cryptography: Basic Applications.* Cambridge university press.
[22] Thore Graepel, Kristin Lauter, and Michael Naehrig. 2012. ML Confidential: Machine Learning on Encrypted Data. In *Information Security and Cryptology.*
[23] Health Information Technology for Economic and Clinical Health Act. 2017. https://en.wikipedia.org. [Online; accessed 01-June-2018].
[24] Wilko Henecka et al. 2010. TASTY: Tool for Automating Secure Two-party Computations. In *ACM CCS.*
[25] Yan Huang et al. 2011. Faster Secure Two-Party Computation Using Garbled Circuits. In *USENIX Security Symposium.*
[26] International Warfarin Pharmacogenetics Consortium. 2009. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *The New England Journal of Medicine* (2009).
[27] Stephen E Kimmel et al. 2013. A pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. *New England Journal of Medicine* (2013).
[28] Yehuda Lindell and Benny Pinkas. 2009. Secure Multiparty Computation for Privacypreserving Data Mining. *Journal of Privacy and Confidentiality* (2009).
[29] Chang Liu et al. 2015. Oblivm: A programming Framework for Secure Computation. In *Security and Privacy.*
[30] Patrick McDaniel and Atul Prakash. 2006. Methods and Limitations of Security Policy Reconciliation. *ACM TISSEC* (2006).
[31] Payman Mohassel and Yupeng Zhang. 2017. SecureML: A system for scalable privacy-preserving machine learning. In *Security and Privacy (SP).*
[32] Olga Ohrimenko et al. 2016. Oblivious Multi-Party Machine Learning on Trusted Processors. In *USENIX Security Symposium.*
[33] Aseem Rastogi et al. 2014. Wysteria: A programming language for generic, mixed-mode multiparty computations. In *IEEE Security and Privacy (SP).*
[34] European Commission Report. 2017. Overview of the National Laws on Electronic Health Records in the EU Member States. http://ec.europa.eu. [Online; accessed 01-June-2018].
[35] Ana C Riekstin et al. 2016. Orchestration of Energy efficiency Capabilities in Networks. *Journal of Network and Computer Applications* (2016).
[36] Reza Shokri et al. 2015. Privacy-preserving Deep Learning. In *ACM CCS.*
[37] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. 2017. Federated Multi-Task Learning. In *NIPS.*
[38] Daniel J Solove and Paul M Schwartz. 2015. *Information Privacy Law.* Aspen.
[39] Marten Van Dijk and Ari Juels. 2010. On the Impossibility of Cryptography Alone for Privacy-preserving Cloud Computing. *HotSec* (2010).
[40] Marten Van Dijk and Ari Juels. 2010. On the Impossibility of Cryptography Alone for Privacy-preserving Cloud Computing. In *USENIX Hot Topics in Security.*
[41] Fang-Jing Wu, Yu-Fen Kao, et al. 2011. From Wireless Sensor Networks Towards Cyber Physical Systems. *Pervasive and Mobile Computing* (2011).
[42] Xi Wu et al. 2015. Revisiting Differentially Private Regression: Lessons from Learning Theory and their Consequences. *arXiv:1512.06388* (2015).
[43] Jun Zhang et al. 2012. Functional Mechanism: Regression Analysis under Differential Privacy. *VLDB* (2012).


# A CURIE POLICY LANGUAGE

This section presents the Backus Naur Form of Curie data exchange policy language.

⟨curie_policy⟩ ::= ⟨statements⟩

⟨statements⟩ ::= ⟨statement⟩ ';' [⟨statements⟩]

⟨statement⟩ ::= ⟨share_clause⟩
| ⟨acquire_clause⟩
| ⟨attribute⟩
| ⟨sub_clause⟩

; share clauses defined as follows:

⟨share_clause⟩ ::= 'share' ':' [⟨members⟩] ':' [⟨conditionals⟩] '::' ⟨selections⟩

; acquisition clauses defined as follows:

⟨acquire_clause⟩ ::= 'acquire' ':' [⟨members⟩] ':' [⟨conditionals⟩] '::' ⟨selections⟩

; attributes are defined as follows:

⟨attribute⟩ ::= ⟨identifier⟩ ':=' '<' ⟨value⟩ '>'
| ⟨identifier⟩ ':=' '<' ⟨value_list⟩ '>'

; user defined sub-clauses defined as follows:

⟨sub_clause⟩ ::= ⟨tag⟩ ':' [⟨conditionals⟩] '::' ⟨selections⟩

; conditionals including data-dependent functions defined as follows:

⟨conditionals⟩ ::= ⟨var⟩'='⟨value⟩ [',' ⟨conditionals⟩]
| 'evaluate' '(' ⟨data_ref⟩ ',' ⟨alg_arg⟩ ',' ⟨threshold_arg⟩ ')' [',' ⟨conditionals⟩] | ''

⟨selections⟩ ::= ⟨filters⟩
| ⟨tag⟩

⟨filters⟩ ::= ⟨filter⟩ [',' ⟨filters⟩]

