

On practical privacy challenges of medical research in the US

Kirill Nikitin
Columbia University
kirill.nikitin@columbia.edu

Gamze Gürsoy
University of Cambridge
gg584@cam.ac.uk

Afsana Ferdous
Barnard College, Columbia University
af3441@barnard.edu

Lucy Simko
Barnard College, Columbia University
lsimko@barnard.edu

Abstract

The progression of medical research depends on access to sensitive health data, creating an inherent tension between patient privacy and public health. National and local policy in the US mandates protection of patient privacy, but largely leaves medical researchers room for interpretation on *how* to do so. While there may be privacy-preserving mechanisms that could suit medical researchers, we recognize a gap between the security and privacy community and the medical research community: we do not know how medical researchers protect the privacy of their data *in practice*, what barriers they face in doing so, and how privacy inhibits research. To understand these underlying privacy tensions, we conducted semi-structured interviews with sixteen researchers who either analyzed medical datasets or served as data custodians. We find that medical researchers interpret and implement privacy regulation with deep respect for patient privacy. However, these privacy protections impede or conflict with research needs and productivity throughout the research lifecycle. Sometimes researchers are able to resolve the tensions with bespoke or commercial tools (e.g., by using their own heuristics for data de-identification), whereas other times privacy tensions remain unresolved. We offer recommendations to tool developers, policy-makers, and privacy researchers for building privacy-enhancing tools for medical research.

Keywords

Privacy, medical research, policy, de-identification, HIPAA

1 Introduction

There is an inherent tension between patient privacy and the progression of medical research for public health. In order for medical researchers to make scientific progress, they need to collect and share large datasets containing sensitive health data, ranging from genetic samples to clinical notes and medical images.

In the US, the federal Health Insurance Portability and Accountability Act (HIPAA) applies to all medical data between a patient and their provider, and includes strict and overarching privacy and security stipulations [160]. These stipulations are largely vague—including phrases like “*any security measures [necessary to] ... implement the specifications*” [160]—and place both the burden and

the privilege of interpretation and implementation on custodians of medical data.

While such flexibility affords data custodians the ability to protect the data and the patients in ways that *they* know best, researchers previously reported that the HIPAA Privacy Rule had caused an increase in the cost of conducting research, created administrative delays, and hindered access to data [107, 109]. This results in calls for easing the privacy restrictions [34, 174].

Additionally, the existence of HIPAA does not protect against all privacy violations in practice: violations of health privacy still occur, e.g., when data are used for commercial purposes or discrimination [35, 63, 114], or through a linkage attack identifying public figures’ health diagnoses [149], or when hospital and insurance data are stolen directly from servers [139], or through leakage from machine learning models trained on medical datasets [50, 87, 93].

The computer security and privacy community recognizes this need for stronger privacy guarantees and protections for medical data: Many studies have developed privacy tools for medical scenarios [10, 75, 108, 137, 175] or suggested that their proposed technology *might* be useful in healthcare [66, 76, 78]. Other studies surveyed challenges related to specific types of medical data [15, 102] or existing privacy tools [170, 173].

In practice, however, medical researchers use few of these tools and technologies. For example, differential privacy is not used much in healthcare data [57], despite healthcare privacy being a strong motivator in some differential privacy work [37, 90]. The diversity of medical data, the complexity of the medical research process, and the ambiguity of the regulations make the straightforward application of technologies, such as homomorphic encryption (HE) or secure multi-party computation (MPC), difficult. When suitable tools do exist, their use might be impractical due to the interdependency of research workflows: privacy protections at one stage of data processing can create barriers elsewhere.

In this paper, a collaboration between researchers in computer privacy, usable security, and bioinformatics, we aimed to understand the tensions between the privacy requirements and the needs of medical research in practice and how the privacy and security community can help with addressing them. We interviewed 16 medical researchers (including 6 data custodians) who work with secondary data about privacy practices, tools, and restrictions throughout the lifecycle of their research, aiming to understand how the implementation and interpretation of privacy policy lead to downstream effects on privacy that hinder research data utility, productivity and innovation. We ask, first: **(RQ1) How does the medical research community interpret and implement privacy regulation like HIPAA *in practice*?** Through this question, we learn how data

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit <https://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–20

© YYYY Copyright held by the owner/author(s).

<https://doi.org/XXXXXXX.XXXXXXX>



custodians and medical researchers translate the flexibility and rigidity of HIPAA and other regulations to work for *their* data and *their* infrastructure. As described in Section 4.2, we find that the interpretation and implementation involve bespoke, manual labor to evaluate research plans as valid and trustworthy, sociotechnically rooted structures of trust in other researchers, and that researchers go *beyond* the minimum necessary de-identification standards in HIPAA because they believe that these standards are insufficient.

From these interpretations and implementations of privacy policy, we then ask: **(RQ2) To what extent are these privacy practices—driven by the interpretation and implementation of HIPAA—in tension with research innovation and research productivity?** And **(RQ3) How do researchers manage these tensions in practice?** Here, we find that the implementation of privacy can indeed impede research innovation and productivity by limiting data utility (Section 4.3) or impairing research workflows (Section 4.4). In response, researchers and data custodians have developed workarounds—such as pseudo-differential-privacy obfuscation to fulfill data sharing requirements, proxy methods and commercial tools for privacy-preserving data deduplication and record linkage, or virtual remote analysis machines hosted by data custodians to facilitate compliance with privacy policy. Yet, some tensions either remain unresolved or rely on labor-intensive and error-prone mitigations; e.g., the necessary release of study data and artifacts, data de-identification, and remote data analysis all require extensive human oversight to ensure privacy compliance. We explore these workarounds, tensions, and mitigations to data-utility and research-workflow challenges in Sections 4.3–4.4.

Through this variety of practices, we find immense space for collaboration with the privacy research community and others adjacent to ours. By rooting our recommendations in the day-to-day problems and practices that medical researchers meticulously explained to us, we see our work and our recommendations as a bridge between the medical research community and the privacy community. With the field of usable security behind us, advocating for research that begins with asking people *what they experience and what they need*, our paper presents the following contributions, towards provably secure, private, and usable tools that make their way to the medical research community and alleviate tension between privacy and research, strengthening both:

- The first experience-focused interview study with medical researchers—who are custodians of and practitioners with medical data—about how and why privacy enables and restricts their medical research.
- Actionable recommendations for researchers in various fields of computer science towards supporting medical research and relieving the tension between patient privacy and medical data use in practice.

2 Background

We describe the types of medical data that researchers study and the relevant privacy regulations. This background provides the context needed to understand our findings. In Section 5, we review related work that a reader might find useful for addressing our findings with their research. Throughout this paper, we mention a number

of organizations and tools; their inclusion is *not* an indication that we drew participants from them.

2.1 Medical research: data and custodians

Medical data are diverse in type and geographically distributed. Some researchers collect data directly via clinical trials or cohort studies; however, many use secondary data—data originally collected for another purpose, e.g., clinical care or billing. Population-level health studies use large-scale data to investigate links between lifestyle factors and chronic diseases; patient-outcome prediction uses clinical data to forecast hospital re-admissions or treatment responses; disease-predictor discovery involves analyzing genomic or imaging data to identify early disease signals.

Medical research data types. Medical research uses a wide variety of data types. Medical facilities collect and store longitudinal information about patients in *Electronic Health Records (EHRs)* during routine healthcare. EHRs can contain diagnoses, treatments, doctor visit details, laboratory tests, prescriptions, and clinical notes. Most information in EHRs (except free-text notes) is structured and, hence, well-suited for statistical analyses or predictive models. *Administrative data* are collected mostly for billing or management purposes. One example is *hospital discharge data*, which medical institutions are required to share with state administrations for research and healthcare improvement [22, 110]. Discharge data typically include patient demographics, diagnosis and procedure codes, and financial information—notably less detailed than EHR data. While discharge data alone are sufficient for some research, it is common to link them to another information source, e.g., to identify correlations between health variables [77]. *Claims data* are another type of administrative data, representing insurance billing claims from healthcare providers on patients’ behalf. Claims data include costs and codes for every ordered test or treatment but not the results or clinical notes. Finally, “*omics*” data is an umbrella term for molecular measurement data, such as genomic, transcriptomic, or proteomic data. With the exception of genetic testing conducted for clinical diagnostics, the majority of omics data are produced within research environments for research purposes.

Each data type carries distinct privacy characteristics, risks of information leakage, and handling requirements.

Data custodians: institutions, consortia, biobanks. Once medical data are generated—through clinical visits, insurance claims, etc.—they can be stored by the initial medical institutions or shared within national consortia of data custodians or with an administrative body. The consortia, often federally funded, govern access and sharing policies and may aggregate identified or de-identified data from multiple institutions. The PCORnet network [123], for example, combines EHRs from different healthcare institutions and makes them available for research. *Health Information Exchanges* [2] collect clinical data for sharing between healthcare entities and also make them available for research. The Healthcare Cost and Utilization Project (HCUP) [3] maintains national administrative health datasets, including data on hospital stays, discharge abstracts, and ambulatory surgeries. *Commercial insurance providers* can share or sell claims [6], whereas the *Centers for Medicare & Medicaid Services (CMS)* maintains claims data for all the government-insured patients

and provides research access for a fee. For omics data, *biobanks*, such as UK Biobank [144] and All of Us [7], are among the largest custodians of genomic data. In addition, numerous domain-specific and public repositories [73, 95, 176] provide access to research-generated omics datasets shared by the scientific community.

2.2 Privacy legislation and medical datasets

The Privacy Rule under the US Health Insurance Portability and Accountability Act (HIPAA) [160] regulates how healthcare entities must handle *Protected Health Information (PHI)*—health information that can be used to uniquely identify individuals. Although the Rule does not directly apply to research, it affects access to medical data and their secondary use [162]. It also defines three legal categories for medical research datasets based on levels of de-identification.

The Rule permits healthcare entities to publicly release *de-identified* health datasets, defined in one of two cases: if an entity removes all 18 HIPAA identifiers (the Safe Harbor method), or if a qualified statistician determines that the risk of re-identification is low. The 18 identifiers defined by HIPAA are: names, sufficiently small geographic subdivisions, dates, telephone numbers, fax numbers, emails, Social Security Numbers, medical record numbers, health plan beneficiary numbers, account numbers, license numbers, vehicle identifiers, device identifiers, URLs, IP addresses, biometric identifiers, full-face photographic images, and “*any other unique identifying number, characteristic, or code*” [160].

The Rule also permits the sharing of *limited* datasets, given the importance of dates and geographic precision in medical and public health research, e.g., for determining disease progression during hospitalization or evaluating mortality in certain areas. *Limited* datasets are de-identified to a lesser extent—dates and addresses up to a full ZIP code are allowed [160]. Unlike fully de-identified data, limited datasets are considered PHI and require a Data Use Agreement (DUA) that mandates the data recipient to protect the data and not to attempt to re-identify the individuals.

Healthcare entities are required to adhere to the Minimum Necessary standard when sharing PHI, i.e., they must share only the information necessary to accomplish the research task [160]. Some researchers, however, might have access to fully identified datasets, either via a data custodian or through their own data collection.

For research-generated omics data, consent terms determine how de-identified data may be shared. Broad consent may permit public release of de-identified raw data for unspecified future research, whereas more restrictive consent limits raw data to controlled-access repositories and allows only derived products, such as aggregate counts, to be shared publicly.

3 Methods

Due to the exploratory nature of our research questions, we conducted semi-structured interviews with medical researchers, enabling us to explore the breadth and depth of our topic and probe into new phenomena by learning from experts—a strength of the methodology [30]. Here, we describe our recruitment, the interview protocol, and our analysis. We describe ethics in the Appendix.

Recruitment. Our goal was to talk with researchers working with medical datasets (including both data consumers and data custodians), a relatively narrow set of potential participants. We leveraged

personal connections at medical schools at multiple institutions and sent cold emails to scientists at institutions that we were not connected to, based on whether their websites (or other descriptions of their work) indicated they might work with medical data. We also used the reporting by state custodians of administrative data on the projects and their principal investigators that had received that state data. Finally, we used snowball recruitment [36]: at the end of each interview, we asked each participant if they would be willing to recommend us to colleagues. Our IRB-approved recruitment messages are in Appendix C. At first, we sought medical researchers doing hands-on research with data, but realized after a few interviews that the perspective of the data custodians was an integral part of the picture. So, we contacted scientists in charge of clinical data consortia and biobanks that provide data access to other researchers (who are also considered medical researchers).

We include 16 interviews in our dataset: we conducted 17 interviews and excluded one because the participant revealed after the interview that they had lied by omission during the interview (see Ethical Considerations). Out of the 16 participants, we recruited nine from our professional network, two by snowballing, and five via cold emails. Two personal network, one snowball, and 24 cold email recruitment attempts were unsuccessful.

We recruited participants with varying research backgrounds and focus on different types of medical data in order to identify the privacy challenges inherent to the research process. The collective expertise spanned various data types, including molecular data, clinical health records, medical images, insurance claims, hospital discharge records, and population-level data. Participants’ current fields of research included cancer care, epidemiology, metabolomics, genetics, psychiatry, as well as the practice of medicine (including pediatrics, anesthesiology, and palliative care). Participants’ research methodologies ranged from qualitative interviews to quantitative surveys to building machine learning models.

Six participants—whom we refer to as *data custodians*—had regulatory roles within their institutions, or acted as a party approving requests for data access from other researchers; all participants had backgrounds as medical researchers working with data. We label the data custodians as C1–C6 and other researchers as R1–R10. Twelve participants held PhDs, two held MDs, one held both a PhD and an MPH, and one was a PhD candidate. Ten were full-time or adjunct university professors, and two were practicing physicians.

