BMJ Health & Care Informatics

# Enabling health data analyses across multiple private datasets with no information sharing using secure multiparty computation

Steven Kerr ,[1] Chris Robertson,[2,3] Cathie Sudlow,[1] Aziz Sheikh[4]

**To cite:** Kerr S, Robertson C, Sudlow C, *et al*. Enabling health data analyses across multiple private datasets with no information sharing using secure multiparty computation. *BMJ Health Care Inform* 2025;**32**:e101384. doi:10.1136/bmjhci-2024-101384

Received 27 November 2024
Accepted 15 May 2025



[1]The University of Edinburgh, Edinburgh, UK
[2]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK
[3]Public Health Scotland, Glasgow, UK
[4]Department of Primary Care Health Sciences, University of Oxford Nuffield, Oxford, UK

**Correspondence to**
Dr Steven Kerr;
steven.kerr@ed.ac.uk

## ABSTRACT

The UK's health datasets are among the most comprehensive and inclusive globally, enabling groundbreaking research during the COVID-19 pandemic. However, restrictions on data sharing between secure data environments (SDEs) imposed limitations on the ability to carry out joint analyses across multiple separate datasets. There are currently significant efforts underway to enable such analyses using methods such as federated analytics (FA) and virtual SDEs. FA involves distributed data analysis without sharing raw data but does require sharing summary statistics. Virtual SDEs in principle allow researchers to access data across multiple SDEs, but in practice, data transfers may be restricted by information governance concerns.

Secure multiparty computation (SMPC) is a cryptographic approach that allows multiple parties to perform joint analyses over private datasets with zero information sharing. SMPC may eliminate the need for data-sharing agreements and statistical disclosure control, offering a compelling alternative to FA and virtual SDEs. SMPC comes with a higher computational burden than traditional pooled analysis. However, efficient implementations of SMPC can enable a wide range of practical, secure analyses to be carried out.

This perspective reviews the strengths and limitations of FA, virtual SDEs and SMPC as approaches to joint analyses across SDEs. We argue that while efforts to implement FA and virtual SDEs are ongoing in the UK, SMPC remains underexplored. Given its unique advantages, we propose that SMPC deserves greater attention as a transformative solution for enabling secure, cross-SDE analyses of private health data.

## INTRODUCTION

The UK has some of the most rich, inclusive and large-scale health datasets in the world. Spurred on by the need to answer urgent questions in relation to the COVID-19 pandemic, these data assets were used to undertake UK-wide analyses on an unprecedented scale.[1–3] This culminated in the first-ever cohort study using routinely collected electronic health records of the entire UK population aged >5 years, in which the association between COVID-19 under-vaccination and severe COVID-19 outcomes (hospitalisation or death) was investigated.[4] Although these analyses have yielded important answers that have shaped scientific responses to the pandemic, restrictions on data sharing between secure data environments (SDEs) have resulted in limitations to the cross-SDE analyses that researchers have been able to undertake. For example, pooled analysis sharing minimal amounts of non-disclosive, aggregated information[1 2] or parallel analyses across SDEs followed by meta-analysis.[3 4]

In this perspective, we review the advantages and disadvantages of methods for enabling joint analyses across multiple private datasets and suggest secure multiparty computation (SMPC) as a promising new approach that can achieve this goal with no information sharing.

## CURRENT APPROACHES TO INDIVIDUAL-LEVEL ANALYSES ACROSS THE UK POPULATION

### Federated analytics (FA)

FA is a paradigm for analysing datasets owned by multiple parties without sharing raw data. FA is sometimes informally described as bringing the algorithm to the data rather than bringing the data to the algorithm. One way this can be implemented is for each party to calculate summary statistics on their data, which are then pooled. This process can be iterated to fit models and is the approach taken by the DataSHIELD software and the Sentinel Initiative.[5 6] The DARE UK TRE-FX driver programme aims to create infrastructure that will enable FA across UK SDEs.[7]

An advantage of FA is that raw data are never shared among parties. However, some data are still shared, typically in the form of aggregated summary statistics. This raises concerns that confidential information could

be disclosed if a sufficient number of summary statistics are shared. It may therefore be necessary to implement automated statistical disclosure control (SDC). However, the complex nature of data and information flows means that automating SDC is an extremely difficult problem to address. A further disadvantage of FA is that users may not have direct access to data held in multiple SDEs, so they cannot harmonise and clean data in ways that would typically be done if data from different sources were accessed in a common environment. It may be necessary for SDEs to implement a comprehensive common data model, which can take significant time and resources to set up and maintain.

### Virtual SDEs

In this approach, the parties can remotely access each other's data in a virtual environment that emulates an SDE containing all parties' data. Cybersecurity techniques are used to protect against unauthorised information disclosure. One way this can be achieved is to have the SDEs connected through a virtual private network (VPN). The DARE UK initiative has implemented this within the TELEPORT driver programme that has linked SAIL Databank in Wales and the Electronic Data Research and Innovation Service (eDRIS) Scottish national safe haven.[8]

A major advantage of this approach is that it can allow parties to access data from multiple SDEs in a way that is similar to using a single SDE. A disadvantage is that data ultimately must be transferred between SDEs, typically via the internet. This may limit what data SDEs can make accessible to each other. For example, if data governance agreements preclude full, raw datasets being sent between SDEs, then researchers may need to use alternative approaches to do pooled analysis. In this case, the main value that virtual SDEs could offer is allowing parties to access a subset of each other's data for initial exploration and formulation of data harmonisation and cleaning strategies.