⟨filter⟩ ::= ⟨var⟩ ⟨operation⟩ ⟨value⟩ | ''

⟨data_ref⟩ ::= '&' ⟨identifier⟩

⟨alg_arg⟩ ::= ⟨algorithms⟩

⟨algorithms⟩ ::= 'Intersection size'
| 'Jaccard index'
| 'Pearson correlation'
| 'Cosine similarity'

⟨threshold_arg⟩ ::= ⟨floating_point_number⟩

⟨operation⟩ ::= '=' | '<' | '>' | '!=' | in |

⟨value_list⟩ ::= '{' ⟨value⟩ '}' [',' ⟨value_list⟩]

⟨members⟩ ::= ⟨member⟩ [',' ⟨members⟩]

⟨member⟩ ::= ⟨identifier⟩ | ''

; for completeness, trivial items defined as follows:

⟨identifier⟩ ::= ⟨word⟩

⟨var⟩ ::= '$' ⟨identifier⟩

⟨value⟩ ::= ⟨string⟩

⟨tag⟩ ::= ⟨word⟩

⟨string⟩ ::= '"' ⟨stringchars⟩ '"'

⟨stringchars⟩ ::= ⟨stringletter⟩ [ ⟨stringchars⟩]

⟨stringletter⟩ ::= 0x10 | 0x13|0x20| ... | 0x7F

⟨word⟩ ::= ⟨char⟩ [ ⟨word⟩ ]

⟨char⟩ ::= ⟨letter⟩ | ⟨digit⟩

⟨letter⟩ ::= 'A' | 'B' | ... | 'Z' | 'a' | 'b' | ... | 'z' | 0x80 | 0x81 | ... | 0xFF

⟨digit⟩ ::= '0' | '1' | ... | '9'

⟨floating_point_number⟩ ::= ⟨decimal_number⟩ '.' [⟨decimal_number⟩]

⟨decimal_number⟩ ::= ⟨digit⟩ [ ⟨decimal_number⟩ ]

# B DIFFERENTIALLY-PRIVATE DOSE ALGORITHM

We presented how members compute a privacy-preserving dose model on negotiated data through their policies. In this section, we consider individual privacy that allows a member to guarantee no information leakage on the targeted individual (i.e., patient) involved in the computation. Specifically, while members compute a secure dose model using the data obtained as a result of their policy negotiations, they also ensure that an adversary cannot infer whether any particular individual is included in computations to build the dose algorithm. In warfarin study, this corresponds to a differentially-private secure dose algorithm on shared data.

To implement a differentially-private secure algorithm, we use a functional mechanism technique [42, 43]. The technique accepts a dataset ($\mathcal{D}$), an objective function ($f_{\mathcal{D}}(\eta)$), and a privacy budget ($\epsilon$) as an input and returns $\epsilon$-differentially-private coefficients $\widetilde{\eta}$ of an algorithm. The intuition behind the functional mechanism is perturbing the objective function of the optimization problem. Perturbation includes both sensitivity analysis and Laplacian noise insertion as opposed to perturbing the results via differentially-private synthetic data generation.

To inter-operate the functional mechanism with the secure dose protocol, members convert each column from [min, max] to [-1,1] before negotiation starts. This processing ensures that sufficient noise is added to the objective function on negotiated data. Then, members proceed with the protocol. At the final stage of the secure algorithm protocol, a member gets clear statistics of $O = X^{\intercal}X$ and $\mathcal{V} = X^{\intercal}Y$ and input dimension $d$ that is the size of $O$ or $\mathcal{V}$. These statistics are the exact quantities that are minimized in the objective of the functional mechanism [43]. Using these statistics, a member

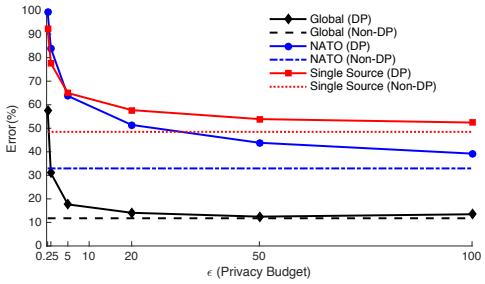


**Figure 1: Non-private secure algorithm (Non-DP) vs. differentially-private secure algorithm (DP) performance of a member in U.S. measured against various policies depicted in Figure 8.**

may (optionally) compute $\epsilon$-differential private secure algorithm without any additional data exchange and computational overhead.

**Differential Privacy Results.** To protect individual privacy in secure dose algorithm, members may compute the differentially-private secure algorithm on their negotiated data. This section presents the results of using the differential-private secure algorithm (DP) instead of using secure dose algorithm (Non-DP). To establish a baseline performance, we constructed non-private secure algorithms of a member. We then build the differential-private secure algorithm for different privacy budgets ($\epsilon$ = 0.25, 1, 5, 20, 50 and 100). Finally, we compare the results of two algorithms through different policies of a member. Figure 1 shows the results of a member in the U.S. that applies both algorithms to predict the dosage. The algorithms are constructed for the single source, NATO, and global consortia. In this, the member dictates acquisition policy for complete data and other members complies with their share policy. The average error over 100 distinct model for each budget value is reported. The use of DP degrades the accuracy as the $\epsilon$ value increases. For instance, the accuracy improvement obtained through NATO policy over single source degrades with the privacy budget less than or equal to 20. We note that other consortia and policies with use of selections and conditionals show similar effect on the dose accuracy.

