HEALTH EVOLUTION FORUM

TRUST FRAMEWORK

Accelerating Responsible Use of De-Identified Data in Algorithm Development

Draft Version 1, April 2022

The **Trust Framework for Accelerating Responsible Use of De-identified Data in Algorithm Development** (hereafter referred to as the Trust Framework) was developed by Fellows of the Health Evolution Forum through the Work Group on Governance and Use of Patient Data in Health IT Products, supported by Health Evolution Forum staff, Caleb Flint and Ye Hoffman, and with research support from Walter Sujansky, MD.

The **Health Evolution Forum** is a collaboration of CEOs and other senior executives of payer, provider, and life science organizations and other industry thought leaders intended to bring about voluntary, industry-led improvement in the health care industry.

The goal of the **Work Group on Governance and Use of Patient Data in Health IT Products** (hereafter referred to as the Work Group) is to articulate standards for how data can be used in developing data tools for the clinical setting between payer, provider, and life science companies, in collaboration with developers of algorithms and other analytics solutions.

The views in this document represent the collective views of the Fellows and do not represent the view of Health Evolution or any specific Fellow or organization, including organizations and individuals providing supporting subject matter expertise, guidance, and research within the Forum.

CONTRIBUTORS TO THIS WORK

Work Group Co-Chairs

- Aneesh Chopra, President, CareJourney
- Cris Ross, Chief Information Officer, Mayo Clinic
- Steve Klasko, former President & CEO, Jefferson Health

Work Group Participants

- Amy Abernethy, President of Clinical Research Business, Verily Life Sciences
- Mike Berger, VP & Chief Data & Analytics Officer, Mount Sinai Health System
- Michael Blum, Chief Digital Transformation Officer, UCSF Health
- Ellen Duffield, SVP & COO, Gateway Health Plan
- Todd Gottula, Co-founder, President & Chief Product Officer, Clarify Health Solutions
- Gary Guthart, President & CEO, Intuitive Surgical
- Eric Hargan, former Acting Deputy Secretary, US Department of Health and Human Services
- David Horrocks, President & CEO, Chesapeake Regional Information System for Patients
- Frank Ingari, President & CEO, Tandigm Health
- Azam Khan, Chief Data Insights Officer, Eli Lilly and Company
- Alice Leiter, VP & Senior Counsel, eHealth Initiative
- Carolyn Magill, CEO, Aetion
- Arien Malec, SVP R&D Clinical and Administrative Networks, Change Healthcare
- Ken Mandl, Director, Computational Health Informatics Program, Boston Children's Hospital
- Deven McGraw, Co-founder & Chief Regulatory Officer, Ciitizen
- Rob Metcalf, CEO, Concert Genetics
- Peter Neupert, Lead Director & Board Member, Adaptive Biotechnologies
- Brenda Pawlak, Managing Director, Manatt
- Eric Schneider, former SVP, The Commonwealth Fund

TABLE OF CONTENTS

SEEKING FEEDBACK

The Approach

This effort seeks to determine how to balance privacy and security safeguards with innovative use of de-identified data in algorithm development. This is a challenging balance, which requires industry stakeholders to engage in good faith in establishing trust between institutions that produce data and those that use data.

The Work Group acknowledges significant work by subject matter experts in promoting responsible use of de-identified data in current initiatives. However, controversies and conflicting ideas have limited agreement on and adoption of industry standards, which slowed the development of algorithms being used in clinical settings to improve health outcomes.

Accelerating responsible use of de-identified data requires a set of standard approaches that will gain widespread acceptance. To that end, the Trust Framework is a necessary first step to collaborate with industry stakeholders in establishing fair and achievable guidelines.

Next Steps

The draft Trust Framework will be circulated among health care and technology leaders to ensure that it represents a comprehensive understanding of the challenges and that potential standards being explored are well informed.

The Work Group will proceed in an iterative fashion to engage industry stakeholders in developing and refining guidelines within each of the Trust Framework's six principles. The next step of this process involves **requesting input on what key questions should be answered and whom to engage in answering those questions**.

Request for Information

The Work Group is seeking initial input through **July 31, 2022**. Please email Ye Hoffman, Director, Forum (YeH@healthevolution.com) with your feedback on:

- ☐ Whether the Trust Framework principles are sufficient to address the overarching goal
- ☐ What are key questions that the Trust Framework must answer within the six principles
- ☐ Which industry organizations and subject matter experts to prioritize for soliciting input

INTRODUCTION

The Challenge

Digitization of health care continues to yield ever-growing data sources that offer, together with advancements in analytics and machine learning, significant opportunities for breakthroughs in care delivery and bio-medical discovery. Since no single entity holds a complete view of individual or population-level health, innovation necessitates cooperation to build large-scale longitudinal data sets.