Interviews. The lead researcher conducted every interview, sometimes with one or two other researchers joining. According to the interviewee’s preference and/or physical location, interviews were either in person in the participant’s office or via teleconferencing. Each interview began with a process of informed consent. We described the study topic and (minimal) risks, provided each participant with our IRB-approved consent form, and invited participants to read the form. To ensure consent was *informed*, we asked participants to summarize the purpose of the study as well as the risks and benefits to them. We also asked for consent to audio record.

Interviews began with a broad question about their research and expanded into topics about data access, data analysis, and privacy concerns or non-concerns. Our interview protocol can be found in Appendix A; we intentionally had a broad swath of topics and questions covered by our IRB application, but inevitably only

covered a subset in each interview, guided by the participant’s experience and our aim to fill gaps from previous interviews [30]. We conducted one pilot interview to refine the interview protocol and did not include the pilot data, as the interview protocol changed, focusing more on practices rather than security protocols.

Interviews lasted 1-2 hours. After an hour, we regularly checked with participants about their time; participants who took the most time appeared eager to continue talking with privacy-focused computer scientists. We offered participants a \$50 Amazon gift card for their time; two participants refused the compensation. We stopped interviewing when we stopped seeing new information arise in the final two interviews, which defines thematic saturation [56].

Analysis. We followed a constructivist epistemology, in alignment with the exploratory nature of our research question. Our goal was not statistical generalizability, but learning *how* and *why* privacy mechanisms restrict or enable medical research. We therefore do not report counts of participants, which imply positivist quantitative conclusions from our qualitative data.

To analyze our interview data, we conducted reflexive thematic analysis, following Braun and Clarke’s six-stage process [18]. We first transcribed each interview by using the offline Whisper library [118] or the Microsoft Teams HIPAA-compliant transcriber, and hand-corrected each transcript. During this process, three researchers wrote memos [30] to reflect on our initial thoughts and to enrich our later conversations, aligned with the first step of reflexive thematic analysis: *familiarizing yourself with the data* [18].

After each of the three researchers had reviewed the transcripts, we began the iterative process of developing our thematic codebook both synchronously—through joint brainstorming discussions—and asynchronously. As the second step of thematic analysis (*generating initial codes*), we first created open codes—some drawn from our memos. We then iteratively and collaboratively created axial and hierarchical codes, with a hierarchical structure three levels deep, increasing specificity with depth. When the codebook seemed stable, each of the three researchers independently applied it to a single interview and then discussed code additions, modifications, and deletions, as well as checking our shared understanding of code applications. We repeated this process for eight interviews. Through these discussions we realized that the subcodes (the deepest level of our codebook) were extremely detailed and most useful for individual annotation and interpretation; we thus focused on building consensus for the *themes* and *codes* but not the *subcodes*. After our eighth interview, we had reached consensus on the necessary codes *before* our discussion, and so the lead author coded the remaining interviews. Because of this collaborative consensus-building process, we did not calculate interrater reliability, consistent with best practices [100]. Our codebook can be found in Appendix B. This iterative and reflective process aligns with steps 3-5 of thematic analysis, which Braun and Clarke emphasize may be an ongoing process: *searching for themes*, *reviewing themes*, and *defining and naming themes*. Finally, we consider our writing process an extension of the analytic process (as do Braun and Clarke), through which we have continued to interpret the themes in our data.

Limitations. Our work should be read with a few key limitations. Methodologically, all data are self-reported, and although we did

our best to ensure both participant confidentiality and data privacy, we know that at least one participant—the one excluded—did not fully report their privacy concerns, due to a fear of institutional retaliation. It is possible that other participants did the same, and hence, that our work *underreports* privacy and security issues. Additionally, several study participants were recruited from the professional network of one of the authors. Although our participants’ research areas are diverse, this recruitment strategy may have skewed demographic representation, e.g., more recruited participants were affiliated with universities than with state agencies. Our work is also intentionally US-centric because of its unique regulatory privacy frameworks, and so it is possible that medical researchers in other countries encounter privacy differently due to different legislative landscapes and resources available.

4 Results

Our work explores how researchers’ implementations of privacy affect research in practice, through the entire research lifecycle. We find that the flexibility in HIPAA means that researchers and data custodians develop their own privacy practices to suit their data and infrastructure best. These practices can favor patient privacy at the expense of the utility of the data, and the time and resources of the researchers. Here, we explore how these tensions are resolved with workarounds, as well as the spaces in which they are unresolved, and might benefit from research and development from computational research communities.

We begin with an overview of the medical data research pipeline. This overview emerged directly from our interviews, and serves as important context for the privacy challenges, trade-offs, and workarounds described in Sections 4.2–4.4. With this understanding of the breadth of steps of medical research, we then discuss researchers’ interpretations and implementations of privacy policy (§4.2) and how these implementations of privacy cause tension with both the utility of the data (§4.3) and productivity (§4.4). The summary of our main findings is shown in Section 4.1.

4.1 The medical data pipeline

Conducting medical research on secondary data involves several key stages, described below and summarized in Figure 1.

Access negotiation. As many medical researchers work with secondary data—originally collected for other purposes and stored by data custodians—they must first negotiate access to (a portion of) the data. Both the data custodians and the researchers’ institutions have policies governing access control. In most cases, researchers must submit applications to both their Institutional Review Board (IRB) and the data custodians (e.g., a consortium) with the purpose of the study, the justification for the requested data elements, and potentially their plans for data privacy and security. The researchers and the data custodians negotiate a Data Use Agreement (DUA) that defines the permitted usage of data, e.g., what kind of analysis or dissemination is permitted.

Data de-identification. Once access is negotiated, custodians may need to *de-identify* the data by removing all protected health information (PHI) before sending a slice to the researchers. De-identification aims to protect data subjects from re-identification

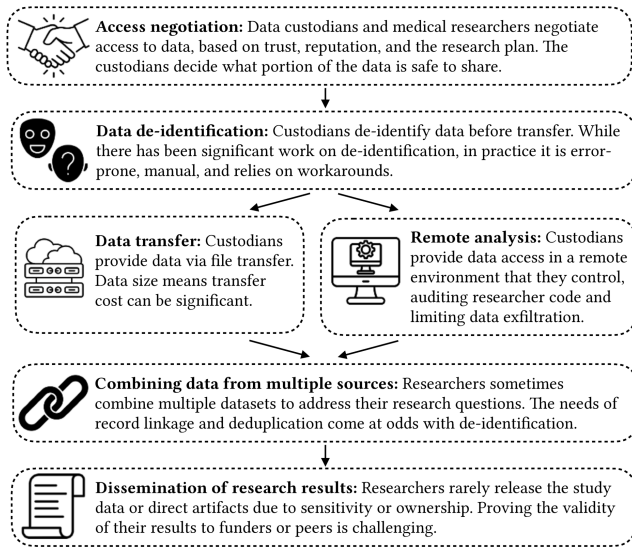


Figure 1: An overview of the medical data pipeline.

and the exposure of sensitive information about them, e.g., their health condition or their mere presence in the dataset. A hospital or a health information exchange typically stores fully identified data and extracts a de-identified or limited dataset from them upon a researcher’s request. A biobank or a consortium usually stores de-identified data, although with preserved crosswalks to the identifiers, e.g., for future study recruitment. Researchers who collect their data first-hand need to de-identify them before dissemination.

Sending and receiving data. In order to actually give a researcher technical access to the data, data custodians follow one of two paths: they either send the data directly to the researchers (e.g., via SFTP, a cloud provider, or physical mail), or they host the data in a remote virtual machine on which the researchers must conduct all their work. When receiving data directly, researchers are permitted to analyze and store the data locally for the duration of the study.

Data access via remote VMs. Some data custodians make their datasets accessible only via remote servers—virtual machines that researchers log into to perform queries and analysis on datasets, without directly downloading data. With remote data access, data analysis is different than on a local database because data custodians control personnel access, the development environment, and upload/download. Researchers are often required to purchase access to the data analysis environment on these remote virtual machines.

Combining data from multiple sources. After receiving access to de-identified secondary data, researchers might combine datasets to search for correlations between different data modalities or to increase the sample size. Studying correlations requires finding and linking the same individuals in de-identified data. When researchers combine data of the same type from multiple sites or perform distributed analysis, the data require the removal of overlapping individuals, whose presence otherwise can invalidate research findings.

Dissemination of research results. The last stage of the data pipeline is the dissemination of research findings. In addition to publishing

findings, researchers commonly share the details of their analysis, including analysis code, and summary statistics about study data. When possible, they also try to publish the study data themselves for transparency and potential secondary analysis, or due to the requirements of funders or publication venues.

4.2 Interpreting and implementing regulations

The HIPAA Privacy Rule, state laws, and institutional policies regulate the privacy requirements for the secondary use of medical data. To answer **RQ1**, we explore the interpretation and implementation of these laws governing medical privacy. We find that researchers’ and institutions’ interpretation and implementation of these policies are varied, well-intentioned, in support of patient privacy—but at times in tension with research productivity and functionality *because of their implementations*. HIPAA is *intentionally* flexible in its security requirements, stating that “covered entities and business associates may use **any security measures** that allow the covered entity or business associate to reasonably and appropriately implement the standards and implementation specifications as specified in this subpart” [160] (emphasis ours). This flexibility reflects the impossibility of setting a rigid national standard across the fragmented and extremely varied organizations that are custodians of health data. However, it also means that medical data custodians must figure out *their own interpretation of privacy and security*, subject to their own resources, which can be difficult in practice.

Researchers’ and custodians’ interpretations of privacy are grounded in a pervasive culture of respect for patient data. Participants described a culture of *respect* for participant data, deeply aware that each data point represented a human being. R1, whose PhD was in computer science, described a “culture switch” they experienced between computer science—“we should just get access to data ‘cause we’ll find cool things’”—and medicine: “data represent human lives ... We’re privileged to see it and to have a view into someone’s lives. And so we have to be careful about it” (R1). Some participants extended the principle of respect for patients as human beings to *research justice* by “engag[ing] patients and other communities in the research process, so that the research ... is truly patient centered and designed to serve the populations that it would impact” (C5). The mental model of privacy in medicine stems from the perspective that research should benefit patients, as opposed to other fields, such as computer science, where we aim to understand the world around us. We return to this patient-centered view in Section 6.2.

Researchers are acutely aware of re-identification risks. Participants were also aware of privacy risks for patients, suggesting adversaries such as an insurance company denying patient claims based on information from a sensitive dataset or a political body prosecuting a patient for having a medical procedure, or a stalker learning the dates of their target’s doctor visits. Participants highlighted patients with intersectional identities, famous people, people from small communities, people with stigmatized or rare diseases, and patients with an increased risk of a disease as possible targets of re-identification. A big driver of de-identification efforts, however, was compliance with the regulations, as “the fines for a HIPAA breach are tremendous” (R1), demonstrating the power of financial incentives in regulatory oversight.

Table 1: Summary of research challenges, current practices, barriers to privacy, and needs in medical research in the US.

	Research challenges & tensions	Current practices	Barriers to privacy	Open problems & needs
(§4.2) Interpreting and implementing regulations	Privacy regulations lack guidance on reconciling necessary access restrictions with researchers’ need for detailed data.	Researchers rely on research plans, institutional and individual histories, and the community’s ethical norms to make these determinations.	Determinations largely rely on personal judgment.	Clearer regulatory guidance, utility-measuring techniques, and sociotechnical support mechanisms.
	Researchers bear the responsibility of determining the degree of de-identification and summarization for safe data release.	Researchers adhere to existing guidelines, e.g., Safe Harbor or cell suppression policies, or voluntarily redact extra information they deem sensitive.	Implementations vary; researchers express uncertainty about privacy and anonymization in practice.	Techniques for measuring the privacy guarantees of data de-identification and anonymization.
(§4.3) Privacy vs. data utility	Researchers need to preserve data utility through data de-identification.	Researchers employ ad hoc data transformations, such as proxy variables and randomized date-shifting.	Transformations produce policy compliance but not strong privacy protection.	Data transformation techniques with provable privacy guarantees.
	Combining structured datasets requires record linkage or deduplication without violating patient privacy.	Some researchers employ commercial tools for privacy-preserving record linkage (PPRL) and deduplication.	Commercial PPRL tools offer opaque privacy guarantees.	Transparent, provably secure PPRL tools; analysis of commercial solutions.
	Privately deduplicating de-identified data is challenging due to the lack of direct identifiers.	Researchers use proxy methods, such as relatedness estimation for genomic data, for deduplication.	Methods are error-prone and data-specific; removal of a subject re-identifies them.	Techniques for privacy-preserving deduplication of unstructured data.
	Transparency requirements are at odds with data ownership and concerns about privacy leakage via study artifacts.	Researchers heuristically evaluate the re-identification risks of their datasets and artifacts prior to dissemination.	Risk evaluations are uncoded and subjective; the limited ability to share study data hinders reproducibility.	Tools to audit artifacts for privacy leakage and validate research findings without raw data access.
(§4.4) Privacy vs. research productivity	De-identification tools miss sensitive identifiers or incorrectly alter benign data.	Researchers manually audit de-identified data and/or write custom verification scripts.	Manual audits are unscalable and error-prone, and custom scripts fail to generalize.	Context-aware de-identification tools for unstructured data.
	Securing remote analysis environments against data exfiltration and unauthorized access to granular data.	Remote platforms limit and inspect data egress and ingress; analysis code is occasionally manually audited.	Manual audits are unscalable and lack insight; automated checks are overly sensitive.	Nuanced automated code-auditing mechanisms.

