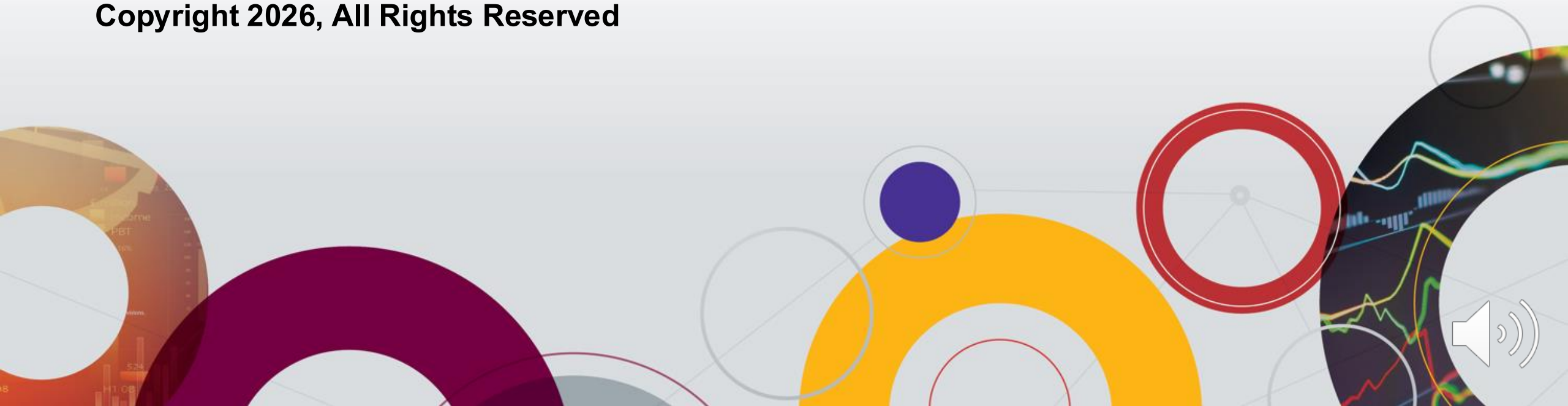


7.4a Privacy and Security

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Privacy and security

Definitions

Concerns

- Privacy
- Security

Tools for protecting health information

HIPAA

- Privacy Rule
- Security Rule



"On the Internet, nobody knows you're a dog."



"Remember when, on the Internet, nobody knew who you were?"

Definitions

Privacy – right to control personal information and keep it to yourself

Confidentiality – right to share personal things about you only to those who you designate

Security – mechanisms to protect your personal information

Individually identifiable health information (IIHI) – any data that can be correlated with an individual

Protected health information (PHI) – IIHI as defined by HIPAA Privacy Rule

Consent – (in context of privacy) written or verbal permission to allow use of your IIHI

HIPAA – Health Insurance Portability and Accountability Act

Concerns about privacy

Personal privacy vs. common good

Continued breaches and disclosures

Newer technologies exacerbate problem

De-identified data

Personal privacy vs. the common good

Concerns expressed in (famous) American Civil Liberties Union (ACLU) [video on ordering pizza](#)

There is a spectrum of views

- One end holds that while personal privacy is important, there are some instances when common good of society outweighs it, such as in biosurveillance (Gostin, 2002; Hodge, 1999)
- Other end holds that personal privacy trumps all other concerns is [Privacy Rights Clearinghouse](#)
- AMA extensive privacy principles (2020)
- Proper protections may encourage use of data (McGraw, 2021)
- Small risk of privacy disclosure, especially re-identification from de-identified data sets, is outweighed by benefit of new advances possible with data in precision medicine and machine learning (Seastedt, 2023)