Industry leaders are eager to participate in data collaborations. They also believe it is essential to protect data against intentional or unintentional risks that could compromise their patients or organizations. While are detailed regulatory and industry guidelines for handling individually identifiable data, de-identified data is typically not subject to privacy laws.

There is no standard approach for using de-identified data in algorithm development. Organizations struggle to evaluate the relative risks and benefits of participating in collaborative efforts using de-identified data. Leaders must navigate complex data de-identification practices, privacy and security considerations, governance and oversight mechanisms, and many other challenges. Even the most well-resourced and capable organizations face significant uncertainties in establishing de-identified data sharing and use agreements.

The Solution

Leaders participating in the Health Evolution Forum recognized the need to develop guidelines that set a baseline for accelerating responsible use of de-identified data in algorithm development. However, industry stakeholders – those that produce data, and those that use data – lack a framework to determine how to establish fair and balanced data sharing and use agreements that foster trust in collaboration.

The Trust Framework for Accelerating Responsible Use of De-identified Data in Algorithm Development defines the fundamental principles that organizations must address in tandem to promote mutual interests among stakeholders. By providing a nuanced understanding and action-oriented approach, the Trust Framework fosters cross-industry collaboration necessary for tackling a complex and critical challenge. This effort engages subject matter experts to identify the most pressing questions that must be addressed, and subsequently find practical answers to move the industry forward.

VALUE OF THIS WORK

For patients

With a better understanding of how de-identified data is governed and used, patients can have greater confidence in data safeguards and responsible use in algorithms developed to improve health outcomes.

For clinicians

With industry momentum in making more de-identified data available for research and algorithm development, clinicians will gain faster access to improvements in care delivery while having greater confidence in their validity and relevance to their patients.

For producers of data

With clear and achievable guidelines, organizations that provide data (such as health systems and health plans) will be empowered to participate in collaborations leveraging de-identified data for algorithm development. Leaders can also proactively address potential backlash due to misunderstanding by policymakers or the media.