Data custodians closely evaluate researchers’ research plans in order to comply with HIPAA’s requirement for only releasing the minimum amount of data necessary. Access to secondary data requires an application from the researchers to the data custodian, during which the custodian evaluates the validity of the research and the minimum data necessary to answer the research question, as required by HIPAA: “...[data custodians] must make reasonable efforts to **limit protected health information to the minimum necessary** to accomplish the intended purpose...” [160] (emphasis ours). HIPAA does not, however, offer concrete guidelines on determining the necessary minimum, so the task falls on the data custodians who then rely on subjective judgment: “*It’s not like it’s a super clear ‘Yes, this; no, that’. An investigator or a team needs to lay out their study design, the outcome variables, the exposure variables, the inclusion criteria, and then all of the co-variables that they’re interested in. And that defines what you can ask for ... There’s no hard and fast rule*” (C5). The custodians fulfill the requirements by sharing, for example, only a subset of data fields or a randomly sampled portion of a dataset. This flexibility provided by HIPAA is both a

burden and an asset, allowing data custodians, who are often deeply familiar with their data and the types of research questions they can address, to use their expertise to tailor data release and protect patient privacy. However, it can be extraordinarily time-consuming, unscalable, and may not operate on provable guarantees of privacy. Indeed, R6 remarked that “*It’s just a process and it takes a long time; my initial [Medicare data] request probably took about a year*” (R6).

Access control is also sociotechnical: Data custodians consider institutional and individual history when granting access to data, in addition to the research plan itself. Beyond the concept of only sharing the minimum necessary, data custodians also evaluate the trustworthiness of the researchers themselves—both to be stewards of the data, and to responsibly conduct valid medical research. C4 described relying on the requester’s reputation when deciding how much sensitive data to share: “*What we do is [we] actually credential researchers, and then they become trusted researchers. They are authorized to use the data at a particular level of sensitivity*” (C4). If a researcher has a history of non-compliance in their past use of data,

they might become subject to more intense scrutiny. Some institutions, however, might credential larger groups of researchers, automatically (or more quickly) granting access to individuals within vetted groups: *“I think within [our institution] they gave kind of a waiver; we didn’t have to have every single thing reviewed, but we still had to report what we’re downloading so they could audit us any time”* (R7). HIPAA does not explicitly cover these mechanisms of sociotechnical access control, which results in varying implementations across institutions.

Privacy policy might sometimes be used as an excuse for not sharing data. Some participants expressed their frustration with the difficulty of acquiring enough data, e.g., to build large-scale ML models, even though that data existed. These barriers might be a result of *“a strong perception of monetary value in that data”* (R2), and hence the unwillingness of some institutions to share: *“Maybe they use privacy as an excuse not to share it. But I think mostly they don’t want to give access”* (R9). Certain facilities might be apprehensive of the reputational or even regulatory risks and attempt to weaponize privacy restrictions to avoid data sharing for research: *“It’s the same thing with these drug fees treatment facilities ... they use the SAMHSA rule¹ as a way to sort of make sure that nobody can tell what they’re doing”* (C3). These participants point towards the politics and adversarial actors involved in accessing healthcare data, which we discuss in Section 6.2.

Even when privacy guidelines are explicit, some data custodians consider them insufficient in practice and implement extra protections. HIPAA requires one of two methods for de-identification: either an *Expert Determination*—using statistical methods—that the data are not identifiable within some small acceptable level of risk, or the *Safe Harbor* method, which pre-defines 18 types of personally identifiable information that researchers can remove to comply with de-identification (see Section 2). Although standard practice amongst our participants was to use the Safe Harbor method and remove the specified 18 patient identifiers, some data custodians remove additional data because they do not believe that the HIPAA guidelines suffice for actual de-identification of their data: *“I look at the content, it’s not really de-identified. It’s just formally de-identified—it meets the requirements”* (C1).

Participants described removing the names of a referring physician, other clinicians, or places, such as institutions or hospitals—the *“things that are not technically identifiable, but you feel like conceptually is the right thing to do to remove those”* (R6). A rare disease can be another identifying feature: *“We need to be careful that we don’t include in that shareable part of something that might identify a patient.... If you share a rare cancer and then the ZIP code of the patients, it becomes probably [identifying]”* (C1). Privacy considerations extended beyond PII and the identities of individuals. Several participants reported removing institutions’ names due to reputational risks: *“We recognize that provider is a really important variable when looking at healthcare outcomes, but we also don’t want a study to identify an individual provider as doing exceptionally poor”* (C6). As we discuss below and in Section 6, these practices indicate space

for formal tools to provide measurable guarantees, in line with HIPAA’s option for *expert determination* via statistical methods.

Data custodians differ in how they implement anti-reidentification rules about the display of aggregate data over small sample sizes. In order to prevent re-identification of patients who are part of a small sample, data custodians prohibit the disclosure of summary statistics on these small pieces of data. In contrast to classic cell-suppression algorithms which use integer programming [32], these rules impose a simple threshold on individual counts, e.g., they prohibit the display of the number of individuals from the custodian’s data with a certain disease if that number is less than the threshold. The choice of the threshold value, however, differs among data custodians. One of the data custodians in our participant group reported increasing their threshold after conversations with data subjects who expressed discomfort with the originally proposed value. In the context of elective participation, the privacy dilemma is whether such a threshold must be chosen to *guarantee* protection from re-identification—assuming we can find this number—or to make *“people feel comfortable,”* even though stronger protections might hinder innovation.

4.3 Privacy vs. data utility

Given the backdrop of how medical researchers interpret and implement privacy regulations, we now address the innovation angle of **RQ2** and **RQ3** by exploring how these interpretations put privacy at odds with data utility. Indeed, the requirement to minimize access to data, for patient privacy, is directly in tension with the explorative spirit of research which relies on insights from data for producing new knowledge. C5 explained that when denying access to data, i.e., when researchers request more data than strictly necessary, data custodians have sympathy: *“It’s rarely that I would necessarily suspect ill intent. It’s mostly that people from a research perspective are like, the more data I collect, the more stuff I can analyze”* (C5). But, as discussed in Section 4.2, data custodians and researchers *must* follow the mandate to maintain patient data privacy.

Here, we find that researchers and data custodians both understand and respect patient privacy. In some cases, they have developed workarounds for data that cannot be shared but are critical for advancing medical science—including custom obfuscation for data fields deemed identifiable which preserves data utility, the use of commercial tools to link identifiable records without revealing patient identity to researchers, and the use of certain data visualizations over others (i.e., so as not to show individual data points). We also find manifestations of this tension between privacy and data utility that are unresolved, stifling medical research, including challenges with deduplication across multiple de-identified datasets.

Data custodians employ pseudo-differential-privacy data transformations to obfuscate dates so that data are both HIPAA-compliant and still useful for analysis. Date elements, such as procedure or birth dates, and addresses are considered PHI and must be redacted for data to be considered de-identified by HIPAA [160]. Medical-research questions, however, often require temporal and geographical information, so the responsibility of determining how to provide it while complying with the privacy regulations falls on data custodians. One approach participants described was to derive and share

¹SAMHSA is a privacy rule that provides extra protection for the records of patients seeking treatment for drug and alcohol dependency, e.g. excluding them from any legal discovery [161].

a proxy variable that addressed the research question but did not directly reveal the dates: “They wanted to know the date of birth and the date of the first COVID test. And we couldn’t give them that. But we could give them [that] the date of the first COVID test was within X days of the date of birth” (C3).

Multiple participants interpreted random date-shifting as compliant with HIPAA’s requirement to remove dates. C2 explained: “Date shifting that stays consistent within one person’s record so that you still can get the right order of things. But like my dates might be shifted by 30 days and your dates might be shifted by 45 days” (C2). However, R2 remarked that, in longitudinal data, even shifted dates, alongside laboratory measurements or other high-precision values, can pose an identifiability risk: “If you shift all the dates by a certain amount, then you can still look and see, oh, this patient has events separated by exactly this precise time duration.” As we discuss in Section 6, this use case would be an excellent candidate for provable perturbation techniques, such as differential privacy [44], that ensure the privacy of patient data and enable analysis over sensitive data, rather than the methods in practical use, which are not mathematically guaranteed to provide privacy.

Some researchers employ commercial tools to link or deduplicate datasets in a privacy-preserving way. Some participants reported using commercial privacy-preserving record linkage (PPRL) [26] services to create joint datasets between different sources, enabling new and explorative analyses over joined data without exposing identifiers linking the datasets. C5 explained that “they put the identifiable information and it generates these tokens or hash that are de-identifiable, but they’re unique to each person based on the person’s identifiable data” (C5). As we discuss in Section 6, the auditing and analysis of such tools are necessary to understand their privacy properties, which are not clear from public documents due to the lack of sufficient details, e.g., Datavant [39] (referenced later in this section) appears to *either* operate as a trusted linkage broker or to support only deterministic linkage (hence, being unable to handle errors or missing fields that are common in medical data).

Insufficient deduplication of de-identified datasets challenges research validity, an issue sometimes resolved by record linkage tools. In addition to linking datasets, researchers need to deduplicate de-identified datasets without re-identifying the subjects. Participants explained that patients who sought care at different medical institutions within the same geographic area or participated in data collection for multiple biobanks might appear in multiple datasets, confounding study validation. C1 explained that “Sometimes [researchers] say ‘Oh, I saw this patient. Oh, there is another patient with the same thing.’ And I say ‘No, that’s the same patient’” (C1). Inadequate deduplication could produce statistically unsound results if some patients are represented multiple times, yet it would be a violation of patient privacy to identify them.

Some custodians of EHR data used commercial record linkage tools for deduplication, to address the issue of patients appearing in data samples of multiple healthcare providers. “We all use Datavant... It creates the technical capacity for deduplication and overlap analysis. It’s a hash token based on name, sex, date of birth and maybe SSN” (C6). Others described the research need to combine private identifiable datasets with public de-identified data and the challenge of deduplicating them without violating subject privacy because

“even knowing whether somebody participated in a research study is re-identifying them. And so even if you have identified information on your end and you’re totally allowed to use that, you’re not allowed to know whether this person participated in [another] study” (C2). Future privacy-preserving deduplication tools must address the challenge of removing overlapping subjects with a known identity without violating their privacy, e.g., by obfuscating the fact of the removal from the researcher (see Section 6).

Data custodians use proxy methods for deduplicating medical data without identifiers suitable for PPRL tokens. Non-tabular medical data, such as genomic data, also require deduplication but the standard PPRL tools are not suitable due to the absence of direct identifiers to match. C2 described a workaround for deduplicating genomic data when combining datasets—“There are some methods that [the consortium] actually worked out to do this, like using a small number of genetic variants and running checksums on those files and then comparing the checksums and just removing anybody who has a match” (C2)—but admitted that “that’s what the current practice is in some networks, but otherwise there isn’t a good solution” (C2). C4, whose custody also included genomic data, mentioned that they “enabled certain types of relatedness analyses to help with that deduplication,” which was another proxy linkage method based on familial sharing of genetic segments. More robust and comprehensive deduplication methods are needed.

Publicly releasing data and study artifacts is the norm, yet there are necessarily many privacy considerations. Medical researchers publish both findings and—when possible—their datasets, for transparency or secondary analysis, sometimes required by funders. However, regulation (including data use agreements (DUAs) from funders) restricts what findings and data they are allowed to publish, creating tension between privacy, innovation, and reproducibility.

When data release was allowed, researchers based their sharing decisions on the data’s *type* and what they *perceived* to be safe to share. In the case of genomic data, some participants said that they would release only count matrices and not the full sequencing data, whereas others would upload sequencing data to one of the data publishing platforms, such as cBio Portal [24], but only for mutated samples: “You don’t have the germline that would identify the patient, but you only have the somatic ones² ... They are specific to the patient but usually it’s more difficult to re-identify the patient” (C1). Prior studies have shown that DNA data can reveal identities of individuals [89, 172, 182], but re-identifiability risks become less clear for summary data, such as count matrices, or for genomic data with randomly occurred mutations, e.g., in cancer tissue. Researchers have to make data publishing decisions without the privacy expertise to evaluate the corresponding risks and, as a result, often err on the side of caution, even if it harms the research progress by limiting reproducibility or secondary analysis.

Researchers were generally comfortable releasing summary statistics about their study cohorts or data—e.g., “age distribution ... incidents of and prevalence of codes, like individual codes, like the number of people in your data set that have code one, two, three, four, five” (R1). They were, however, often cautious about releasing

²Germline mutations are inherited from parents and are present in every cell of the body; somatic mutations occur spontaneously and are localized to a specific tissue.

study artifacts. R1, who trained models on clinical notes, said that they would not publicly release model weights, due to information-leakage concerns: *“I’m too NLP-focused here because there are texts [that are used for training these models]”* (R1). In contrast, R9, who worked with medical imaging data, said that they sometimes released the weights of trained image-based models *“because we want other researchers to use our tools and benefit from that”* (R9).

Researchers also try to share their analysis code for transparency and replicability, but R1 cautioned that the code must be carefully checked for PHI as it could become an unexpected source of PHI leakage: *“We’re very careful with code because you never know what’s in your code. Like there’s supposed to be no data, but that’s a classical mistake. Suddenly, there’s a medical record number in your code somewhere”* (R1). Although analysis code might not immediately appear to be medically sensitive, it may require the same privacy treatment, and hence privacy tools, as unstructured medical data, such as clinical notes or images.

