Pioneering the Integration of Highly-sensitive Demographics Information into an Administrative Claims Database: A Step Forward in the Journey to Supporting All Populations

Rongrong Wang[1], Tu My To[1], Gurleen S. Jhuti[1], Santa Borel[2], Gina Mak[3], Meysam Safari[2]

[1]Genentech, South San Francisco, CA, USA; [2]Privacy Analytics, Ottawa, Ontario, Canada; [3]IQVIA, New York, USA

CONCLUSIONS / IMPACT

The integration of enhanced individual-level demographic data attributes with administrative claims data, electronic health records (EHR), or other real-world health data can fill gaps in existing datasets.

Enriching real-world data with these variables provides a comprehensive understanding of a patient's health journey and enables analyses that were not previously possible, helping to uncover obstacles to healthcare access and improve patient health outcomes for all.

Our innovative methodology sets a precedent for future initiatives aiming to integrate real-world data with additional highly-sensitive demographics information, thereby advancing the landscape of health impact research and its applications in improving healthcare delivery and outcomes.

Call to Action

Leverage this innovative approach to gain deeper insights into health impact research and generate better data to drive improved healthcare policies and practices for all population.

For additional information or to obtain a PDF of this poster

Copies of this poster obtained through Quick Response (QR) Code are for personal use only and may not be reproduced without permission from the lead author of this poster.

ISPOR 2025 Annual Meeting
May 13-16, 2025

BACKGROUND

- Differences in healthcare access and health outcomes exist within all populations. Various demographic and social factors can also have considerable health impact and further exacerbate poorer health outcomes.
- Health outcomes research aims to identify factors that can improve healthcare access and outcomes for all populations by acknowledging and addressing these barriers.
- High-quality data is essential for measuring differences in health outcomes across all populations and identifying opportunities to improve healthcare access. However, the data to fully explore these differences, such as individual-level demographic information and social factors, are not consistently collected within the healthcare system. Furthermore, the need to protect sensitive individual data while utilizing it for research purposes presents a considerable challenge.

CHALLENGES	SOLUTIONS
• Real world data (RWD) assets (both US and ex-US) typically contain only general patient demographic characteristics and are limited in information on individual-level race/ethnicity, income, and education.	• Enrich RWD by linking patient-level data from different data sources to provide a fuller picture of the patient.
• Combining patient-level information (e.g. race/ethnicity) with patient health data (e.g. medical insurance claims data) increases the risk of non-compliance with data privacy obligations.	• Consolidating data users, data providers, and privacy experts to determine optimal strategies for maximizing data usability and insights generation for a better understanding of health outcomes, while ensuring full compliance with data privacy legislation.