## C ANALYSIS OF THE DOSE ALGORITHM

We present security and privacy guarantees of the dose algorithm provided to all members through the share of encrypted integrated statistics, ($\mathcal{O}_i = \mathcal{X}^\intercal\mathcal{X}$ and $\mathcal{V}_i = \mathcal{X}^\intercal\mathcal{Y}$ matrices). Since all data exchange among parties is encrypted through the use of HE, the security of the algorithm against any adversary outside the authorized parties is based on the underlying HE cryptosystem.

**An adversary not involving session initiator.** Assume for now that a session initiator does not collude with other parties. Loosely speaking, since all computations are performed on the encrypted data, none of the parties learn anything about other parties' input.

We consider a party $P_{i+1}$ in Figure 4. The party $P_{i+1}$ has the public key generated by the session initiator $K_i$, the encryption of local statistics of previous parties $M_i = (E(\mathcal{O}_i)_K, E(\mathcal{V}_i)_K)$. Its input is $(\mathcal{V}_{i+1}, \mathcal{O}_{i+1})$ and its output is $M_{i+1} = (E(\mathcal{O}_i+\mathcal{O}_{i+1}), E(\mathcal{V}_i+\mathcal{V}_{i+1}))$. A simulator $S$ selects random values for its own inputs $(\mathcal{V}'_{i+1}, \mathcal{O}'_{i+1})$ and encrypts them using the public key published by the session initiator. Then, the simulator $S$ performs the homomorphic operation on the received message $M_i$ and outputs $M'_{i+1} = (E(\mathcal{O}_i + \mathcal{O}'_{i+1})_K, E(\mathcal{V}_i + \mathcal{V}'_{i+1})_K)$. Here, we assume the underlying HE is semantically secure. Therefore, the output of the simulator $M'_{i+1}$ is computationally indistinguishable from output of the real execution of the protocol $M_{i+1}$ for every input pairs. Therefore, using the definition in [21] the protocol privately computes the function in the presence of one semi-honest corrupted party. The extension to multi-corrupted semi-honest adversaries is straightforward as the only difference is the view of a subset of parties having many encrypted messages. Since the semantic security of the underlying HE is hold for any pair of these many encrypted messages, no information leaks about the corresponding plaintexts.

**Adversary involving session initiator.** We consider the case when the session initiator is corrupted. The corrupted parties including session initiator can infer the input of an honest party if the predecessor (previous party) and successor (next party) of an honest party are both corrupted. We consider the possible cases for data leakage: (1) *2-party:* The session initiator is corrupted, and another party is honest. In this case, predecessor and successor of the honest party are both the corrupted session initiator. Therefore, the input of honest party is learned by the corrupted party, (2) *3-party*: A corrupted session initiator is either predecessor or successor; thus it can learn inputs of the one of the honest party only if another party is corrupted, and (3) *n-party ($n > 3$)*: To learn an honest party's input, at least two parties must be corrupted and placed in previous and next of the honest party.

While the individual raw data of members does not leak, the risk of inappropriate disclosures from local summary statistics exists in some extreme cases [14]. Consider the exchange of plain matrix $V_i = X^\intercal Y$ among two parties; a party may use the extreme values found in $V_i$ to identify particular patients. For instance, in dose algorithm, taking inducers such as Rifadin and Dilantin could indicate high dose prescriptions. If the values of $V_i$ are high, then a party may infer a patient that takes enzyme inducers and the presence of high dosage warfarin intake. Similarly, exchange of $O_i = X^\intercal X$ may leak information about the number of observations and represent the number of 0s or 1s in a column. For instance, for the former first entry in the matrix, $X^\intercal X$, gives the total number of patients. For the latter, $(X^\intercal X)_{j,j}$ gives the number of 1s in the column. This type information lets a party infer knowledge, particularly when binary inputs (e.g., use of the medicine) are used.

## D CURIE DEPLOYMENT DETAILS

We use a dataset collected by the International Warfarin Pharmacogenetics Consortium (IWPC), to date the most comprehensive database containing patient data collected from 24 medical institutions from 9 countries [26]. The dataset does not include the name of the medical institutions, yet there is a separate ethnicity dataset provided for identifying the genomic impacts of the algorithm. We use the race (reported by patients) and race categories (defined by the Office of Management and Budget) to predict the country of a patient[3]. For instance, we consider a medical institution with a high number of Japanese race is located in Japan. We use subsets of patient records that have no missing inputs for accurate evaluation. We split the dataset into two cohorts: training cohort is used to learn the algorithm, and validation cohort is used to assign dose to the new patients based on the consortia and data exchange policies.

[3]The authors indicated via personal communication that they cannot provide the exact name of the institutions due to the privacy concerns.