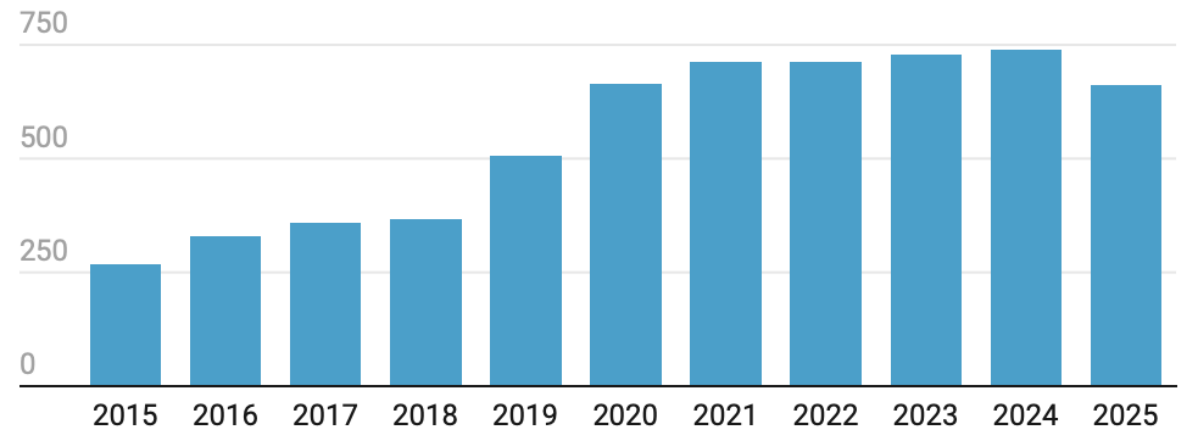
There are no right answers; where do your views fit?

Continuing information breaches and disclosures

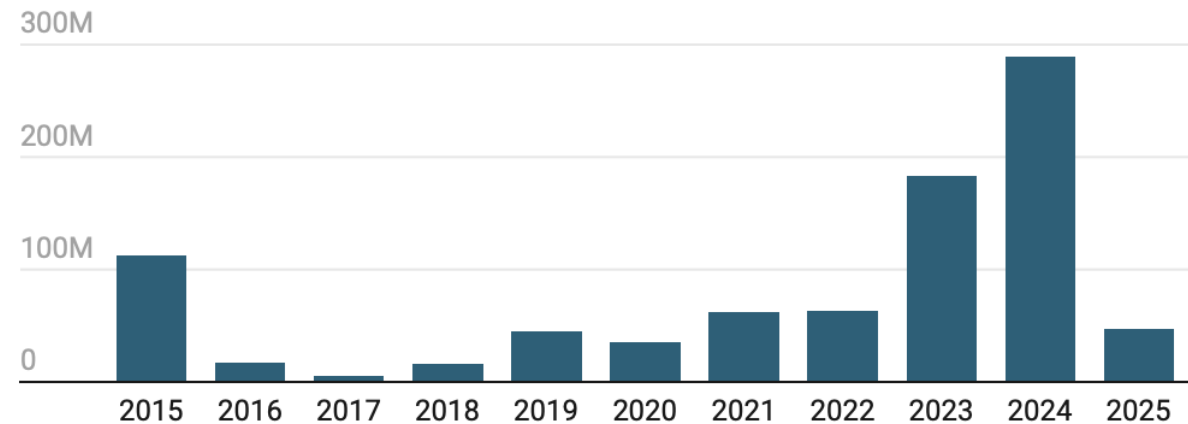
HHS required by HIPAA to post [list of breaches of unsecured PHI affecting >500 individuals](#) (“wall of shame”)

- Numbers of breaches persist to present, with large spike in individuals affected due to major breaches in 2023-2024 (Broderick, 2026), e.g.,
 - In 2024, Change Healthcare cyberattack with 276 million records representing 81% of the US population
 - OHSU with 4 entries