## SMPC OFFERS A POTENTIAL WAY FORWARD

SMPC allows several parties to carry out a joint computation over private datasets with zero information sharing. There are many techniques that can be used for SMPC. For example, secret sharing allows a private value to be distributed across multiple parties, where no party on its own has information about the secret, but some threshold number of parties can together recover the secret. Secret sharing has the advantage of information-theoretic or perfect security, meaning the protocol is secure against adversaries that have an infinite amount of computational power. A separate class of techniques called garbled circuits uses encryption to carry out calculations securely. For introductions to SMPC, see Escudero (2024) and Evan *et al* (2018).[9 10]

SMPC has the major advantage that no information is shared. In particular, there are mathematically rigorous proofs that SMPC protocols are secure and do not leak any information. This eliminates the need to determine whether information could be disclosed and obviates the need for data-sharing agreements or SDC between SDEs. On the other hand, it can come with significant additional computational complexity. In particular, a large volume of non-disclosive communication can be required between parties. In secret sharing, this involves the parties sending random numbers to each other that can be combined to recover the secret. In garbled circuits, the parties send encrypted messages and keys to each other. However, the additional computational demand of SMPC may not present a significant barrier in a wide range of epidemiological studies. Like FA, SMPC does not allow parties to access each other's data, and so this approach shares similar problems with data harmonisation and cleaning.

## CONCLUSION

FA, virtual SDEs and SMPC all offer ways of securely carrying out joint analyses across SDEs, each with different advantages and disadvantages. Virtual SDEs can in principle allow all data to be shared and pooled. However, data must ultimately travel between SDEs, which poses security risks that may lead to restrictions on what data can be shared. FA only shares summary statistics between SDEs. SMPC shares no information, providing provable security. However, these latter two approaches do not allow parties to see each other's data in a way that is conducive to data harmonisation and cleaning.

While there are significant efforts underway to implement FA and virtual SDEs in the UK, SMPC has received relatively little attention. One reason for this is that SMPC may not be as well known and uses cryptographic techniques that are not widely understood. However, SMPC offers similar capabilities to FA but with the additional security guarantee that the parties do not learn anything about each other's data beyond the final result of the calculation. SMPC is also now sufficiently developed that it is practical to implement in health data analyses. For these reasons, we believe SMPC is promising and warrants greater attention as a solution for enabling pooled analysis across private health datasets.

**Contributors** This piece was conceptualised by SK who drafted the manuscript. AS, CR and CS commented on several drafts. SK is the guarantor.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**ORCID iD**
Steven Kerr http://orcid.org/0000-0002-3643-7859

## REFERENCES

1 Kerr S, Joy M, Torabi F, *et al*. First dose ChAdOx1 and BNT162b2 COVID-19 vaccinations and cerebral venous sinus thrombosis: A pooled self-controlled case series study of 11.6 million individuals in England, Scotland, and Wales. *PLoS Med* 2022;19:e1003927.
2 Kerr S, Bedston S, Bradley DT, *et al*. Waning of first- and second-dose ChAdOx1 and BNT162b2 COVID-19 vaccinations: a pooled target trial study of 12.9 million individuals in England, Northern Ireland, Scotland and Wales. *Int J Epidemiol* 2023;52:22–31.
3 Bedston S, Almaghrabi F, Patterson L, *et al*. Risk of severe COVID-19 outcomes after autumn 2022 COVID-19 booster vaccinations: a pooled analysis of national prospective cohort studies involving 7.4 million adults in England, Northern Ireland, Scotland and Wales. *Lancet Reg Health Eur* 2024;37:100816.
4 Kerr S, Bedston S, Cezard G, *et al*. 2024 Under-vaccination and severe COVID-19 outcomes: meta-analysis of national cohort studies of over 64 million people in England, Northern Ireland, Scotland and Wales. *The Lancet*.
5 Gaye A, Marcon Y, Isaeva J, *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014;43:1929–44.
6 Sentinel Initiative. Available: https://www.sentinelinitiative.org/
7 Giles T, Soiland-Reyes S, Coulridge J, *et al*. TRE-fx: deliver a federated network of trusted research environments to enable safe sata analytics. 2023 Available: https://zenodo.org/records/10055354
8 Orton C, Thompson S, Lee A, *et al*. TELEPORT: Connecting researcher to big data at light speed. 2023 Available: https://zenodo.org/records/10055358
9 Escudero D. An introduction to secret-sharing based secure multiparty computation. Cryptology ePrint Archive; 2024. Available: https://ia.cr/2022/062
10 Evans D, Kolesnikov V, Rosulek M. *A pragmatic introduction to secure multi-party computation*. Now Foundations and Trends.2018. Available: https://ieeexplore.ieee.org/document/8584398