OBJECTIVE

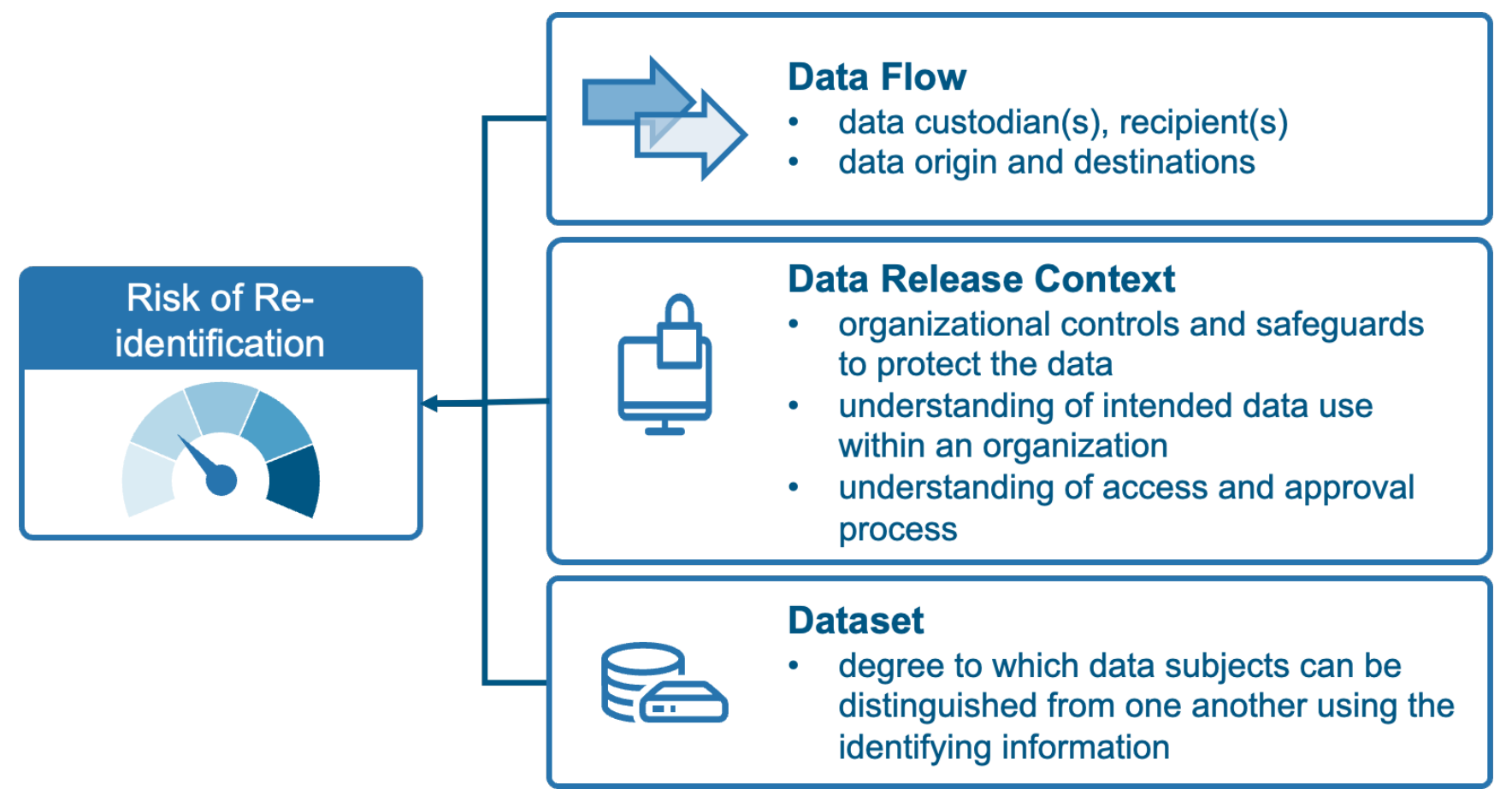
- The goal of this project is to enhance a large administrative healthcare claims database with additional sensitive individual-level demographic variables, including race/ethnicity, education level, income level. These variables are not typically included as part of standard data offerings and are sourced from a leading consumer database.
- An innovative methodology is applied to uphold patient privacy standards while preserving full usability of all other important variables in the administrative claims database.

METHOD

In accordance with the HIPAA Privacy Rule in the US[1], an Expert Determination method for de-identification of Protected Health Information (PHI) was applied to assess the potential for individuals to be reidentified in claims data linked to sensitive individual-level demographic characteristics data (e.g. race/ethnicity, education, and income) obtained from a consumer database.

- This method entails the application of statistical or scientific principles for rendering protected health information not individually identifiable.
- This is to ensure that there is a very small risk that the information could be used alone, or in combination with other reasonably available information, to identify an individual.

Data flow (describing the origination, usage, and safeguarding of the data), data release context (appropriate control on data access and use), and information contained in the data (classified by identifiability of the data fields) were examined to determine the overall re-identification risk and ensure it was below an acceptable threshold.



Data Flow
- data custodian(s), recipient(s)
- data origin and destinations

Data Release Context
- organizational controls and safeguards to protect the data
- understanding of intended data use within an organization
- understanding of access and approval process

Dataset
- degree to which data subjects can be distinguished from one another using the identifying information

RESULTS

Safe setting: strong organizational control (e.g., IT security, data access controls) was confirmed through Context Assessment Questionnaires.

- Privacy and Security questionnaire reviews and assesses the safeguards and controls protecting the data.
- Recipient Trust questionnaire reviews and assesses the motives, capacity, and contractual obligations in effect.

Safe people: strong compliance expectations were verified by reviewing the organizational privacy and security protocols, data access granting policy, and data control mechanisms.

Overall contextual protections on the data are very high:

- Strong controls in place, considering the high privacy and security and high recipient trust scores;
- Strong policies protecting the data;
- Demonstrating the benefits to patient populations that will be realized through the research performed on the integrated data.

Due to demonstrable beneficial outcomes to patient population and society, the invasion of privacy was considered to be low. Therefore, the acceptable reidentification risk threshold for this dataset was adjusted from 0.04% (minimum bar) to 0.08% (beneficial and expected use case).

With proper data access and use control (see figure below), linking claims data to additional sensitive patient demographic variables was below this threshold and resulted in a very small risk of re-identification for patients.