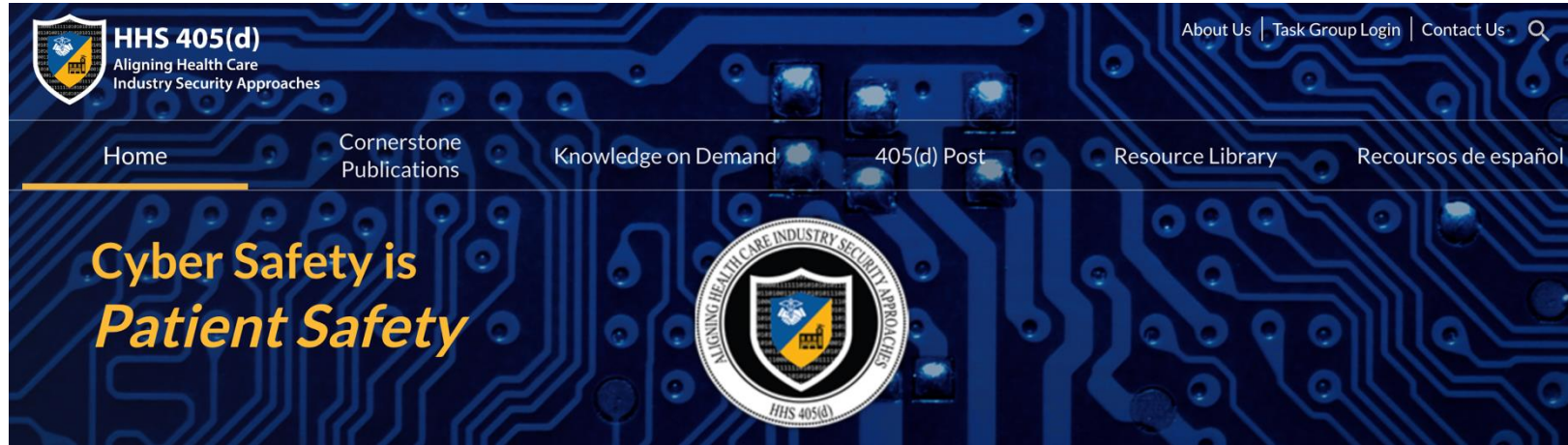
Number of breaches



Individuals affected



Top threats – HHS 405(d) program



Social Engineering



Ransomware



Loss or Theft of
Equipment or Data



Insider, Accidental or
Malicious Data Loss



Attacks Against
Network Connected
Medical Devices

Breaches and disclosures not limited to biomedicine and health

Privacy Rights Clearinghouse provides searchable [Chronology of Data Breaches](#) – not limited to healthcare

- Over 73K breaches affecting 81B records in searchable database

ACM Tech Brief
recommendations for data
privacy (Altman, 2024)

DATA PRIVACY PROTECTION: BY THE NUMBERS	
4.8	Billions of individuals worldwide whose personal information is for commercial sale. ¹
462	Projected 2031 total revenue of worldwide “data broker” market in billions of U.S. dollars. ²
156	Comparable revenues for that global market in 2012 in billions of U.S. dollars. ²
81	Percentage of Americans concerned about how companies use their data. ³
38.4	Percentage of 2010 U.S. Census responses that can be reidentified with high confidence. ⁴
77	Percentage of Americans concerned that they don’t understand what the government does with their data. ³
95	Percentage of individuals in a study of 1.5 million Europeans who could be uniquely characterized from just four random location records without using other personal data. ⁵
10,000	Minimum number of data types available for purchase from a major data broker. ^{1,6}
6.14	Millions of person-years estimated to be required annually to read all of the privacy policies for web services used and sites visited by U.S. internet users. ⁷
350	Number of consumer privacy bills considered by U.S. state legislatures in 2023. ⁸
3,519	Number of computer science papers on privacy posted to arXiv in 2023. ⁹

Breaches also costly in money

Analysis of 1461 PHI breaches in US over prior 10 years indicated 169M patients affected (Jiang, 2020)

- 71% affecting 159M patients (94%) contained sensitive demographic or financial information that could be exploited for identity or financial fraud
- 2% affecting 2.4 million patients (1%) contained sensitive medical information, potentially threatening clinical privacy

Cost of a breach to a healthcare organization averaged \$10.9M, twice next highest industry (IBM, 2023)

Notable privacy attacks

Ransomware attacks from 2016-2021 – 374 reported, exposing PHI of nearly 42 million patients. disrupting care and threatening patient outcomes and safety (Neprash, 2022)

Not limited to US; in Singapore

- 1.5 million records from SingHealth, including Prime Minister Lee Hsien Loong (Mahmud, 2018)
- Subsequent release of names of 14,200 patients with HIV (Griffiths, 2019)

23andMe confirmed hackers stole ancestry data on 6.9 million users (Franceschi-Bicchierai, 2023)

Another big breach affected MOVEit, a secure file transfer firm, with at least 60 million individuals affected, and many healthcare organizations (Page, 2023)

NextGen Healthcare reported hackers accessed personal data of more than 1 million patients (Page, 2023)

Notable privacy attacks (cont.)

