National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Value of a Virtual Pooled Registry process to improve data quality in Central Cancer Registries

NAACCR
June 15, 2015
Lynne Penberthy MD, MPH
Patrick Mergler PhD

Challenge

- Mobility of the US population varies from 5-11% of residents moving to another state each year
  - Variation by age, demographics and economy
- This mobility may result in
  - Lack of complete information on each cancer case
  - Duplication of case reporting
  - Inability to accurately assess multiple primary cancer incidence

Challenge

- Contiguous states routinely perform data exchange for cases who are residents of the exchanging states
- But…
  - This does not provide de-duplication either among contiguous nor among states that do not routinely perform exchange
  - Nor does this capture cases or information where the patient provides as their residence in two different states over time

Importance

- We do not have accurate data assessing the magnitude of duplicate cases (especially for non-contiguous states)
- Nor do we have accurate estimates of multiple primary (MPC) incidence.
  - This is especially important in the era of precision medicine and genomic classification of tumors

Proposed solution

- Provide a secure mechanism to allow State Registries to identify patients with potential multi-reporting to
  - enhance QC efforts and
  - to de-duplicate and assess MPC incidence across all State Registries

- Utilize a De-Identified, hashing process to detect potential multi-state reported primary cancers

Proposed solution: steps & process

- Have participating State Registries setup and utilize a locally installed “hashing” application
  - Establishes a federated, hashed form of master person indexing across a network ***without*** sharing PHI outside of a registry firewall

- A hash is a string or number generated from a string of text.
  - The same input will always produce the same output.
  - Thus, if two registries hash the same patient identifier, they will generate matching hashed values (but that contains none of the patient ID).
  - That way two registries could know if they have the same patient, and know which one it is by looking only at their own data.
  - It is not possible to go from the output back to the input.