Safe Projects
- ✓ Requests are reviewed to ensure requirements around compliance, contractual agreements, privacy and data use purpose are met

Safe People
- ✓ Data users are trained
- ✓ Users subject to contractual controls, including no attempts to re-identify people in the data

Safe Data
- ✓ RRDs are performed to ensure that the re-identification risk is sufficiently low
- ✓ Transformations are made to the data where required to reduce re-identification risk

Safe Settings
- ✓ Access is limited to approved users
- ✓ Approval is tracked and recorded
- ✓ Systems are very secure
- ✓ High privacy and security controls and measures are in place

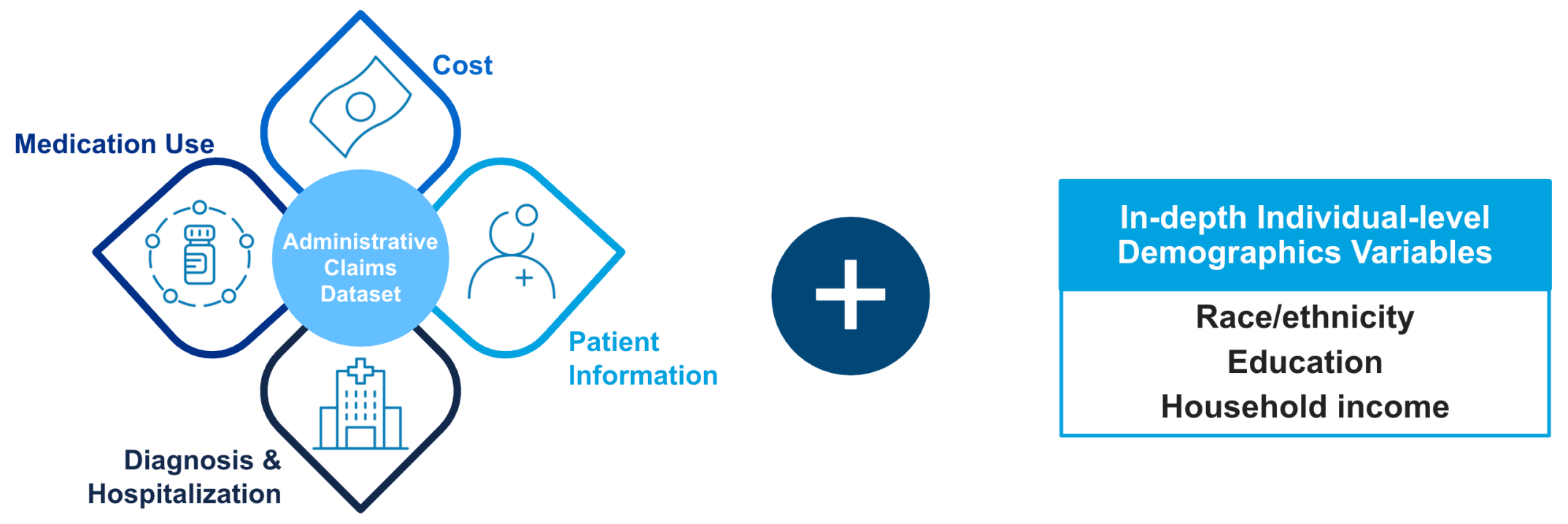
Safe Outputs
- ✓ Monitoring of data use
- ✓ Limits on data being extracted
- ✓ Tracking of data to ensure access is time banded

Real-world datasets can be enhanced with additional sensitive patient demographics like race/ethnicity, household income, education level, mortality, and more—compliantly and without compromising the usability of other variables.

USE CASES OF THE INTEGRATED DATA

Evidence suggests that non-medical variables considerably affect health outcomes and managing disease effectively requires addressing non-medical factors alongside medical treatment.

The integrated data of administrative health claims and enhanced demographic characteristics can facilitate a deeper understanding of the impact of non-clinical factors on disease progression and inform actions and interventions to enhance patients' health.



In-depth Individual-level Demographics Variables
Race/ethnicity
Education
Household income

Understand the heterogeneity in disease treatments and outcomes

- The impact of individual-level demographic characteristics (i.e. social factors) on access to high-quality healthcare, including disease-modifying therapies (DMTs).
- The impact of individual-level demographic characteristics (i.e. social factors) on future severity and patient-reported symptoms in both relapse-onset and progressive-onset disease.
- Describe the bidirectional relationship where a disease can affect these demographic characteristics and vice versa.

Inform actions and interventions to enhance patients' health:

- Risk Stratification: Assess particular demographic characteristics to identify individuals at risk for severe symptoms and consider earlier, more aggressive treatment for high-risk populations.
- Personalized Care: Deliver tailored care by adopting a more holistic view of the patient, including their financial limitations, when discussing treatment options.
- Promoting Comprehensiveness in Research: Ensure research studies reflect all local demographic characteristics to generate more generalizable findings.

Reference

1. Office for Civil Rights, "Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," Department of Health and Human Services, Washington, DC, 2012. Accessed: Jul. 07, 2021. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Disclosure

Rongrong Wang, Tu My To, Gurleen S. Jhuti: employees of Genentech, Inc. and stockholders in Roche.
Santa Borel, Meysam Safari: employees of Privacy Analytics Inc., an IQVIA company.
Gina Mak: employee of IQVIA, Inc.