Ransomware attack associated with disruptions at adjacent emergency departments (Dameff, 2023), including increased incidence of cardiac arrest (Pham, 2024)

Healthcare hit hardest by ransomware in 2023, FBI Internet crime report (McKeon, 2024)

Change Healthcare – largest healthcare breach ever (Berryman, 2025)

- Data of 190M Americans exposed
- Financial impacts on hospitals, pharmacies, and clinical practices (Hyperproof, 2024)
- Data likely still accessible on dark Web

Medical identity theft

A growing concern, emanating from general identity theft, defined as use of IHI for obtaining access to property or services, as noted by [OIG](#) and [FTC](#)

- Victims not only individuals but also healthcare providers and plans as well as society at large
- Physicians at risk as well (Agrawal, 2011)

Electronic health records (EHRs) deemed extremely valuable, e.g., 100 times more than credit card numbers (Akpan, 2016)

Challenges from proliferation of technologies and applications

Growing use of electronic data in clinical workflows

Health information exchange (HIE) moves data across networks

Cloud computing changes perimeter of data protection

New models of care (e.g., accountable care organizations) require more members of team to access information

Clinicians want to bring their own devices (e.g., personal laptops, smartphones, etc.)

Mobile devices send data continuously, on average about every 4.5 minutes (Leith, 2021)

Some technologies can worsen the problem

USB or thumb drives run programs when plugged into USB port; can be modified to extract data from computer (Wright, 2007)

10% of hard drives sold by a second-hand retailer in Canada had remnants of PHI (El Emam, 2007)

Medical devices

- Software on implanted devices (Burns, 2016; Kramer, 2017; FDA, 2018)
- Vendors of networked CPAP units, heart monitors, blood glucose meters, lifestyle apps, etc. may sell data to third parties (Kravitz, 2018)

Reach of Internet may exacerbate privacy problems

Tracking tools

- Widespread use on hospital Web sites, sending data to Google, Facebook, and other ad-serving sites (Feathers, 2022; Friedman, 2023)
- Medical journal sites (Gupta, 2022)

Mobile or other apps

- Review of 20 common medical apps found none followed all best security practices (Papageorgiou, 2018)
- Growing use of and concern over medical records (O'Brien, 2018)
- Many telehealth apps send personal data to tech companies (Palmer, 2022)

Selling de-identified EHR data that can potentially be re-identified using other data (Ross, 2022; Kim, 2023)

Offshore data beyond reach of US law enforcement but healthcare entities may be responsible for their business associates (Miliard, 2016)

Privacy issues also arise from changing laws

Dobbs v. Jackson Women's Health Organization case overturning Roe v. Wade allows some states to place legal restrictions on abortion available in others

- Not only banning abortion within states but from seeking or helping others obtain procedure, even in other states

Significant impact from informatics standpoint (Clayton, 2022; Khanna, 2022; Arvisais-Anhalt, 2022; Spector-Bagdady, 2022)

- Access to data may be eased by advances in informatics, e.g., data interoperability and health information exchange (HIE)
- HIPAA may not protect due to exceptions for law enforcement and other reporting allowed to governments

Ultimate “personal identifier” may be one’s genome

Even “de-identified” data may compromise privacy (Malin, 2004)

Genome of family members can identify siblings (Cassa, 2008)

Data from genome databases can reveal individual-level information (Lumley, 2010; Gymrek, 2013)

Misuse of consumer DNA testing products (e.g., Ancestry.com, 23AndMe) may compromise privacy of people or their relatives (May, 2018; Kim, 2018)

- Privacy laws apply ambiguously to such services (Hazel, 2018)

As breadth of data and uses grow, need sociotechnical approaches to protect privacy (Wan, 2022)

What to do when consumer sites go bankrupt, e.g., 23AndMe (Prince, 2025)

What about de-identified data?