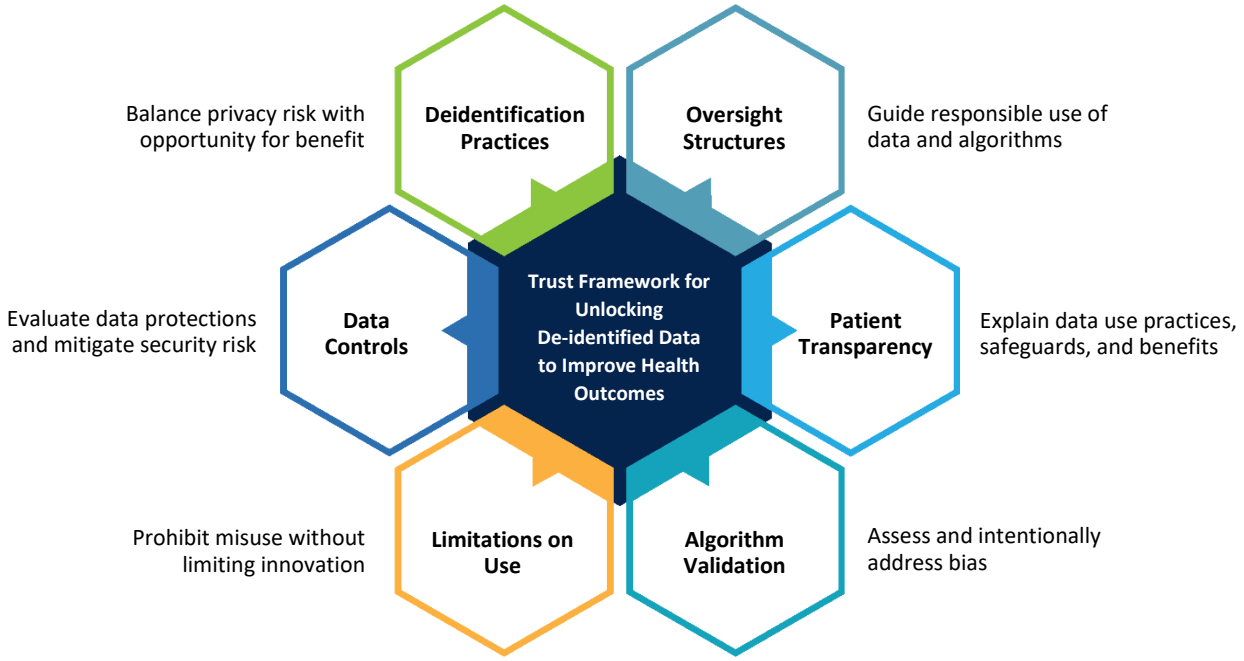
For users of data

With a common language about goals and friction points in launching data-sharing collaborations, data users (such as research coalitions, technology developers, and analytic services) can better support prospective collaborators in winning buy-in from key decision-makers across their organization.

For policymakers

While this effort is not aimed at policymaking or changes to regulations, policymakers can engage with multi-stakeholder consensus building efforts to establish more trusted data access and use, which reduces harms without limiting research and innovation.

TRUST FRAMEWORK



This effort is designed to build consensus for how industry stakeholders establish de-identified data sharing and use agreements within the existing regulatory landscape.[1] It is intended to set a minimum set of requirements for industry data sharing and is not designed to replace the use of advanced privacy-preserving technologies or other sophisticated approaches that go further than these principles but are not widely available to all.

The initial draft of the Trust Framework was developed during the 2021-2022 Forum Fellowship year by the Work Group on Governance and Use of Patient Data in Health IT Products. This draft describes each of the Trust Framework's six principles, including the goal, context, and challenges.

In the next phase of this effort, the Work Group is seeking input from industry stakeholders on all components of the Trust Framework, with particular attention to pinpointing the key questions that should be answered to promote widespread adoption. **The Work Group has identified an initial set of critical questions and welcomes feedback on any additional questions for consideration.**

[1] Kenneth D. Mandl, M.D., M.P.H., and Eric D. Perakslis, Ph.D. "HIPAA and the Leak of 'Deidentified' EHR Data," https://www.nejm.org/doi/full/10.1056/NEJMp2102616

I. De-identification Practices

GOAL

Balance privacy risk with benefit. When de-identifying data and contracting with authorized data recipients, organizations must minimize the risk of data being re-identified while preserving its analytical and investigative value.

CONTEXT

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule stipulates two methods for data de-identification[2]:

- **Safe Harbor**. This method requires removing or modifying 18 specific data elements that identify the individual to whom the data corresponds, including full dates and full ZIP codes.
- **Expert Determination**. This method applies "generally accepted statistical or scientific principles" to analyze and, if necessary, modify the data to ensure that the risk of re-identification is "very small." The HIPAA Privacy Rule does not specify exact methods for expert determination, nor quantifies the notion of "very small" risk.

CHALLENGES

No single de-identification practice is best across all potential use cases. There is no way to de-identify data such that the risk of re-identification is zero. The re-identification risk depends on the data set itself, along with what other data can be cross-referenced, with the general trend of more data becoming available over time.

Identifying the most protective practice that supports the intended use case. The "safe harbor" method is more straightforward and easily verified, but in practice may be unsuitable for analyses that require retention of complete dates or timestamps. Thus the "expert determination" method may be required for the most useful analyses. Generally, performing the "expert determination" relies on expertise in at least the following areas:

- Statistical Disclosure Limitation/Control Theory & Practices
- Privacy-Preserving Data Publishing and Mining
- Data Privacy Computer Science (e.g., Differential Privacy, Homomorphic Encryption)
- Biostatistics/Epidemiology
- HIPAA/HITECH and Data Privacy Law
- Medical Informatics and Medical Coding/Billing Systems
- Geographic Information Systems
- Machine Learning/Artificial Intelligence
- Cryptography

KEY QUESTIONS

- How should organizations assess the risk of re-identification or categorize levels of risk?
- What methods of expert determination should be used to address different levels of risk?
- Where can organizations access the appropriate expertise needed to perform a risk assessment and apply expert determination practices?

[2] HHS Guidance Regarding Methods for De-identification of Protected Health Information, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

II. Data Controls

GOAL

Evaluate data protections and mitigate security risk. Data producers must hold data users responsible and accountable for implementing appropriate data security safeguards.