Tension between funders’ requirements, peer review, and privacy. The inability of researchers to publish study data was at odds with the requirements of funders, such as the National Institutes of Health (NIH), for increased transparency, and with the peer review process. For example, R6 described releasing the transcripts of interviews they conducted for clinical studies, as it was mandated by the study funder: *“The NIH requirement is that you share data ... If you are funded, you are supposed to do that”* (R6). But researchers were also often unable to publicly share their data due to either privacy concerns, as discussed above, or because they were prohibited from doing so by a DUA in the case of secondary data use. As R6 put it: *“I can’t actually share data because the data that I [would] share, I don’t own”* (R6). The inability to share then caused friction. R1 reported that *“It’s a problem sometimes when we get reviewers who are like ‘I can’t believe they’re not releasing the data’.... we did have papers that have been rejected because of it”* (R1). R3 and R9 shared that they were able to avoid the repercussions of not releasing the data by relaying the DUA restrictions to their funders and editors: *“In one of the papers, one of the reviewers was really pressing that we should release the data, but then the clinic didn’t want to do it. Then we discussed it with the editor and then the editor was like ‘no, it’s OK, you don’t need to share the data’”* (R9). As we discuss in Section 6, both potential techniques for relieving the reproducibility tensions and for privacy audits of study artifacts require investigation.

4.4 Privacy vs. research productivity

The interpretation and implementation of privacy are also in tension with the practical and logistical side of research, as the enforcement of specific privacy policies and requirements impairs productivity and workflow through custodian oversight, storage and access requirements, and ill-suited privacy tools. In this section, we focus on the research-productivity angle of **RQ2** and **RQ3** and explore both the challenges that data custodians face in implementing oversight and bureaucracy and how researchers navigate the imposed restrictions. We find that inaccurate de-identification tools require manual review and correction; data storage and transfer requirements increase costs and can prevent collaboration; and remote analysis environments restrict data movement, research outputs, and researchers’ preferred development tools. Although

these environments allow data custodians to retain control over sensitive data, monitoring researchers’ activity often depends on resource-intensive manual audits that cannot reliably capture the intent behind analysis code. Thus, practices designed to protect patient privacy can delay research, disrupt established workflows, and place significant oversight burdens on data custodians.

De-identification tools built are not entirely accurate, requiring researcher oversight and manual correction. To de-identify data by removing data by type (either only the PII types defined in the Safe Harbor method or, additionally, the data types that might be potentially identifying, as described in Section 4.2), some researchers use commercial tools, whereas others develop custom scripts. Participants reported that neither approach was entirely accurate, so they must manually check and correct tool output.

C6, whose custom de-identification tools were developed in-house, described their auditing process: *“We have these algorithms, we do some keyword searches ... then we have our internal study teams do a random selection of notes and fully read them to see if anything has gotten through there”* (C6). Despite the partially automated process, researchers still have to manually check the outputs, which neither scales nor guarantees complete de-identification outside of the randomly selected examples.

Complex natural-language processing tools also struggle with context and disambiguation of data elements and, as a result, require auditing by researchers: *“The dates are supposed to be scrubbed, but ages are okay to keep. But those NLP pipelines have a hard time telling the difference between an age and a date. ... Half the time [age] shows up as the de-identified bracket, which tells you 10 to 20, so it gives you a range, but you lose the granularity around the age”* (C2). If uncorrected, these imprecise redactions also degrade data utility. The pipelines would sometimes miss even unambiguous elements, such as names: *“we did actually often find information in the clinical notes that should have been scrubbed that got missed. We would frequently find names of patients, or family members, or physicians. And there was a process to report that”* (C2). R6, whose research involved interviewing doctors, also noted the poor performance of some commercial tools: *“If you’re doing a transcription, service will de-identify for you. Although I’ll tell you that I’m using a service right now and I would say half of the stuff is de-identified, half of it is not.”* These experiences contrast with tool performance reported in the academic literature [104, 115].

Participants also shared that de-identifying imaging data was error-prone. R8 and R9 had received medical images with inserted patient text reports or even PHI written over them: *“They can squeeze in a scanned PDF of radiation dose exposure with the name of the patient on. Or on the image itself, somebody may annotate on the packs”* (R8). In one example, the automated tools were set to crop the side of an image with PHI on it but missed any accidental PHI insertion elsewhere. Both free-text and image de-identification require locating PHI before it can be redacted and hence can miss sensitive identifiers if they appear in an unexpected context.

Privacy policy regulates data storage location and transfer, and can impede research feasibility and increase cost. Regulatory bodies (such as the researchers’ IRBs, the data custodians, etc.) can dictate data storage locations to enforce privacy standards appropriate for their data, but the burden of practically working with (and

sometimes building) these different data storage mechanisms falls on researchers, at times causing researchers to decide not to pursue certain work. They might have different requirements for certain types of data (e.g., data from clinical trials), or require that data are stored in a certain geographic area. R4 explained that “*data that comes from some clinical trials where we extract some samples and measurements from patients ... are only allowed to [be] store[d] on [institution] servers*” (R4). Regulation can also restrict health data from being stored outside a US state’s borders (C3) or transferred internationally (R9), which can stifle collaboration. R9 recalled a time when a potential collaboration with international researchers required infrastructure that they could not justify building: “*It was a clinic in a different country, and they said ‘our data should not leave the country’ but we didn’t want to implement all our software and all our tools in their servers because it would take a long time. ... we were like ‘OK, we are not gonna do it then’*” (R9).

When data transfer is allowed, the process still can affect research productivity. Participants described their experience with data transfer via secure FTP, via a cloud (including commercial providers, such as AWS), or even by physical delivery of hard drives, affected by the size of the data. Datasets that contain raw sequencing data, EHRs, or imaging data can reach multiple terabytes, so transferring them might take days: “*The real difficulty is in the transfer of the data cause imaging data is very heavy so it’s a process to ... make sure we can get them over at a reasonable time*” (R8). Participants also voiced concern about the resources needed to transfer that amount of data: “*if we have to send one terabyte out of data out to 100,000 researchers, that would cost the amount of power needed to run New York City or something like that*” (R5).

Sending physical hard drives by mail (with an emailed password) also sometimes occurs, though participants said it is uncommon now except for extremely large datasets (such as medical images). The CMS, for example, used to mail physical hard drives, whilst having “*ten to twenty million beneficiaries and each of them has a lot of data. So it’s multiple terabytes, many, many terabytes of data. I think they send it in like ... 100 terabyte drives*” (R6).

Policy restrictions for data storage location create an opportunity for privacy-preserving data analysis, e.g., by using secure multi-party computation or homomorphic encryption. The sheer size of medical datasets, however, can make privacy-preserving approaches challenging due to bandwidth and computational constraints. We discuss this tension in Section 6.1.

Remote analysis environments provide a solution to data transfer and storage issues but impose their own restrictions. Rather than navigate regulatory and logistical hurdles for transferring data to and storing data at the researchers’ institutions, some data custodians maintain thin clients (remote virtual machines) on which researchers access and analyze data. R6 described that in these environments, researchers “*buy a seat, and then that seat allows you access to log into the system where your data is*” (R6). All analysis happens in the remote environment: “*you... process the data [and] run your scripts in a cloud environment they provide for you*” (R5).

C6 explained that “*we don’t want [the data] to leave our institution and our safeguards*” (C6) for the sake of maintaining control of the data’s storage, access, and security. However, data custodians have to ensure that researchers do not violate privacy policy when

using these remote development environments, which necessitates imposing restrictions. C6 acknowledged that “*most researchers in this space would prefer to have the data themselves to work in the environments at their institution, the environments that they created, their lab servers*” (C6). We explore how custodians and researchers navigate these tensions throughout the rest of the section.

Remote analysis environments restrict downloaded and uploaded data for privacy, shaping research timelines and output. The remote environments allow data custodians to control any data leaving the system and thus to protect patient privacy by preventing researchers from downloading “*any PII*” (C5) or even entire datasets, yet can delay researchers’ progress due to manual oversight, or restrict their summary outputs (e.g., visualizations).

Restrictions are sometimes enforced manually, taking time, and other times automatically, risking inaccuracy. A team may manually inspect data requested for download from the remote analysis environment: “*You have to put in a data request for downloading. Then, it will take several days for them to go through everything*” (R3). Other platforms perform an automated inspection in the background which sometimes results in opaque enforcement: “*One time I had an issue where I was downloading a few JPEG files at once. And it flagged it for excess like data egress from the platform. And then it blocked my account for a day*” (R10).

Download restrictions extend beyond the dataset itself to results and visualizations. R3 described a case when their download request was rejected because it included a scatter plot: “*Because a scatter plot implies that on a specific XY area, that is an individual... That is one individual data point. So they will not let you publish scatter plots. So what we did was we generated bubble plots*” (R3).

Imported data, which researchers may use to link to other datasets and perform analyses, are also restricted in remote analysis environments. C4’s organization “*do[es]n’t allow people to import any PHI*” (C4). R2 also said that “*[the] external data sets you can bring in are limited.*” They described that while external datasets were not prohibited, it was a “*significant challenge*” to figure out how to solve “*a lot of security questions around how to set up those environments to protect the [local] data.*”

This fine-grained control over the analysis environment enables data custodians to protect the privacy of patients in the dataset itself, but can slow research and shape research output, perhaps revealing space for bespoke privacy automation tools or policy to support productivity and privacy.

The software development experience in remote analysis environments varies due to control by the data custodian rather than the researcher. Data custodians who maintain remote analysis environments necessarily also maintain control over the actual user experience of the researchers writing code and doing data analysis in the environments. While this control is necessary for privacy and security—indeed, part of the very point of this type of data access model—it can hinder researchers by taking them out of their known environments and sometimes disallowing the use of the tools that they are comfortable with.

While some environments had all the researchers needed, R2 and C2 described needing to install additional third-party tools or libraries. R2 said that in some environments, installations were allowed “*through a system like jFrog, which ... claims to be able to*

detect malicious packages.” (R2) In the most restrictive environment, R2 described that “you can’t copy and paste into or out of the environment, you can’t install packages locally..., [and] the tools you have to work with are very limited” (R2). Although R2 acknowledged the very real threat of malicious actors and libraries [184], they were frustrated with how their workflow and analyses were restricted. They discussed hypothetical workarounds for some restrictions, including a script to get around paste restrictions but said that “It’s just about figuring out how you can do your work most effectively despite, in some cases, the blockers that have been put in place for what are claimed to be security reasons and sometimes are security reasons and sometimes are security theater reasons. Or sometimes frankly feel like they’re there to explicitly prevent use of data because [the data custodians] don’t actually want people using their data. They just want to say that their data can be used” (R2).

Mechanisms of analysis oversight by data custodians vary in practice due to the immense resources it requires, though oversight via manual auditing of code prevailed. Remote analysis environments—managed by the data custodian or their contractors—allow data custodians, in theory, complete control and purview over analysis and data import/export. In practice, oversight over code varies because nuanced oversight is incredibly complex and resource-intensive. Nonetheless, several data custodians reported that they manually audited researchers’ code.

Data custodians inspect analysis code in the remote environments, as a privacy and security measure to ensure the researchers are not violating the terms of the Data Use Agreement (DUA) and downloading or analyzing data in a way that is disallowed or violates patient privacy. Participants discussed the challenge of meaningfully auditing code to understand the purpose of the code due to the inherent difficulty in reading and understanding others’ code: *“the one piece that we are missing that is both really helpful for understanding how folks are using our data, but then also from that privacy piece is ‘What are they actually doing? Are they data cleaning? Are they running models? Are they filtering the data down to 5 rows?’” (C6). C4 noted the importance (and implied rarity) of comprehensive code documentation: “unless somebody is really good at commenting their code, it’s really hard to interpret what exactly is going on in somebody’s head” (C4).*

In addition to the function of the code itself, data custodians may be concerned with the *interpretation* of the data—which is not necessarily something that will be written in code. C4 explained that *“even the retrieval of particular data points... doesn’t necessarily tell us exactly what they’re going to ask with those data points” (C4).*

The manual nature of auditing analysis scripts cost resources and drove uncertainty for both data custodians and researchers. C4 remarked on the potential for automation—*“we are definitely flying by the seat of our pants... this is a place I think we would really use a lot of help, to standardize how we go about this” (C4).* (Though they noted that such automation and standardization could just *“create some hurdles for somebody to avoid” (C4).*) C6 also wondered: *“When we think of re-identifiability issues in these data, we theoretically could have a system that says ... ‘You’ve tried to run code that will reveal line level for more than five patients and that’s not permitted. Rewrite your code to do that or you can’t filter down to this level.’ So I think that there is an opportunity there” (C6).* As we discuss in Section 6,

this is an opportunity for the programming languages community to remove privacy-caused practical barriers for researchers, and lift the burden on data custodians.

5 Related Work

Now we turn to a number of related works about privacy, privacy policy, and privacy tools in medical settings.