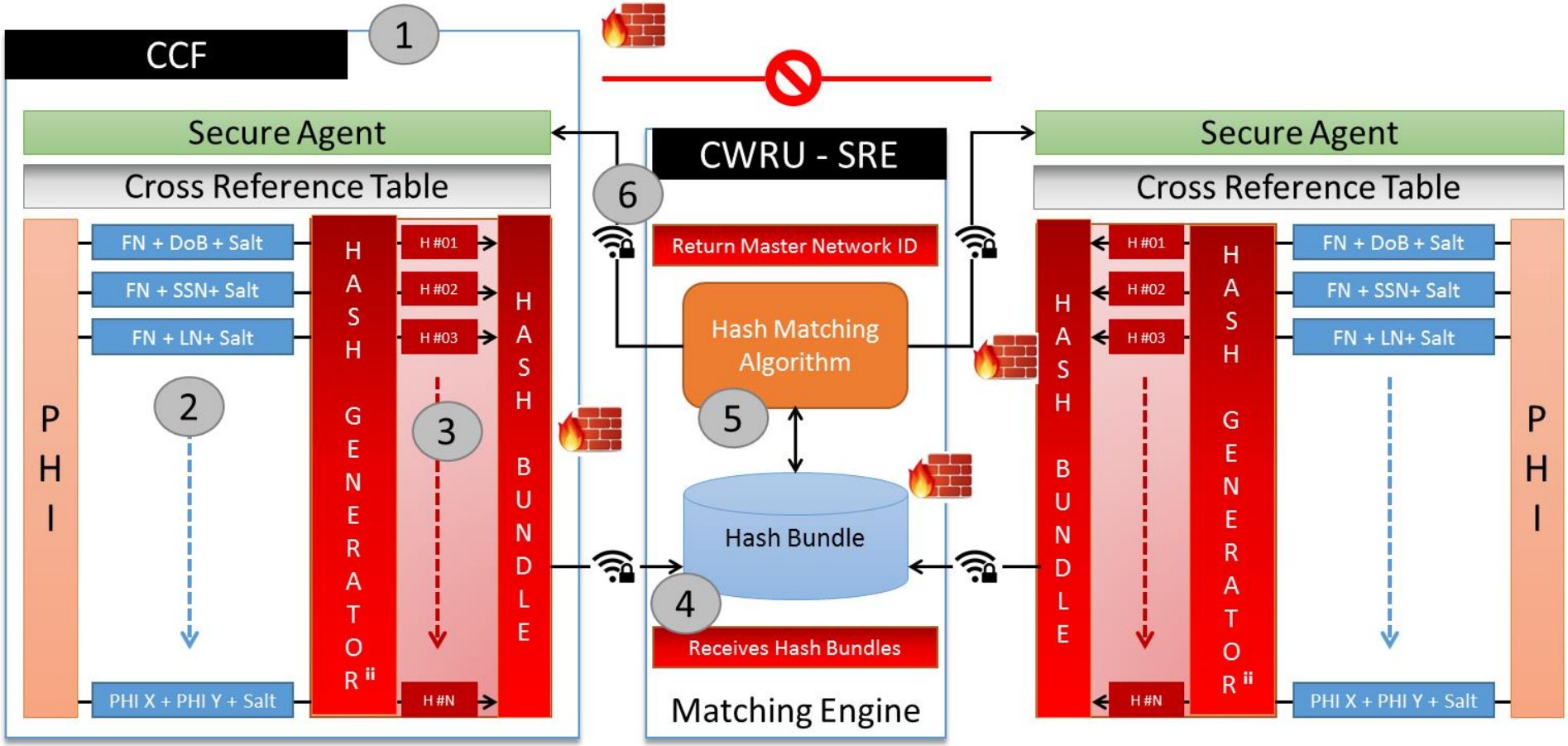
Proposed solution: steps & process

- Provide their hashed, de-Identified data to a central matching service provider
  - Deterministic matching algorithm across all participating registries based on various combinations of hashed values
- Registry receives results of their cases
  - Indicates “Potential Duplicates and Multi-Reported Primary” cases with indication of which registries also reported the same
  - Maintains linkages for participating registries ***behind the registry firewall*** to perform person re-identification

Currently in use

- Supports:
  - Capricorn research network in Chicago (PCORI CDRN: 11 clinical institutions, 2 VAs)
  - CLEARPATH research network in Cleveland *[implementation phase] (*3 hospitals: Cleveland Clinic, University Hospitals, MetroHealth Hospital)

Linkage Process in Detail



[i] Random string of data used as an additional input to a one-way hash function. Defends against dictionary and "rainbow table" attacks.