CONTEXT

Data that has been de-deidentified in accordance with the HIPAA Privacy Rule is no longer considered protected health information (PHI), nor subject to HIPAA Rules. There is no single source of straightforward and comprehensive guidelines for securing de-identified data. Organizations seeking to tailor security practices to de-identified data must draw upon multiple existing standards, such as the HITRUST Common Security Framework (CSF)[3] and CISA[4] best practices. Meanwhile, privacy-enhancing technologies are quickly improving. Industry leaders are looking to establish "clean rooms" that allow data producers to retain control, enhance federated learning models, and advance "behind the glass" analysis techniques that reduce risk by enabling data to stay in place.

CHALLENGES

While less risky than PHI, de-identified data poses a non-zero risk if breached or misused. Security infrastructure, including technical and policy measures, must be mature enough to protect against re-identification by rogue actors, but simple enough to implement so as not to impose undue administrative burdens. Organizations should consider expanding their security perimeter to encompass de-identified data sets (particularly those that are linked) much the same way identifiable data are protected.

Data producers have limited resources to evaluate security safeguards across numerous data-sharing agreements. Leveraging industry-recognized certification would reduce the burden on data producers to understand and conduct regular assessments of external parties' data controls. Data recipients should substantiate their conformance to industry best practices through either a self-audit or independent audit, as deemed appropriate by the data producers. Stakeholders should proactively address the overall administrative burden, aiming to minimize unnecessary delays and costs while striving towards best practice.

Privacy risk is tied to security risk. Given the high sensitivity and value of health data, industry stakeholders must safeguard data from security breaches by unauthorized parties who seek to re-identify patients for non-health care or malicious uses.

KEY QUESTIONS

- Which industry-recognized security standard(s) should algorithm developers use?
- How should data producers enforce developer compliance with security standards?
- What procedures should be followed in the case of a security breach involving de-identified data?

[3] Specifically, the subset of the HITRUST CSF designated as applicable to the safeguarding of de-identified data in Appendix A of the HITRUST De-Identification Framework, https://hitrustalliance.net/product-tool/de-identification/

[4] Cybersecurity and Infrastructure Security Agency, For example, CISA Ransomware Guide 2020, Part 1: Ransomware Prevention Best Practices, https://www.cisa.gov/sites/default/files/publications/CISA_MS-ISAC_Ransomware%20Guide_S508C_.pdf

III. Limitations on Use

GOAL

Prohibit misuse, including re-identification, without limiting innovation. Re-identification of data that an organization has collected, de-identified, and shared with other organizations can impose regulatory and reputational costs.

CONTEXT

The HIPAA Privacy Rule does not require a Data Use Agreement (DUA) for sharing de-identified data. In contrast a DUA is required for releasing Limited Data Sets including stipulations that prohibit re-identification.[5] Regardless of the method chosen to de-identify data, data producers should establish a DUA with the data recipient; while the HIPAA Privacy Rule does not dictate how a covered entity may disclose de-identified data, establishing a DUA allows the data producer to stipulate terms designed to protect the data.

CHALLENGES

Data producers must determine the allowable use cases to include in the data sharing and use agreement. Appropriate and permissible use cases are highly variable and may evolve with emerging technology and/or new industry and institutional priorities. Data producers must ensure terms and conditions flow down to any third parties with whom the data recipient engages that can access the data.

Linking data sets warrants reassessment of potential risk for re-identification. Increasingly, the highest-value use cases require linking multiple de-identified data sets for a more complete view of health. The more data sets are linked together, the more likely that a combined data set could be used to re-identify the patient. However, linking data sets to derive more value is different than re-identification, which requires an individual or entity to act intentionally on the decision to break the veil of anonymity using that data asset.

Resharing of data has significant implications for potential re-identification. When a data recipient reshares the data with a third party, the original data producer may lose control of how data are used and secured, which has significant implications for potential re-identification. Two options for maintaining oversight of data use are: 1) strict prohibition of redistribution, or 2) redistribution only with permission. In addition to wholescale redistribution, data producers must also consider how to approach derivative works, including the complexity of determining when a data set is sufficiently modified to no longer be considered redistribution.

KEY QUESTIONS

- What contractual terms should be used to enforce limitations on use?
- Under what circumstances might data producers allow resharing data with third parties, either wholesale and/or in derivative works?

[5] Further information about data use agreement requirements for the release of Limited Data Sets can be found on the OCR website, https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

IV. Algorithm Validation

GOAL

Assess and intentionally address bias. Every step of the algorithm development lifecycle has a human touch point that can introduce bias. For example, developers may ask algorithms to consider the wrong questions or use historical data sets embedded with bias.