Medical privacy in practice. Substantial prior work has found that patients care deeply about the privacy of healthcare data but that they also are largely willing to trust medical researchers with the data [31, 35, 41, 63]. Patients, especially those with rare diagnoses, *want* their data to be used for medical research [31]. Copious work has also focused on patient privacy in certain geopolitical case studies, like during the COVID-19 pandemic [9], or currently about reproductive healthcare in the US [94, 99]. There is little work, however, on the experiences of medical researchers with privacy technologies. Guo et al. studied de-identification in practice [57]—more broadly than medical researchers—and found that differential privacy (DP) usage varied. Ngong et al. [112] studied the usability of DP tools. Raisaro et al. [127] deployed a homomorphic encryption-based tool for genomic analysis in a hospital and surveyed the physicians who tested it. Kayaalp et al. [79] described their experience with manual annotation and de-identification of clinical texts without clear standards. Krzyzanowski and Manson [82] examined HIPAA’s Safe Harbor provision on geographic data and its misinterpretations by epidemiologists. Our research builds on this body of work by adding to our understanding of privacy in medical research—which patients set as a condition for the use of their data.

Navigation of privacy. A recurring theme in privacy and policy scholarship is that privacy becomes harder to protect as the volume of available information grows. Early studies [106, 116, 138] argued that treating personally identifiable information as a fixed set of removable attributes was misguided, as quasi-identifiers could still enable re-identification. The rise of “Big Data” further increased the risk of sensitive inferences [33, 142, 168], which in turn spurred discussions on privacy harms and how to define them [23, 33, 126, 141]. While early studies used health data illustratively, later work examined health privacy policies, including HIPAA, and identified their limited coverage of entities [21, 29] and data sources [132].

De-identification techniques for medical data. Efforts to develop tools for de-identification of medical records began with emerging legal requirements for sharing [47] and protecting [158] clinical data. The first systems used template search and basic heuristics to find and redact PHI [147]. Researchers later shifted to developing algorithms and tools based on k-anonymity [4, 68, 70, 86, 125, 148, 151, 157] or differential privacy [85, 146]. Early tools for automated de-identification of free text were similarly based on regular expressions, look-up tables, and contextual rules [108, 156]. The progress then shifted to using natural language processing [155, 177], neural networks [40, 92, 104], and eventually to language models [91, 179]. Relevant works on de-identification of medical imaging data, such as X-ray or ultrasound scans, rely on neural networks to detect and remove burned-in PHI from images [103, 111, 165, 183].

Re-identification risks and measures. In the first famous attack demonstration, Sweeney [149] re-identified the Governor of Massachusetts in a dataset of hospital-discharge records. The work was followed by more publications of re-identification attacks on health data [19, 46, 72, 150, 153, 154] as well as their analysis [45, 62]. Other works [12, 38] estimated the risks of sharing de-identified or limited health datasets when they could be linked to another data source. Finally, additional studies used a game-theoretic approach [171, 178] and generative models [131] for the risk assessment.

Studies on re-identifying individuals from genomic data began with evaluating the identifiability of genetic variations [48, 54, 67, 89, 96, 97, 120], and progressed to recognizing individuals in mixed genetic samples [17, 28, 64, 134] or summary-level genetic data [11, 20, 60, 71, 135, 166, 169, 172, 182]. We refer readers to reviews on privacy challenges and opportunities in genomic data processing and sharing [5, 15, 102] for a more extensive analysis.

Privacy-preserving record linkage (PPRL). Early PPRL protocols [16, 27, 117] for health data compared hash identifiers derived from personal information. This approach, however, was only somewhat effective in practice, as real-world health data often contained discrepancies and reporting errors which led to mismatches. More recent approaches rely on probabilistic record matching by using Bloom filters [25, 43, 84, 137, 163], locality-sensitive hashing [175], differential privacy [61, 129], and secure multi-party computation [69, 83]. We refer the reader to Christen et al. [26] for an extensive overview of PPRL. Finally, another body of recent work [13, 81, 98, 101, 113] evaluated the effectiveness of existing PPRL protocols in practical clinical settings. Despite the progress in PPRL design—and as supported by the experiences of our participants—the linking challenges with data reporting differences, scalability, and privacy leakages require further research.

Relevance to non-medical sensitive data. Although we focus on medical research, the challenges that our participants faced and the relevant techniques extend to other domains that share and analyze sensitive data. De-identification and anonymization support the disclosure of sensitive microdata, including financial and population statistics [14, 42, 119, 140], as well as narrative data [74, 105]. Similarly, PPRL applies to humanitarian-aid distribution [130], surveys [136], and counterterrorism [55]. Finally, code auditing that captures researchers’ intent relates to techniques for automated comment generation [65] and code review [88, 180].

Medical research privacy beyond the US. Privacy laws, such as Canada’s PIPEDA [121] and the EU’s GDPR [49], create challenges that are similar to what medical researchers face in the US. Prior work has examined the GDPR’s effects on consent and data sharing [58, 133, 167] and applied PPRL to medical records in Australia [113, 128] and Germany [59, 143]. We leave a cross-jurisdictional comparison of medical privacy practices for future work.

6 Discussion

Through our interviews, we have the opportunity to identify key avenues for future research and development to alleviate tensions between privacy and research that arise in practice in medical research. We remark that most participants were unabashedly excited

to talk to us, as privacy researchers; some even referenced having mentally prepared notes of what they wanted to tell us about privacy issues, and a few of our recommendations are drawn directly from interviews (with our expansion and interpretation). This shows that there is space for collaboration between our communities, and we strongly encourage future researchers (§6.1) to talk directly with medical researchers when developing new tools. Our interviews also surfaced tensions and barriers that may be alleviated with policy interventions (§6.2).

6.1 Recommendations for technologists

We identify a few key concrete directions for tools or research that would directly remove practical barriers for the medical research community. Even though some of these technologies may exist already, our work shows that they have not made their way to the medical research community; we urge technologists and researchers to consider developing and designing their tools with the specifics of medical data in mind.

Automated code auditing. Remote analysis environments run by data custodians are common, so that the custodians can oversee data analysis and prevent download. However, participants told us that the audit process could be inefficient, spotty, and subject to errors due to its manual nature (Section 4.4). As participants explained, there were complex rules to enforce—not just, for example, whether the code returns a single row of data, but also whether it runs analysis that could be misinterpreted and harm racially or ethnically minoritized groups—which may not be readily apparent from the code itself. Here, we see an opportunity for automated or semi-automated code auditing, in a way that both helps data custodians understand the analysis code and communicates to medical researchers what is being analyzed and the outcome of the analysis. A code auditor should build on the culture of trust between medical researchers and enable medical research while ensuring the appropriate privacy protections for patients.

Practical and measurable privacy guarantees. Our work shows that medical researchers recognize that simply removing the HIPAA-required identifiers can be insufficient for eliminating the risk of re-identification and call for de-identification protocols with provable guarantees. We recognize that, despite the early vision [148, 151], achieving traditional privacy guarantees, such as k -anonymity [152], for modern real-world medical datasets can be challenging, due to their large size and dynamic membership. As an alternative, we suggest focusing on developing easy-to-apply techniques with measurable privacy properties. For instance, participants told us that dates were often important to their analyses. However, HIPAA includes dates as identifiers [160]. In response, some researchers find proxies for dates, and some researchers perturb the dates within a certain range by using custom scripts developed in-house—approximating differential privacy (DP). Indeed, prior work has found that only some disciplines use DP [57], and here we remark that DP would be an asset to medical researchers and would also improve privacy guarantees. We encourage the DP community to work with medical researchers to build tools to apply DP to their datasets.

Semi-automated de-identification of unstructured data. Participants remarked that de-identification of unstructured data (such as

clinical notes or X-ray images) was even more error-prone and time-consuming than de-identification of structured medical records. As we described in Section 5, various approaches exist for removing PHI from free-text data. These tools, however, still have imperfect recall and precision, and particularly struggle with ambiguous values, such as numbers, or rare names [124, 164]. De-identifying imaging data is even more challenging because it requires first finding any burned-in text on an image and then determining whether it is PHI. We observe, however, that in the medical context, unstructured data often are accompanied by structured patient information that tools could use to improve de-identification. In fact, early systems for free-text de-identification already used user-specific profiles [148] for automated PHI lookup. We suggest this as an opportunity for developers: medical researchers need a tool to automate de-identification of unstructured data that could benefit from the extra supplied context and tunable parameters (e.g., setting combinations of certain information as sensitive) to increase its efficacy.

Data-aware privacy tools. Research on privacy-preserving analysis or computation, e.g., that uses HE or MPC, tends to focus on the security and efficiency of developed tools and use complete, properly formatted, and modestly sized datasets to evaluate their performance. As described by our participants, medical research data in practice, however, are often large, contain discrepancies and reporting errors, and can be structured in one of the many possible formats. These issues are exacerbated when the analysis is multi-center and each site wishes to conduct research without directly sharing its data with the study partners. Hence, we first recommend privacy researchers account for data heterogeneity when developing tools for privacy-preserving medical analysis. Second, we suggest the development of techniques for privacy-preserving data harmonization as a vital research direction.

Analysis of commercial privacy-preserving record linkage (PPRL) tools. Participants described their use of commercial tools for PPRL and record deduplication. We find that besides some evaluations of the tools’ linking performance [13, 98, 101], there has been little to no work that analyzes the privacy and security properties that these tools provide. It is in part due to the lack of transparency by the commercial vendors—it is unclear what exact linkage algorithm and threat model they implement. We invite tool vendors to publish their algorithm specifications and work with privacy researchers on improving their tools. We invite researchers to analyze these tools regardless, for the sake of patient privacy in medical research.

Validation of research results without access to data. There is a tension between medical researchers’ inability to publish secondary data that they analyze, due to ownership restrictions or patient privacy, and the growing demand for research transparency. We see this as an opportunity for designing tools that can attest to the correctness of data analysis and usage patterns without revealing the data inputs. The field of verifiable computing [53, 122] offers suitable tools for this task, and there is already work, for example, on using them to attest to the correctness of model training [1, 52, 145] or genomic operations [51]. These tools require further research to be applicable to the complex needs of medical data analysis, and we invite practitioners to investigate this opportunity.

6.2 Recommendations for policy-makers

While privacy-preserving tools are a critical part of addressing the practical barriers to privacy and medical research innovation, it is also important to consider the broader bureaucratic framework of the US healthcare and medical research system.

Update state and national policy to align healthcare and advertising data. Some participants expressed that all the time and effort devoted to patient privacy were not useful, at a population level, because they hypothesized that many diagnoses could be derived from data more publicly available, e.g., through web searches and unencrypted chat history. Indeed, recent work has found that browsing history can leak diagnoses, and that health ads are often deceptive [181]. Additionally, fitness and menstrual trackers are not subject to HIPAA, but store sensitive health information that may indicate, for example, miscarriage in a state where women have been arrested for miscarrying [8, 99].

While we cannot advocate for removing privacy protections for health data just because they may be available elsewhere, this broader view of the monetization of data should direct future policy both about health data and health proxy data that are outside the realm of HIPAA. Additionally, future work on the public nature of health data outside doctor-patient communications is necessary, in order to justify the privacy barriers required for medical research.

Structural and systemic changes to reduce the power of adversaries. Participants identified insurance companies as acting adversarially by, for example, increasing prices or denying coverage to patients with certain diseases or disease risk factors. Participants worried that if insurance companies had access to their data, they could glean new information and adjust coverage. Policy-makers have the opportunity to limit this privacy risk by removing the value of such data to insurance companies and more strongly legislating how insurance companies must equitably protect patients.

7 Conclusion

Through 16 semi-structured interviews, we explored how bureaucratic and technical mechanisms exacerbated tensions between patient privacy and medical research. We found that medical researchers encountered a number of privacy restrictions when accessing and handling data, and that while they deeply respected patient privacy, they were frustrated by some of these restrictions. We identified avenues for future research to develop policy and technology to *relieve* this tension without compromising patient privacy or research quality.

Acknowledgments

We are very grateful to our participants for their time and insight. We thank Nafisa Yeasmin for her contributions to this project. This work was supported by the Warren Alpert Foundation (G.G.) and the National Science Foundation under Grant No. 2445401 (L.S.). This work was also supported by the Barnard College Summer Research Institute through The Class of 1982 and the Office of the Provost. We follow standard authorship convention in usable security, with the final two authors in ascending order of leadership.

Ethical Considerations

We deeply considered ethics throughout each stage of this project, and we believe that we conducted our work ethically.

Ethical principles. Our work is aligned with the ethical values in the Menlo and Belmont reports [80, 159]. In line with the principles of *justice* and *respect for law and public interest*, our Institutional Review Boards (IRBs) approved our study. There was a minimal power imbalance with interviewees because they were also mostly researchers with PhDs or MDs. Given the average pay of medical researchers, \$50 was a fair but not overpowering incentive. We aligned with *beneficence*—maximizing benefit and minimizing harm—by minimizing the risk of data leakages by transcribing with locally hosted tools and redacting transcripts before analysis. To *respect participants’ autonomy*, we conducted a thorough informed consent process, including asking participants to briefly recite to us the (minimal) risks, benefits, and the study purpose before we began the interview. We also checked in with them throughout the interview and before asking sensitive questions, reminding them that they could tell us *hypothetical* stories.

Stakeholders, benefits, and harms. We identify medical researchers, data custodians, patients, health care providers, and health care institutions as the stakeholders in our work. Our participants (medical researchers and data custodians) were affected by research procedures—they may have been harmed by feeling stress during the interview, and they also may have (as a few said directly) benefited from discussing privacy concerns and questions with us. Compensation (\$50) also modestly benefited those participants who accepted it (we offered it to all). We do not expect other stakeholders were affected by our research procedures.