[ii] Utilizes NSA SHA-2 512 hashing algorithm

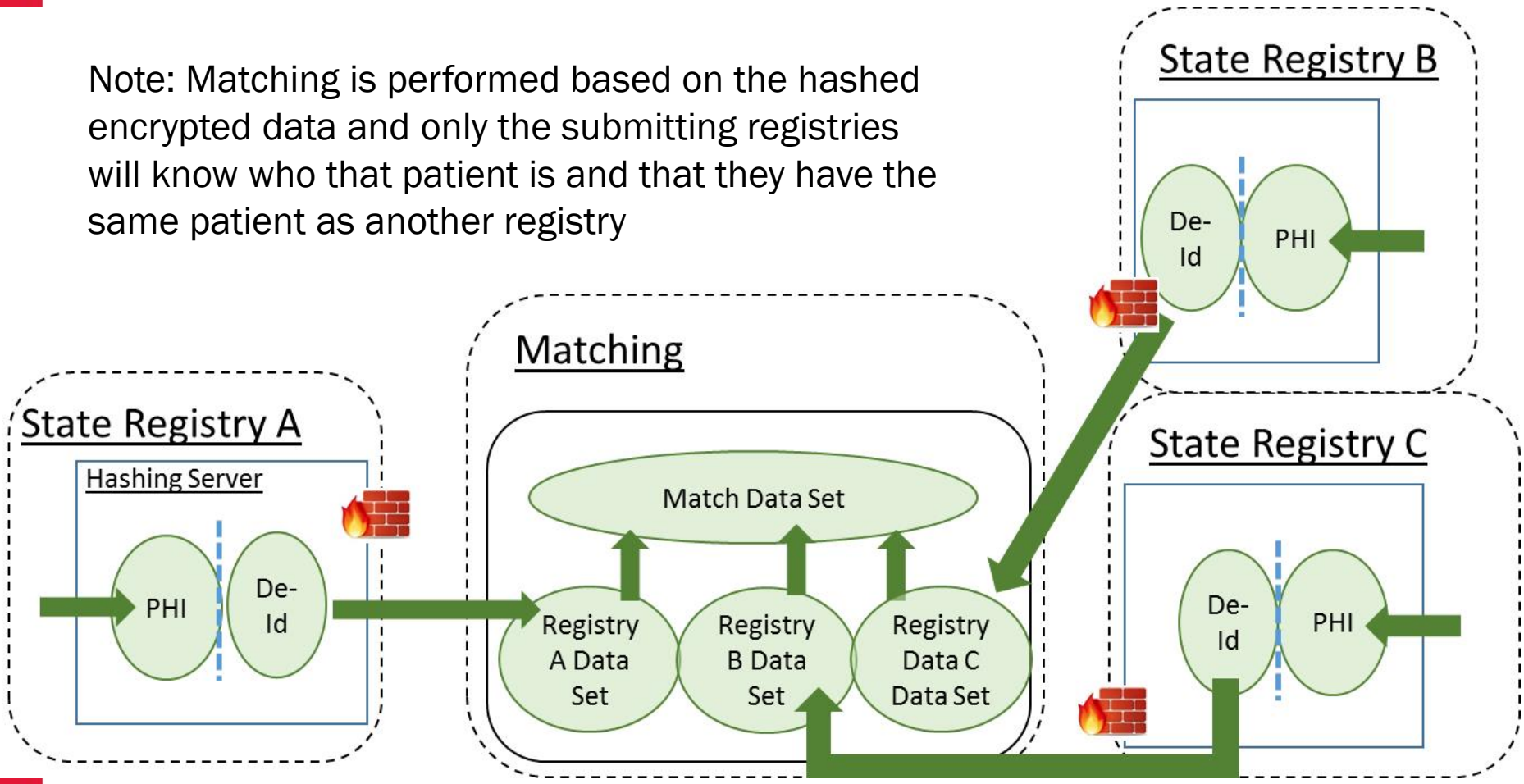
9

Implementation Details

- Registry Environment Setup:
  - IMS to provide hashing client for registry install
    - provided as a Virtual Machine or on a registry a physical server
  - Registry installs within their environment:
    - Opens Firewall to enable data to be sent to SALT Provider and Matching Provider
- Data Mapping:
  - Registry data is “prepped” with appropriate fields (SSN, DOB, Name, etc.)
  - Registry provides “prepped” or formatted file copied to hashing server
- Matching Provider:
  - Encrypted hashed data provided to high throughput computing facility for matching
  - Matches from all sites and then returns back a “master ID” with encrypted data and information on where their may be a duplicate case

Conceptual Model

Note: Matching is performed based on the hashed encrypted data and only the submitting registries will know who that patient is and that they have the same patient as another registry



Implementation Details

- Through this process
  - patient identifying information will be maintained behind the registry firewall
    - the PII and the encryption key will be maintained by the registry behind their firewall
  - The encrypted de-identified data will use a sophisticated deterministic matching algorithm to (initially):
    - identify the same patients who are reported in more than one state
    - Provide the information on the “potential duplicate cases” to the relevant states

Next Steps

Coincident with this effort a NAACCR workgroup will:

- Develop new processes for registries to perform exchange and sharing
- Develop rules to determine how multiple primary cancers would be captured and reported
- Simplifying this process NAACCR is working to have registries sign a universal agreement to simplify data exchange across all registries

Next Steps:

- A validation study of the encrypted match process
  - to determine the accuracy of encrypted matches
  - By a subset of states using unencrypted PII

- If the initial pilot is successful subsequent additional development will be done to enable matching of patients based on both encrypted PII in combination with the cancer site and date of diagnosis

- This process could be repeated annually to assure continued de-duplication and accurate assessment and identification of MPCs

In conclusion

This method based on secure encryption and high throughput computing has the potential to

- provide more accurate estimates of multiple primary cancer incidence and
- to assure that our incidence estimates are not impacted by duplicate case reporting that is currently not accounted for

THANK YOU

16