CONTEXT

Standards for algorithm validation are an emerging area among regulators, industry coalitions, technology companies, and standards organizations. For example, NIST is developing a voluntary Artificial Intelligence (AI) Risk Management Framework[6] to "incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems." Algorithm validation will need to be a recurring activity as algorithms will continue to learn and apply new patterns so they must be continuously tested as the algorithms iterate. Regulatory frameworks have not historically addressed iterative validation, and this will be a necessary element of future regulation.

CHALLENGES

Biases not only reduce algorithm performance but can also result in an unfair allocation of health care resources or stigmatization. Developer of algorithms should equip end-users with information on their approach to addressing biases in algorithm development. However, there is no existing industry standard for how best to contextualize this information in the clinical workflow.

Data recipients must build processes for monitoring and addressing algorithm bias in product design. As part of the vendor selection process, algorithm procurers should ask whether developers have tested their product for validity and bias. At a minimum, evaluate algorithms for non-discrimination against protected and sensitive classes such as race, ethnicity, age, gender, and socioeconomic status. The health care workforce must be educated on the importance of determining whether an algorithm has been tested for the specific population they are serving, and who may be over- or under-represented in the data sets on which the algorithm is being trained.

KEY QUESTIONS

- How can the validity of an algorithm developed with de-identified data be communicated to clinicians?
- What role should industry stakeholders (regulators, standards organizations, providers, payers, or other entities) have in assessing algorithm validity?

[6] National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework

V. Patient Transparency

GOAL

Explain data use practices, safeguards, and benefits. To build trust with patients, industry stakeholders must promote understanding about how data are used to improve care.

CONTEXT

The HIPAA Privacy Rule does not require patient consent to share de-identified data. Patients are accustomed to HIPAA Notice of Privacy Practices but may perceive the intent as a legal disclosure primarily to protect the organization from liability. Transparency around robust de-identification practices is necessary but insufficient to build patient trust. Organizations should consider alternative communication methods to address de-identified data use and safeguards against re-identification.

CHALLENGES

Provider organizations must engage patients in dialog about how data is being used to improve care delivery. The responsibility for patient transparency most appropriately rests on the provider organization (as the data source) rather than downstream recipients and users of that data**.**

Transparency can cause conflict unless there is also a foundation of patient trust and understanding. Provider organizations must take care not to cause an unnecessary burden for front-line staff and clinicians – strategic communication with patients should highlight value of data use and anticipate appropriate avenues for directing patient concerns and questions.

KEY QUESTIONS

- Should patients be given disclosures about how organizations use de-identified data in algorithm development?
- What patient-centered language should provider organizations use to educate patients about data use, and what is the appropriate channel to promote transparency?
- What public disclosures should provider organizations make when engaging with algorithm developers using de-identified data?

VI. Oversight Structures

GOAL

Guide responsible use of data and algorithms. Ensure that oversight is aligned with broader organizational and industry initiatives, including evolving standards for responsible practices in machine learning, artificial intelligence, and other secondary uses of data.

CONTEXT

Because the HIPAA Privacy Rule does not limit the disclosure of de-identified data, legal review is not necessarily an obstacle since sharing of de-identified data sets is legally permissible. Gaining buy-in from privacy officers is critical to sharing and using de-identified data sets. The main concern is the reputational impact on the organization in the event of real and/or perceived breaches of privacy and confidentiality.

CHALLENGES

Organizations looking to maximize patient value must also minimize conflicts of interest especially where monetization is involved. Leaders should establish a multidisciplinary approach to data sharing that draws upon clinical, research, and other business stakeholders including informatics, innovation, privacy, compliance, and legal functions. Organizations must win buy-in across constituents who have varying levels of risk tolerance.

Data governance boards must consider patient perspectives and research ethics. Patients continue to become more engaged with their health data and more aware of data privacy and security concerns. Provider organizations can seek input from Patient and Family Advisory Councils, especially around communication and education about data use. Decisions about de-identified data must also address equity and fairness, such as through consultation with experts in research ethics, commonly found within the purview of Institutional Review Boards (IRBs) or departments overseeing sponsored human subjects research.

KEY QUESTIONS

- Which leader(s) in the organization should be responsible for establishing oversight for de-identified data sharing arrangements?
- When should the Board be aware or involved in algorithm development using de-identified data?
- In what capacity should patients be represented in oversight of algorithm development using de-identified data?