All stakeholders might be affected by the publication of our work. Our intention is that all stakeholders will benefit indirectly from the publication of our paper, as it can lead to increased collaboration and understanding between the medical research community and the privacy community, leading to tools and policies that better fit researchers’ needs, and therefore better serve patients, health care providers, and health care institutions in improving patients’ health more quickly and safely. However, it is also possible that our work will lead to intangible harms to some of our stakeholders. If we inadvertently publish quotes that reveal confidential information that our participants told us (e.g., revealing PII about themselves or patients, or rule-breaking), harms could extend to participants (for revealing information), patients (a violation of their privacy), and health care institutions (sanctions for rule-breaking, or consequences of proprietary information being leaked).

Mitigations. Given that harms would mostly arise from publishing information in quotes that should be kept private, we have redacted all quotes, being conservative about what remains in the paper. If we felt anecdotes were extremely specific to one person or one institution, we either redacted them or chose not to include them. We also invited each participant to read our paper before publication (after submission) in case they identified any additional information requiring redaction. In response, two participants requested *minor* modifications to the text, and three direct quotes were redacted.

Additionally, early in the course of our interviews, one participant (whose data were excluded) indicated immediately after the interview that they had significantly lied by omission. After this interview, we re-evaluated our consent process to consider whether it was a failure of consent that affected this participant. Both researchers present for the interview agreed that the consent process had seemed rushed due to the interviewee jumping to interview topics during the consent process, but agreed that the participant had understood the risks and we had not pressured them into saying anything, as we had checked in about comfort and topic multiple times. We also noted that the participant had (seemingly genuinely) said, multiple times, that they were really excited to participate and contribute data to our project. After a lengthy discussion about our consent process, and consultation with peers (non-authors), we decided not to change the formal consent process. For all future participants, we emphasized *even more* that we were not judging them, and that if they *were* going to reveal something that could get them in trouble, we were comfortable hearing it, but they were welcome to use hypotheticals or to keep it to themselves.

Decision. Given the above stakeholder analysis, alignment with the principles of the Belmont and Menlo reports [80, 159], careful consideration of harms and benefits, and mitigations, we used a consequentialist lens through which to evaluate the ethics of our research. We considered whether the potential benefits of our work outweighed the harms—and we believe they do, as the main harms stem from publishing quotes (which we have very carefully gone through), and potential benefits include public health.

Open Science

We provide our interview protocol in Appendix A, the codebook from our qualitative analysis in Appendix B, and the recruitment messages that we used in Appendix C. As our participants were often the leading experts in narrow sub-fields, we are unable to release the audio recordings and transcripts of our interviews due to the risks of participant re-identification via their areas of expertise or work descriptions.

AI Use

We used Whisper [118], open-source LLM software from OpenAI, and the Microsoft Teams HIPAA-compliant transcriber to create initial transcriptions from the audio recordings of our interviews. Whisper operates entirely offline, so no data are sent back to OpenAI. The HIPAA-compliant transcriber stores the data only in the researcher’s account within the institution’s controlled infrastructure. We manually checked transcriptions as part of our analysis process. We also used AI to occasionally refine phrasing and to proofread the final text for typos and grammatical errors. We did not use generative AI in any other part of our research (including but not limited to literature review, study design, and analysis).

References

- [1] Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos. 2024. Zero-knowledge proofs of training for deep neural networks. In *ACM Conference on Computer and Communications Security*.
- [2] Julia Adler-Milstein, David W Bates, and Ashish K Jha. 2011. A survey of health information exchange organizations in the United States: implications for meaningful use. *Annals of internal medicine* 154, 10 (2011).

- [3] Agency for Healthcare Research and Quality. 2025. Healthcare Cost and Utilization Project (HCUP). <https://www.ahrq.gov/data/hcup/index.html>
- [4] Gagan Aggarwal, Rina Panigrahy, Tomás Feder, Dilys Thomas, Krishnam Kenthapadi, Samir Khuller, and An Zhu. 2010. Achieving anonymity via clustering. *ACM Transactions on Algorithms* 6, 3 (2010).
- [5] Mete Akgün, A Osman Bayrak, Bugra Ozer, and M Şamil Sağıroğlu. 2015. Privacy preserving processing of genomic data: A survey. *Journal of biomedical informatics* 56 (2015).
- [6] Isabelle Rose I Alberto, Nicole Rose I Alberto, Arnab K Ghosh, Bhav Jain, Shruti Jayakumar, Nicole Martinez-Martin, Ned McCague, Dana Moukheiber, Lama Moukheiber, Mira Moukheiber, et al. 2023. The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet Digital Health* 5, 5 (2023).
- [7] All of Us Research Program Investigators. 2019. The All of Us research program. *New England Journal of Medicine* 381, 7 (2019).
- [8] Cary Aspinwall. 2025. These States Have Investigated Miscarriages and Stillbirths as Crimes. <https://www.themarshallproject.org/2025/04/02/law-pregnancy-california-ohio-georgia-alabama> Accessed: 2025-11-12.
- [9] Oshrat Ayalon, Dana Turjeman, and Elissa M Redmiles. 2023. Exploring privacy and incentives considerations in adoption of COVID-19 contact tracing apps. In *USENIX Security Symposium*.
- [10] Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. 2013. Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *ACM Workshop on Privacy in the Electronic Society*.
- [11] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. 2016. Membership privacy in MicroRNA-based studies. In *ACM Conference on Computer and Communications Security*.
- [12] Kathleen Benitez and Bradley Malin. 2010. Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association* 17, 2 (2010).
- [13] Elmer V Bernstam, Reuben Joseph Applegate, Alvin Yu, Deepa Chaudhari, Tian Liu, Alex Coda, and Jonah Leshin. 2022. Real-world matching performance of deidentified record-linking tokens. *Applied Clinical Informatics* 13, 04 (2022).
- [14] Jelke G Bethlehem, Wouter J Keller, and Jeroen Pannekoek. 1990. Disclosure control of microdata. *J. Amer. Statist. Assoc.* 85, 409 (1990).
- [15] Luca Bonomi, Yingxiang Huang, and Lucila Ohno-Machado. 2020. Privacy challenges and research opportunities for genomic data sharing. *Nature genetics* 52, 7 (2020).
- [16] Hocine Bouzelat, Catherine Quantin, and Liliane Dussierre. 1996. Extraction and anonymity protocol of medical file. In *AMIA Annual Symposium*.
- [17] Rosemary Braun, William Rowe, Carl Schaefer, Jinghui Zhang, and Kenneth Buettow. 2009. Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS genetics* 5, 10 (2009).
- [18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006).
- [19] John S Brownstein, Christopher A Cassa, and Kenneth D Mandl. 2006. No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine* 355, 16 (2006).
- [20] Diyu Bu, Xiaofeng Wang, and Haixu Tang. 2021. Haplotype-based membership inference from summary genomic data. *Bioinformatics* 37 (2021).
- [21] Mary Butler. 2017. Is HIPAA outdated? While coverage gaps and growing breaches raise industry concern, others argue HIPAA is still effective. *J. AHIMA* 88, 14-7 (2017).
- [22] California Code of Regulations. 1996. Patient Data Reporting Requirements. 22 C.C.R. § 97213.
- [23] Ryan Calo. 2011. The boundaries of privacy harm. *Ind. LJ* 86 (2011).
- [24] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* 2, 5 (2012).
- [25] Yuan Chi, Jun Hong, Anna Jurek, Weiru Liu, and Dermot O'Reilly. 2017. Privacy preserving record linkage in the presence of missing values. *Information Systems* 71 (2017).
- [26] Peter Christen, Thilina Ranbaduge, and Rainer Schnell. 2020. *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer International Publishing.
- [27] Tim Churches and Peter Christen. 2004. Some methods for blindfolded record linkage. *BMC medical informatics and decision making* 4, 1 (2004).
- [28] David Clayton. 2010. On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics* 11, 4 (2010).
- [29] I Glenn Cohen and Michelle M Mello. 2018. HIPAA and protecting health information in the 21st century. *Jama* 320, 3 (2018).
- [30] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.
- [31] Sandra Courbier, Rebecca Dimond, and Virginie Bros-Facer. 2019. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection-quantitative survey and recommendations. *Orphanet journal of rare diseases* 14, 1 (2019).
- [32] Lawrence H Cox. 1980. Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.* 75, 370 (1980).
- [33] Kate Crawford and Jason Schultz. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev* 55 (2014).
- [34] MA Crook. 2013. The risks of absolute medical confidentiality. *Science and engineering ethics* 19, 1 (2013).
- [35] Laura J Damschroder, Joy L Pritts, Michael A Neblo, Rosemarie J Kalarickal, John W Creswell, and Rodney A Hayward. 2007. Patients, privacy and trust: patients' willingness to allow researchers to access their medical records. *Social science & medicine* 64, 1 (2007).
- [36] Johnnie Daniel. 2011. *Sampling essentials: Practical guidelines for making sampling choices*. Sage publications.
- [37] Fida Kamal Dankar and Khaled El Emam. 2013. Practicing differential privacy in health care: A review. *Transactions on Data Privacy* 6, 1 (2013).
- [38] Fida Kamal Dankar, Khaled El Emam, Angelica Neisa, and Tyson Roffey. 2012. Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making* 12, 1 (2012).
- [39] Datavant. 2025. Datavant Connect: Tokenization Software for Health Data. <https://www.datavant.com/white-papers/datavant-connect-overview-tokenization-technology-structured-data> Accessed: 2025-09-03.
- [40] Franck Deroncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24, 3 (2017).
- [41] Rosie Dobson, Helen Wihongi, and Robyn Whittaker. 2023. Exploring patient perspectives on the secondary use of their personal health information: an interview study. *BMC Medical Informatics and Decision Making* 23, 1 (2023).
- [42] George Duncan and Diane Lambert. 1989. The risk of disclosure for microdata. *Journal of Business & Economic Statistics* 7, 2 (1989).
- [43] Elizabeth Durham, Yuan Xue, Murat Kantarcioglu, and Bradley Malin. 2010. Private medical record linkage with approximate matching. In *AMIA Annual Symposium*.
- [44] Cynthia Dwork. 2006. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*.
- [45] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A systematic review of re-identification attacks on health data. *PLoS one* 6, 12 (2011).
- [46] Khaled El Emam and Patricia Kosseim. 2009. Privacy interests in prescription data, part 2: patient privacy. *IEEE Security & Privacy* 7, 2 (2009).
- [47] Mark H Epstein and Barbara S Kurtzig. 1994. Statewide health information: a tool for improving hospital accountability. *The Joint Commission Journal on Quality Improvement* 20, 7 (1994).
- [48] Yaniv Erlich, Tal Shor, Itsik Pe'er, and Shai Carmi. 2018. Identity inference of genomic data using long-range familial searches. *Science* 362, 6415 (2018).
- [49] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). Official Journal of the European Union, L 119.
- [50] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An End-to-End case study of personalized warfarin dosing. In *USENIX Security Symposium*.
- [51] Alexander Frolov, Maurice Shih, Rob Patro, and Ian Miers. 2026. Icefish: Practical zk-SNARKs for Verifiable Genomics. In *USENIX Security Symposium*.
- [52] Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guru-Vamsi Policharla, and Mingyuan Wang. 2023. Experimenting with zero-knowledge proofs of training. In *ACM Conference on Computer and Communications Security*.
- [53] Rosario Gennaro, Craig Gentry, and Bryan Parno. 2010. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In *Annual Cryptology Conference*.
- [54] Jane Gitschier. 2009. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *The American Journal of Human Genetics* 84, 2 (2009).
- [55] Shanti Gomata and Michael D Larsen. 2004. Record linkage and counterterrorism. *Chance* 17, 1 (2004).
- [56] Judith Green and N Thorogood. 2005. Analysing qualitative data. *Principles of social research* (2005).
- [57] Wentao Guo, Paige Pepitone, Adam J. Aviv, and Michelle L. Mazurek. 2025. How Researchers De-Identify Data in Practice. In *USENIX Security Symposium*.
- [58] Dara Hallinan. 2020. Broad consent under the GDPR: an optimistic perspective on a bright future. *Life sciences, society and policy* 16, 1 (2020).
- [59] Christopher Hampf, Lars Geidel, Norman Zerbe, Martin Bialke, Dana Stahl, Arne Blumentritt, Thomas Bahl, Peter Hufnagl, and Wolfgang Hoffmann. 2020. Assessment of scalability and performance of the record linkage tool E-PIX® in managing multi-million patients in research projects at a large university hospital in Germany. *Journal of Translational Medicine* 18, 1 (2020).
- [60] Arif Harmanci and Mark Gerstein. 2016. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods* 13, 3 (2016).