Growing numbers of repositories of de-identified data used for research purposes, e.g., Vanderbilt Research Data Warehouse (Danciu, 2014) and Observational Health Data Sciences and Informatics (OHDSI) (Hripcsak, 2015)

Structured data easier to de-identify, though methods developed for text (Li, 2017; Heider, 2020)

Various methods can be used to ensure population databases do not allow identified of individuals, e.g., when cell numbers small (O'Keefe, 2018)

Can be implemented on very large scale (Radhakrishnan, 2023)

Best practices for protecting health data using de-identification and perturbation of structured datasets (Lulamba, 2025)

LLMs may be helpful in de-identification (Wiest, 2025)

De-identified data can sometimes be re-identified

87% of US population uniquely identified by five-digit zip code, gender, and date of birth (Sweeney, 2002)

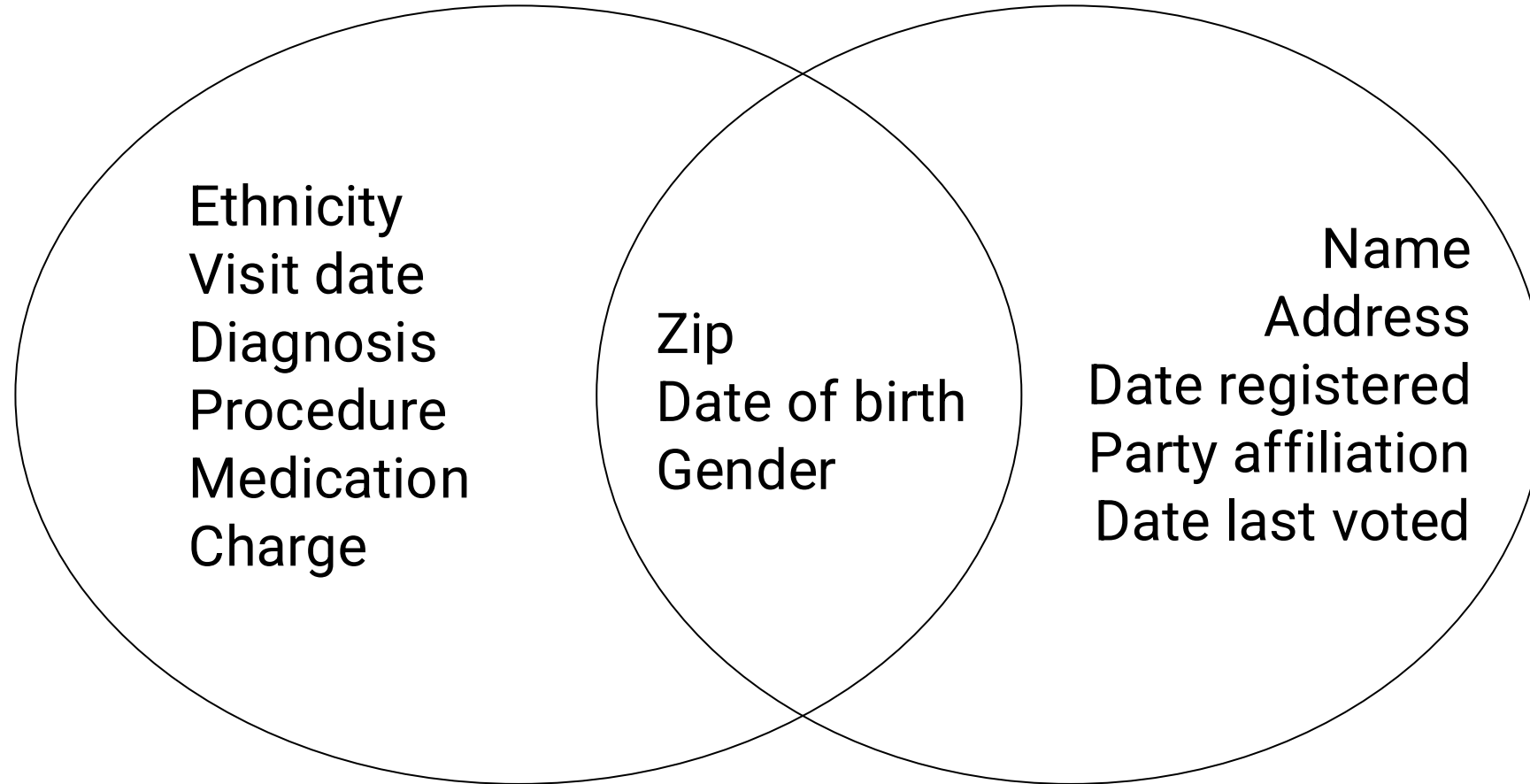
Computer scientists have long noted “broken promises” of data anonymization (Ohm, 2009)

