

Portfolio Question 10.1.P - Opting Out of NHS Data

Daniel Lawson University of Bristol

Portfolio 10.1.P (v4.0.0)

Portfolio Question 10.1.P - Opting Out of NHS Data

- ▶ Should individuals be able to opt out of their health data being used for research that benefits everyone?
 - ▶ Contrast the **individual** right to privacy with the **collective** benefit of complete data, and consider whether technical safeguards can change the answer.

Why NHS data matters

- ▶ The NHS holds lifelong, near-complete health records for a whole population. Very few countries have anything comparable.
- ▶ During COVID-19, **OpenSAFELY** analysed the GP records of 17 million adults to identify risk factors for death¹.
 - ▶ Results arrived within weeks, and directly informed who was prioritised for shielding and vaccination.
 - ▶ Analysts **never saw the raw records**: code was sent to the data, and only aggregate outputs came back.
- ▶ Uses include: drug safety monitoring, planning hospital capacity, finding rare disease patterns, and training clinical prediction models.
- ▶ The value comes largely from **completeness**: everyone, not just those who volunteer.

¹Williamson et al. "Factors associated with COVID-19-related death using OpenSAFELY." Nature 584 (2020): 430-436.

A history of controversy

- ▶ **care.data (2013-2016)**: a plan to centralise GP records in England. A leaflet drop was widely missed, there was confusion over whether data could be sold to insurers, and the scheme was scrapped.
- ▶ **Royal Free and DeepMind (2017)**: a London hospital trust shared about 1.6 million patient records with Google DeepMind to develop a kidney-injury alert app.
 - ▶ The ICO ruled that the trust had **failed to comply with data protection law**: patients would not reasonably have expected their data to be used this way.
- ▶ **GDPR (2021)**: NHS Digital planned to extract GP records for around 55 million patients in England, with a six-week window to opt out.
 - ▶ After public backlash, **over a million** people opted out within weeks, and the scheme was paused indefinitely.
- ▶ The common thread: projects with real public benefit failed on **trust** and **transparency**, rather than on the science.

How opting out works

- ▶ In England there are two main mechanisms:
 - ▶ **Type 1 opt-out**: tells your GP practice not to share your record beyond your direct care.
 - ▶ **National Data Opt-out**: stops your confidential data being used for research and planning, with some exceptions (such as anonymised data and legal requirements).
- ▶ Recall from the lecture: GDPR treats health data as a **special category**. Consent is not the only lawful basis, though; public interest and medical research can also justify processing.
- ▶ So the law does **not require** consent here. Opt-outs are a policy choice, made to preserve public trust.

The statistical cost of opting out

- ▶ Opting out is not random. If opt-out rates differ by age, ethnicity, region, health condition or attitude to authority, the remaining data is a **biased sample**.
 - ▶ A rare-disease study loses power quickly if even a few patients leave.
 - ▶ A model trained on the remaining data may perform worse for the groups who opted out, which can be the groups who already distrust the system most.
- ▶ This creates ethical tension, “benefits to the group above benefits to the individual”:
 - ▶ Taken literally, is opting out **unethical**?
 - ▶ Or is the right to say no exactly what makes the system trustworthy enough for most people to stay in?

Can technology resolve the dilemma?

- ▶ The 2022 Goldacre Review (“Better, broader, safer”) recommended **Trusted Research Environments** (TREs) as the default for NHS data:
 - ▶ Data stays in one secure place; approved researchers run code against it; only checked, aggregate outputs leave.
 - ▶ All code and access is logged and auditable.
- ▶ This is often described with the **Five Safes**: safe people, safe projects, safe settings, safe data, safe outputs.
- ▶ But questions remain:
 - ▶ Who decides what counts as a “safe project”? What about commercial partners? In 2023 Palantir was awarded the contract for the NHS Federated Data Platform, which renewed the controversy.
 - ▶ Even aggregate outputs can leak information (see Part 2, Statistical Disclosure).

Discussion

- ▶ Should NHS data use be **opt-in**, **opt-out**, or **compulsory** (like the census)? Does the answer change for:
 - ▶ Academic research vs. commercial drug development vs. training AI models?
 - ▶ Identifiable vs. pseudonymised vs. aggregate data?
- ▶ If a TRE makes re-identification extremely unlikely, is there still a reason to let people opt out?
- ▶ Who should bear the cost when opt-outs bias the data, and how could a data scientist correct for it?
- ▶ What would it take for **you** to trust the NHS with your data?