- [61] Xi He, Ashwin Machanavajjhala, Cheryl Flynn, and Divesh Srivastava. 2017. Composing differential privacy and secure computation: A case study on scaling private record linkage. In *ACM Conference on Computer and Communications Security*.
- [62] Jane Henriksen-Bulmer and Sheridan Jeary. 2016. Re-identification attacks—A systematic literature review. *Journal of Information Management* 36, 6 (2016).
- [63] Elizabeth M Hill, Emma L Turner, Richard M Martin, and Jenny L Donovan. 2013. "Let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. *BMC medical research methodology* 13, 1 (2013).
- [64] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (2008).
- [65] Xing Hu, Ge Li, Xin Xia, David Lo, and Zhi Jin. 2018. Deep code comment generation. In *Conference on Program Comprehension*.
- [66] Zhicong Huang, Wen-jie Lu, Cheng Hong, and Jiansheng Ding. 2022. Cheetah: Lean and fast secure Two-Party deep neural network inference. In *USENIX Security Symposium*.
- [67] Mathias Humbert, Kevin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. 2015. De-anonymizing Genomic Databases Using Phenotypic Traits. *Proceedings on Privacy Enhancing Technologies* 2015, 2 (2015).
- [68] Anco Hundepool and LCRJ Willenborg. 1996. μ - and τ -argus: Software for statistical disclosure control. In *International seminar on statistical confidentiality*.
- [69] Ali Inan, Murat Kantarcioglu, Elisa Bertino, and Monica Scannapieco. 2008. A hybrid approach to private record linkage. In *IEEE Conference on Data Engineering*.
- [70] Vijay S Iyengar. 2002. Transforming data to satisfy privacy constraints. In *ACM Conference on Knowledge Discovery and Data Mining*.
- [71] Kevin B Jacobs, Meredith Yeager, Sholom Wacholder, David Craig, Peter Kraft, David J Hunter, Justin Paschal, Teri A Manolio, Margaret Tucker, Robert N Hoover, et al. 2009. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature genetics* 41, 11 (2009).
- [72] Victor Janmey and Peter L Elkin. 2018. Re-identification risk in HIPAA de-identified datasets: the MVA attack. In *AMIA Annual Symposium*.
- [73] Mark A Jensen, Vincent Ferretti, Robert L Grossman, and Louis M Staudt. 2017. The NCI Genomic Data Commons as an engine for precision medicine. *Blood, The Journal of the American Society of Hematology* 130, 4 (2017).
- [74] Samantha Joel, Paul W Eastwick, and Eli J Finkel. 2018. Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018).
- [75] Aaron Johnson and Vitaly Shmatikov. 2013. Privacy-preserving data exploration in genome-wide association studies. In *ACM Conference on Knowledge Discovery and Data Mining*.
- [76] Nikola Jovanovic, Marc Fischer, Samuel Steffen, and Martin Vechev. 2022. Private and reliable neural network inference. In *ACM Conference on Computer and Communications Security*.
- [77] Douglas P Jutte, Leslie L Roos, and Marni D Brownell. 2011. Administrative record linkage as a tool for public health research. *Annual review of public health* 32, 1 (2011).
- [78] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. 2018. GAZELLE: A low latency framework for secure neural network inference. In *USENIX Security Symposium*.
- [79] Mehmet Kayaalp, Allen C Browne, Pamela Sagan, Tyne McGee, and Clement J McDonald. 2015. Challenges and insights in using HIPAA Privacy Rule for clinical text annotation. In *AMIA Annual Symposium proceedings*, Vol. 2015.
- [80] Erin Kenneally and David Dittrich. 2012. The Menlo Report: Ethical principles guiding information and communication technology research. Available at SSRN 2445102 (2012).
- [81] Abel N Kho, John P Cashy, Kathryn L Jackson, Adam R Pah, Satyender Goel, Jörn Boehnke, John Eric Humphries, Scott Duke Kominers, Bala N Hota, Shannon A Sims, et al. 2015. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. *Journal of the American Medical Informatics Association* 22, 5 (2015).
- [82] Brittany Krzyzanowski and Steven M Manson. 2022. Twenty years of the Health Insurance Portability and Accountability Act Safe Harbor Provision: unsolved challenges and ways forward. *JMIR medical informatics* 10, 8 (2022).
- [83] Mehmet Kuzu, Murat Kantarcioglu, Ali Inan, Elisa Bertino, Elizabeth Durham, and Bradley Malin. 2013. Efficient privacy-aware record integration. In *Conference on Extending Database Technology*.
- [84] Pierre KY Lai, Siu-Ming Yiu, Kam-Pui Chow, CF Chong, and Lucas Chi Kwong Hui. 2006. An efficient Bloom filter based solution for multiparty private matching. In *Conference on Security and Management*.
- [85] Hyukki Lee and Yon Dohn Chung. 2020. Differentially private release of medical microdata: an efficient and practical approach for preserving informative attribute values. *BMC Medical Informatics and Decision Making* 20, 1 (2020).
- [86] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. 2005. Incognito: Efficient full-domain k-anonymity. In *ACM Conference on Management of Data*.
- [87] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- [88] Zhiyu Li, Shuai Lu, Daya Guo, Nan Duan, Shailesh Jannu, Grant Jenks, Deep Majumder, Jared Green, Alexey Svyatkovskiy, Shengyu Fu, et al. 2022. Automating code review activities by large-scale pre-training. In *Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.
- [89] Zhen Lin, Art B Owen, and Russ B Altman. 2004. Genomic research and human subject privacy. *Science* 305, 5681 (2004).
- [90] WeiKang Liu, Yanchun Zhang, Hong Yang, and Qinxue Meng. 2024. A survey on differential privacy for medical data analysis. *Annals of Data Science* 11, 2 (2024).
- [91] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. *arXiv preprint arXiv:2303.11032* (2023).
- [92] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics* 75 (2017).
- [93] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *IEEE Symposium on Security and Privacy*.
- [94] Alan F Luo, Phoebe Moh, Cora Sula, Michelle L Mazurek, and Nora McDonald. 2026. Perceived Privacy Risk and Mitigation Post-Roe. In *IEEE Symposium on Security and Privacy*.
- [95] Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, et al. 2007. The NCI dbGaP database of genotypes and phenotypes. *Nature genetics* 39, 10 (2007).
- [96] Bradley Malin and Latanya Sweeney. 2000. Determining the identifiability of DNA database entries. In *Proceedings of the AMIA Symposium*.
- [97] Bradley Malin and Latanya Sweeney. 2004. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics* 37, 3 (2004).
- [98] Keith Marsolo, Daniel Kiernan, Sengwee Toh, Jasmin Phua, Darcy Louzao, Kevin Haynes, Mark Weiner, Francisco Angulo, Charles Bailey, Jiang Bian, Daniel Fort, Shaun Grannis, Ashok Kumar Krishnamurthy, Vinit Nair, Pedro Rivera, Jonathan Silverstein, Maryan Zirkle, and Thomas Carton. 2022. Assessing the impact of privacy-preserving record linkage on record overlap and patient demographic and clinical characteristics in PCORnet®, the National Patient-Centered Clinical Research Network. *Journal of the American Medical Informatics Association* 30, 3 (11 2022).
- [99] Nora McDonald, Alan Luo, Phoebe Moh, Michelle L Mazurek, and Nazanin Andalibi. 2025. Threat modeling healthcare privacy in the United States. *ACM Transactions on Computer-Human Interaction* 32, 1 (2025).
- [100] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019).
- [101] Lisa B Mirel, Dean M Resnick, Jonathan Aram, and Christine S Cox. 2022. A methodological assessment of privacy preserving record linkage using survey and administrative data. *Statistical Journal of the IAOS* 38, 2 (2022).
- [102] Alexandros Mittos, Bradley Malin, and Emiliano De Cristofaro. 2019. Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective. *Proceedings on Privacy Enhancing Technologies* 1 (2019).
- [103] Eriksson Monteiro, Carlos Costa, and José Luis Oliveira. 2017. A de-identification pipeline for ultrasound medical images in DICOM format. *Journal of medical systems* 41, 5 (2017).
- [104] Karthik Murugadoss, Ajit Rajasekharan, Bradley Malin, Vineet Agarwal, Sairam Bade, Jeff R Anderson, Jason L Ross, William A Faubion, John D Halamka, Venky Soundararajan, et al. 2021. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns* 2, 6 (2021).
- [105] Claire A Myers, Shelby E Long, and Faye O Polasek. 2020. Protecting participant privacy while maintaining content and context: Challenges in qualitative data de-identification and sharing. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020).
- [106] Arvind Narayanan and Vitaly Shmatikov. 2010. Myths and fallacies of "Personally Identifiable Information". *Commun. ACM* 53, 6 (2010).
- [107] Sharyl J. Nass, Laura A. Levit, and Lawrence O. Gostin. 2009. *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. National Academies Press.

- [108] Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making* 8, 1 (2008).
- [109] Roberta B Ness, Joint Policy Committee, et al. 2007. Influence of the HIPAA Privacy Rule on health research. *JAMA* 298, 18 (2007).
- [110] New York State Department of Health. 2019. Statewide Planning and Research Cooperative System (SPARCS) Data Reporting Requirements. N.Y. Comp. Codes R. & Regs. Tit. 10 § 400.18.
- [111] Wayne Newhauser, Timothy Jones, Stuart Swerdlhoff, Warren Newhauser, Mark Cilia, Robert Carver, Andy Halloran, and Rui Zhang. 2014. Anonymization of DICOM electronic medical records for radiation therapy. *Computers in biology and medicine* 53 (2014).
- [112] Ivoline C Ngong, Brad Stenger, Joseph P Near, and Yuanyuan Feng. 2024. Evaluating the usability of differential privacy tools with data practitioners. In *Symposium on Usable Privacy and Security*.
- [113] Long Nguyen, Mark Stoove, Douglas Boyle, Denton Callander, Hamish McManus, Jason Asselin, Rebecca Guy, Basil Donovan, Margaret Hellard, Carol El-Hayek, et al. 2020. Privacy-preserving record linkage of deidentified records within a public health surveillance system: evaluation study. *Journal of Medical Internet Research* 22, 6 (2020).
- [114] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004).
- [115] Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, Gabriela Schmajuk, Dana Ludwig, Theodore Goldstein, and Atul J. Butte. 2020. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine* 3, 1 (2020), 57.
- [116] Paul Ohm. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA L. Rev.* 57 (2009).
- [117] Christine M O'Keefe, Ming Yung, Lifang Gu, and Rohan Baxter. 2004. Privacy-preserving data linkage protocols. In *ACM Workshop on Privacy in the Electronic Society*.
- [118] OpenAI. 2025. Whisper. <https://github.com/openai/whisper>.
- [119] Gerhard Paass. 1988. Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics* 6, 4 (1988).
- [120] Andrew J Pakstis, William C Speed, Rixun Fang, Fiona CL Hyland, Manohar R Furtado, Judith R Kidd, and Kenneth K Kidd. 2010. SNPs for a universal individual identification panel. *Human genetics* 127, 3 (2010).
- [121] Parliament of Canada. 2000. Personal Information Protection and Electronic Documents Act (PIPEDA). S.C. 2000, c. 5.
- [122] Bryan Parno, Jon Howell, Craig Gentry, and Mariana Raykova. 2013. Pinocchio: Nearly Practical Verifiable Computation. In *IEEE Symposium on Security and Privacy*.
- [123] PCORI. 2025. PCORnet: The National Patient-Centered Clinical Research Network. <https://pcorntest.org/>. Accessed: 2025-10-09.
- [124] Dzung Pham, Peter Kairouz, Niloofar Miresghallah, Eugene Bagdasarian, Chau Minh Pham, and Amir Houmansadr. 2025. Can Large Language Models Really Recognize Your Name? *arXiv:2505.14549* (2025).
- [125] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A Kuhn. 2020. Flexible data anonymization using ARX—Current status and challenges ahead. *Software: Practice and Experience* 50, 7 (2020).
- [126] W Nicholson Price and I Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature medicine* 25, 1 (2019).
- [127] Jean-Louis Raisaro, Paul J McLaren, Jacques Fellay, Matthias Cavassini, Catherine Klersy, Jean-Pierre Hubaux, Swiss HIV Cohort Study, et al. 2018. Are privacy-enhancing technologies for genomic data ready for the clinic? A survey of medical experts of the Swiss HIV Cohort Study. *Journal of biomedical informatics* 79 (2018).
- [128] Sean M Randall, Anna M Ferrante, James H Boyd, Jacqueline K Bauer, and James B Semmens. 2014. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics* 50 (2014).
- [129] Fang-Yu Rao, Jianneng Cao, Elisa Bertino, and Murat Kantarcioglu. 2019. Hybrid private record linkage: Separating differentially private synopses from matching records. *ACM Transactions on Privacy and Security* 22, 3 (2019).
- [130] Tim Rausch, Sylvain Chatel, and Wouter Lueks. 2026. xDup: Privacy-Preserving Deduplication for Humanitarian Organizations using Fuzzy PSI. In *IEEE Symposium on Security and Privacy*.
- [131] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications* 10, 1 (2019).
- [132] S Trent Rosenbloom, Jeffery RL Smith, Rita Bowen, Janelle Burns, Lauren Riplinger, and Thomas H Payne. 2019. Updating HIPAA for the electronic medical record era. *Journal of the American Medical Informatics Association* 26, 10 (2019).
- [133] John Mark Michael Rumbold and Barbara Pierscionek. 2017. The effect of the General Data Protection Regulation on medical research. *Journal of medical Internet research* 19, 2 (2017).
- [134] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. 2009. Genomic privacy and limits of individual detection in a pool. *Nature genetics* 41, 9 (2009).
- [135] Eric E Schadt, Sangsoon Woo, and Ke Hao. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature genetics* 44, 5 (2012).
- [136] Rainer Schnell. 2014. Linking surveys and administrative data. In *Improving Survey Methods*. Routledge, 273–287.
- [137] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. 2009. Privacy-preserving record linkage using Bloom filters. *BMC medical informatics and decision making* 9, 1 (2009).
- [138] Paul M Schwartz and Daniel J Solove. 2011. The PII problem: Privacy and a new concept of personally identifiable information. *NYUL rev.* 86 (2011).
- [139] Adil Hussain Seh, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. 2020. Healthcare data breaches: insights and implications. *Healthcare* 8, 2 (2020).
- [140] Chris J Skinner and Mark J Elliot. 2002. A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64, 4 (2002).
- [141] Daniel J Solove. 2023. Data is what data does: Regulating based on harm and risk instead of sensitive data. *Nw. UL Rev.* 118 (2023).
- [142] Alicia Solow-Niederman. 2022. Information privacy and the inference economy. *Nw. UL Rev.* 117 (2022).
- [143] Sebastian Stammmler, Tobias Kussel, Philipp Schoppmann, Florian Stampe, Galina Templer, Stefan Katzenbeisser, Kay Hamacher, and Martin Lablans. 2022. Mainzelliste SecureEpiLinker (MainSEL): privacy-preserving record linkage using secure multi-party computation. *Bioinformatics* 38, 6 (2022).
- [144] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* 12, 3 (2015).
- [145] Haochen Sun, Tonghe Bai, Jason Li, and Hongyang Zhang. 2024. zkDL: Efficient zero-knowledge proofs of deep learning training. *IEEE Transactions on Information Forensics and Security* (2024).
- [146] Zongkun Sun, Yinglong Wang, Minglei Shu, Ruixia Liu, and Huiqi Zhao. 2019. Differential privacy for data and model publishing of medical data. *IEEE Access* 7 (2019).
- [147] Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system. In *AMIA Annual Symposium*.
- [148] Latanya Sweeney. 1997. Guaranteeing anonymity when sharing medical data, the datafly system. In *AMIA Annual Symposium*.
- [149] Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25, 2-3 (1997).
- [150] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671, 2000 (2000).
- [151] Latanya Sweeney. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002).
- [152] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002).
- [153] Latanya Sweeney. 2015. Only you, your doctor, and many others may know. *Technology Science* 2015092903, 9 (2015).
- [154] Latanya Sweeney, Ji Su Yoo, Laura Perovich, Katherine E Boronow, Phil Brown, and Julia Green Brody. 2017. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. *Technology Science* 2017 (2017).
- [155] György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. 2007. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association* 14, 5 (2007).
- [156] Ricky K Taira, Alex AT Bui, and Hooshang Kangarloo. 2002. Identification of patient name references within medical documents using semantic selectional restrictions. In *AMIA Symposium*.
- [157] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. 2015. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software* 67 (2015).
- [158] United States. Congress. Office of Technology Assessment. 1993. Protecting privacy in computerized medical information.
- [159] United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1978. The Belmont report: ethical principles and guidelines for the protection of human subjects of research.
- [160] U.S. Department of Health and Human Services. 2000. The HIPAA Privacy Rule. 45 CFR Parts 160 and 164.
- [161] U.S. Department of Health and Human Services. 2024. Confidentiality of Substance Use Disorder Patient Records; Final Rule (42 CFR Part 2).
- [162] U.S. Department of Health and Human Services, National Institutes of Health. 2003. Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule. https://privacypolicyandresearch.nih.gov/pdf/HIPAA_Priva