One analysis identified Governor William Weld of Massachusetts in health insurance database for state employees by purchasing voter registration for Cambridge, MA for \$20 and linking zip code, gender, and date of birth to “de-identified” medical database (Sweeney, 1997)

Re-analysis found prominence of Gov. Weld aided in re-identification and current privacy laws markedly decrease risk of de-identification (Barth-Jones, 2012)

- But same voter registration data allowed re-identification of participants in (Massachusetts-based) Personal Genome Project (Sweeney, 2013)

How Governor Weld was re-identified



Other examples of re-identification of de-identified data

Genomic data can aid re-identification in clinical research studies (Malin, 2005; Lumley, 2010)

Laboratory orders and test results in combination may be unique (Cimino, 2012; Johnson, 2019)

Re-identification from other sources

- Public claims database in Washington State was used to re-identify patients through news reports of hospitalizations (Sweeney, 2013)
- Motor vehicle accidents in news can be linked with claims databases (Janmey, 2018)

Analysis of 210 public data sets found 99.9% of Americans likely to be re-identified with use of 15 demographic variables (Rocher, 2019)

De-identified EHR data can be re-identified with other data, e.g., claims data (Ross, 2022)

De-identified data not necessarily a panacea

Many ethical issues, from re-identification to harm caused by it (Rothstein, 2010)

If complete success is not achievable, how much accuracy is acceptable?

- Some PHI elements may be more acceptable than others, e.g., parts of street addresses vs. names (Prasser, 2016)

By same token, “over-scrubbing” may eliminate data important to analysis being done (Pantazos, 2017)

De-identified data may also (Meystre, 2010)

- Eliminate temporal information useful in longitudinal analysis
- Undermine accuracy of techniques such as machine learning and natural language processing

Public trust is essential to balance beneficial use vs. protection

Survey of US adults found patients amenable to secondary use of their EHR data; more toward research than quality or marketing uses (Grande, 2013)

Breast cancer patients expressed low level of concerns about privacy of data but do want it protected (Rogith, 2014)

Patients want more granular control of EHR data (Caine, 2013; Schwartz, 2015; Crotty, 2015) but physicians have concern over lack of complete access for providing care (Tierney, 2015)

- Pros and cons (Caine, 2015)

93% of patients willing to have data shared with academics and 82% with industry researchers (Mello, 2018)

From two academic medical centers, 87% willing to have their de-identified data shared and 56% willing to allow identified data to be shared for research, with somewhat lower proportions for those with mental health or abuse issues (Weng, 2019)

Balancing beneficial use vs. protection

Patients more willing to share data with local researchers and when well-designed consent forms used (Kim, 2019)

- Higher age and health literacy also associated with willingness to share

AMA survey found large majorities of patients wanted all users of their data to be held accountable and, for tech companies, to allow opting out (AMA, 2022)

50% of smart device owners willing to share and 32% would consider sharing smart device data for research purposes (Shandi, 2024)

Other methods may help

- Privacy-preserving record linkage (PPRL) (Kho, 2020; Lim, 2022; Marsolo, 2023)
- Differential privacy (Zheng, 2022; Altman, 2024)