- cy_Rule_Booklet.pdf. NIH Publication No. 03-5388.
- [163] Sirintra Vaiswri, Thilina Ranbaduge, and Peter Christen. 2018. Reference values based hardening for Bloom filters based privacy-preserving record linkage. In *Australasian Conference on Data Mining*.
 - [164] Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Language Resources and Evaluation Conference*.
 - [165] Petr Vcelak, Martin Kryl, Michal Kratochvil, and Jana Kleckova. 2019. Identification and classification of DICOM files with burned-in text content. *Journal of medical informatics* 126 (2019).
 - [166] Peter M Visscher and William G Hill. 2009. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS genetics* 5, 10 (2009).
 - [167] Antonia Vlahou, Dara Hallinan, Rolf Apweiler, Angel Argiles, Joachim Beige, Ariela Benigni, Rainer Bischoff, Peter C. Black, Franziska Boehm, Jocelyn Céraline, George P. Chrousos, Christian Delles, Pieter Evenepoel, Ivo Fridolin, Griet Glorieux, Alain J. van Gool, Isabel Heidegger, John P.A. Ioannidis, Joachim Jankowski, Vera Jankowski, Carmen Jeronimo, Ashish M. Kamat, Rosalinde Masereeuw, Gert Mayer, Harald Mischak, Alberto Ortiz, Giuseppe Remuzzi, Peter Rossing, Joost P. Schanstra, Bernd J. Schmitz-Dräger, Goce Spasovski, Jan A. Staessen, Dimitrios Stamatialis, Peter Stenvinkel, Christoph Wanner, Stephen B. Williams, Faiez Zannad, Carmine Zoccali, and Raymond Vanholder. 2021. Data sharing under the General Data Protection Regulation. *Hypertension* 77, 4 (2021).
 - [168] Sandra Wachter and Brent Mittelstadt. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.* (2019).
 - [169] Conor R Walker, Xiaoting Li, Manav Chakravarthy, William Lounsbey-Scaife, Yoelim A Choi, Ritambhara Singh, and Gamze Gürsoy. 2024. Private information leakage from single-cell count matrices. *Cell* 187, 23 (2024).
 - [170] Zhiyu Wan, James W Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley A Malin. 2022. Sociotechnical safeguards for genomic data privacy. *Nature Reviews Genetics* 23, 7 (2022).
 - [171] Zhiyu Wan, Yevgeniy Vorobeychik, Weiyei Xia, Ellen Wright Clayton, Murat Kantarcioglu, Ranjit Ganta, Raymond Heatherly, and Bradley A Malin. 2015. A game theoretic framework for analyzing re-identification risk. *PLoS one* 10, 3 (2015).
 - [172] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In *ACM Conference on Computer and Communications Security*.
 - [173] Shuang Wang, Xiaoqian Jiang, Siddharth Singh, Rebecca Marmor, Luca Bonomi, Dov Fox, Michelle Dow, and Lucila Ohno-Machado. 2017. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. *Annals of the New York Academy of Sciences* 1387, 1 (2017).
 - [174] Daniel Wartenberg and W Douglas Thompson. 2010. Privacy versus public health: the impact of current confidentiality rules. *American Journal of Public Health* 100, 3 (2010).
 - [175] Ruidi Wei and Florian Kerschbaum. 2023. Cryptographically secure private record linkage using locality-sensitive hashing. *Vldb Endowment* 17, 2 (2023).
 - [176] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* 45, 10 (2013).
 - [177] Ben Wellner, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leonid Peshkin, Alex Yeh, Janet Hitzeman, and Lynette Hirschman. 2007. Rapidly retargetable approaches to de-identification in medical records. *Journal of the American Medical Informatics Association* 14, 5 (2007).
 - [178] Weiyei Xia, Yongtai Liu, Zhiyu Wan, Yevgeniy Vorobeychik, Murat Kantarcioglu, Steve Nyemba, Ellen Wright Clayton, and Bradley A Malin. 2021. Enabling realistic health data re-identification risk assessment through adversarial modeling. *Journal of the American Medical Informatics Association* 28, 4 (2021).
 - [179] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. RadBERT: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence* 4, 4 (2022).
 - [180] Yongda Yu, Guoping Rong, Haifeng Shen, He Zhang, Dong Shao, Min Wang, Zhao Wei, Yong Xu, and Juhong Wang. 2024. Fine-tuning large language models to improve accuracy and comprehensibility of automated code review. *ACM Transactions on Software Engineering and Methodology* 34, 1 (2024).
 - [181] Eric Zeng, Xiaoyuan Wu, Emily N Ertmann, Lily Huang, Danielle F Johnson, Anusha T Mehendale, Brandon T Tang, Karolina Zhukoff, Michael Adjei-Poku, Lujio Bauer, et al. 2025. Measuring Risks to Users' Health Privacy Posed by Third-Party Web Tracking and Targeted Advertising. In *CHI Conference on Human Factors in Computing Systems*.
 - [182] Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng Wang. 2011. To release or not to release: evaluating information leaks in aggregate human-genome data. In *European Symposium on Research in Computer*

Security.

- [183] Yingxuan Zhu, PD Singh, Khan Siddiqui, and Michael Gillam. 2010. An automatic system to detect and extract texts in medical images for de-identification. In *Medical Imaging: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, Vol. 7628.
- [184] Markus Zimmermann, Cristian-Alexandru Staicu, Cam Tenny, and Michael Pradel. 2019. Small world with high risks: A study of security threats in the npm ecosystem. In *USENIX Security Symposium*.

A Interview Protocol

A condensed version of our interview protocol is below. We have shortened, removed, or collapsed probes, and we have also removed the recommended script for the interview with speech to transition between sections. For all questions, interviewers reworded questions to either express neutrality or build off of something the participant had said previously.

A. Personal background

- Can you describe your background, role at your institution and your responsibilities?

B. Understanding medical data and their use in research.

- Can you please describe your research at a high level?
- What kind of data do you work with?
- Where did you get the dataset?
- Do you collaborate with researchers from other institutions?

C. Privacy/threat modeling for their data.

- Do you consider your data sensitive or private? Why or why not? What about them do you consider sensitive or private?
- Who is in your dataset? What types of people? Where do the data come from?

D. Security for their data.

- Policy
 - What sort of restrictions (or protections) govern your use of your data?
 - How do these restrictions/policies/protections govern how you can / do share data with others?
 - How do these restrictions/policies/protections govern how you can analyze the data?
 - How do these restrictions/policies/protections govern how you publish/communicate about your work?
- Technical security
 - What technical security or privacy measures get used with your data?
 - Are there any other privacy-enhancing tools that you've used in your work? Why did you use them?
 - Is there anything that someone has recommended but you have not done? Why did you not do it?

E. Challenges and how to overcome them.

- Probe for possible scenarios; follow up using a scenario they have already mentioned:
 - Combining datasets from different sources and locating the same de-identified individuals?
 - Issues with releasing study data upon acceptance of a manuscript?
 - Not having enough data or having an insufficient level of detail?

- F. Participants' understanding/mental model of privacy measures.
- You mentioned that your data use [x]. What's your understanding of how that works?
 - What or who does it protect your data from?
 - What or who does it not protect your data from?
 - We've talked about [x] protections from [y]. Do you feel like this matches the sensitivity / privacy of the data you work with? Why or why not?
- G. Restrictions on research from S+P measures
- At a high level, to what extent do you feel like the restrictions on your dataset restrict your research?
 - Is there any type of research that you would like to do but it is currently not possible because of privacy restrictions?
 - Do privacy constraints benefit your research in any way? For example, is there anything you would not have been able to do without the privacy procedures in place?
 - How about the effect of the restrictions on your collaborations?
 - Are there any data that you would like to collect or access but cannot due to privacy or security or other restrictions?
- H. Ethics
- Given that you are working with sensitive and medical data, how do you make ethical decisions about privacy, security, and safety?
- I. Wrapping up
- Anything else you'd like to tell us?
 - Any questions for us?
 - Would you be willing to pass our information to others who might be willing to participate?

B Codebook

Theme	Codes
Participant context	Background, research goals
About the data	What types of data, data generation, data properties
Regulations	Regulatory bodies, Policies / processes, Practices, Opinions about regulation
Threat model	Trusted entities, Adversaries, Threats, Attack vectors, Consequences, Attitudes towards S&P, who is vulnerable, non-threats
Getting/giving access to data	Second-hand data provenance, getting practical access to data, requirements for access, data ownership vs rental
Data storage	privacy issues, location, requirements, storage format, who manages it, opinions
Data preparation	linking with another dataset, de-identification, practical de-identification strategies, challenges and needs
Analysis/computation	analysis types/goals, analysis code checks, analysis code development, workarounds, challenges, allowed to do, not allowed to do
Dissemination	rules: <i>do</i> publish, rules: <i>don't</i> publish, goals, tools

Table 2: Our codebook. Subcodes are not shown, as we used them as an analytic tool but did not build consensus on them.

C Recruitment messages

Our IRB-approved recruitment messages are below.

For someone we knew personally.

Subject: Study on privacy challenges of working with clinical data
Dear [name],

I am part of a team at [institution] seeking to interview medical researchers about their day-to-day work with sensitive clinical data. We would like to learn what privacy barriers medical researchers face, how they think about privacy in general, and what privacy technologies could facilitate their work.

Interviews can be in person or on Zoom; 60 minutes preferred but we are flexible. We are happy to offer a \$50 gift card of your choice as a token of our gratitude for your time and valuable insights.

Thanks,
[researcher]

To someone we do not know personally, or for participants to pass on.

Subject: Study on privacy challenges of working with clinical data

Dear [name],

I am part of a team at [institution] seeking to interview medical researchers about their day-to-day work with sensitive clinical data. We received your contact information from [another researcher who has participated in our study] or [your academic website]. We would like to learn what privacy barriers medical researchers face, how they think about privacy in general, and what privacy technologies could facilitate their work. We are interested in talking in particular to you because of your research experience with [type of clinical data].

Interviews can be in person or on Zoom; 60 minutes preferred but we are flexible. We are happy to offer a \$50 gift card of your choice as a token of our gratitude for your time and valuable insights.

Contact us at [email addresses].

Thanks,
[researcher]

To a mailing list.

Hi [title of mailing list/group],

We are a team of researchers from [institution] seeking to interview medical researchers about their day-to-day work with sensitive clinical data. We would like to learn what privacy barriers medical researchers face, how they think about privacy in general, and what privacy technologies could facilitate their work. The end goal is to identify what tools need to be built to make the lives of medical researchers easier.

Interviews can be in person or on Zoom; 60 minutes preferred but we are flexible. We are happy to offer a \$50 gift card of your choice as a token of our gratitude for your time and valuable insights.

Contact us at [email addresses].

Thanks,
[researcher]

In-person talking points.

- Seeking to interview medical researchers about their day-to-day work with sensitive clinical data.
- Learn about privacy barriers, the perception of privacy, and the needs.
- In-person or Zoom interviews; 60 minutes but we are flexible.