[title] **Enabling discovery in trusted research environments securely with improved tooling**

[standfirst] **Biobanks are introducing Trusted Research Environments to enable secure data access, yet they can increase costs and complexity for biomedical researchers who use them. We discuss challenges, solutions and introduce the Phenofhy community Python tool for Our Future Health.**

Vincent J. Straub[1,*] and Melinda C. Mills[1,2,3*]

[1] Leverhulme Centre for Demographic Science, Nuffield Department of Population Health, University of Oxford and Nuffield College, Oxford, UK

[2] Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands

[3] Department of Economics, Econometrics and Finance, University of Groningen, Groningen, The Netherlands

*Correspondence to: vincent.straub@ndph.ox.ac.uk; melinda.mills@demography.ox.ac.uk

[1925 words]

Driven by the promise of precision medicine and the arrival of new technology, the last two decades have seen tremendous growth in data collection across phenotyping, genomic sequencing, proteomics, metabolomics, and imaging, resulting in vast troves of health data. Biobanks, the entities that collect, store, and manage biospecimens and health data--from national databases to bespoke disease-specific collections--have been at the forefront of this transformation[1].

Prominent examples include the UK Biobank (UKB), the US National Institutes of Health All of Us (AoU) and Our Future Health (OFH), a newer UK study of the UK adult population, aiming to recruit up to 5 million participants[2]. Other notable sources of biomedical data include electronic health records (EHRs), wearables, geolocation data, national registry data, and, more recently, trans-biobank cross-cohort analysis efforts[3]

As these developments continue[4], mechanisms that enable researchers to access de-identified, individual-level health data in an efficient, trusted, and cost-effective manner will be critical to the future success of global health research and genomic medicine[5]. One data access model that has quickly gained momentum and become a key node within biomedical knowledge infrastructures are Trusted Research Environments (TREs)[6]. These are

controlled systems and interfaces that enable users to remotely access and analyse sensitive data in a virtual environment via pay-for-what-you-use cloud computing[7], without ever needing to download the raw data.

TREs are increasingly being adopted by biobanks like OFH, yet they inadvertently create overlooked issues of costing and tooling for biomedical researchers in practice. We detail potential undesirable consequences of these issues for scientific research, and introduce Phenofhy, a user-friendly Python package for phenotypic analysis of OFH, along with practical recommendations to maximise TRE utility across biobanks.

The growth of trusted research environments

Originally created to meet security concerns and the needs of data controllers in fulfilling obligations to data owners, the TRE model has recently been adopted by a growing number of data providers because of the further benefits they promise. These are typically taken to include: decreased access barriers; lower storage costs; and greater collaboration opportunities[3,7].

In the UK context, OFH has been accessible to researchers via its OFH TRE since 2025; users have needed to access the UKB via its UKB Research Analysis Platform (UKB RAP) since the UKB switched to a TRE model in 2024. Other examples include the Genomics England Research Environment and the AoU Researcher Workbench in the US. At the time of writing, the UK's National Health Service (NHS) is also transitioning to a TRE approach.

Yet, a definitive picture of what a 'good' TRE looks like in terms of scientific utility is still missing. This partly reflects their nascency, complexity, and varied user-base (which includes researchers but also IT operators and information governance users), but also mirrors the technology-focused and rapidly-evolving nature of contemporary biomedical research[8].

Multiple consortia and research organisations have put forth initial specifications to standardise TREs, including Health Data Research UK (HDRUK) and Data and Analytics Research Environments (DARE); their efforts should rightly be commended for recognising the importance of ensuring TREs are usable to researchers[7]. Yet some nontrivial consequences of the turn towards TREs are arguably overlooked in current technical considerations, including an easily forgotten reality of biomedical discovery (especially when it comes to working with large, real-world data): it is inherently messier than the streamlined nature of TREs and cloud computing might suggest.

Growing pains of TREs and their impacts on research

From the perspective of a researcher-analyst, what arguably matters most in today's computational biomedical era, especially for early career researchers, are the mundane realities of working with data: getting started, exploring variables, playing around with the data, pre-processing, making mistakes, finding errors, testing pipelines, and configuring environments. Yet, this iterative, experimental nature of data analysis can be at odds with the reality of working in the current generation of TREs.

In practice, researchers are often faced with a number of constraints including significant time delays when launching interactive analyses; a lack of up-to-date tooling or inability to import custom software; a complex process to work out expected compute costs; and a sometimes substantial learning curve to use a new TRE comfortably and efficiently[6], especially for novice users, as each often comes with its own platform-specific nomenclature.

Figuring out cost estimates, for instance, can be a significant time burden for researchers and institutions alike, as currently users typically have to: (1) consult platform-specific rate cards; (2) estimate job runtimes and storage requirements; and (3) manually calculate costs across compute, storage, and data egress—all of which assume a pre-existing level of cloud computing expertise often missing from existing curricula[7]. Platforms have begun responding through community forums and updated pricing documentation, but no standardised cost estimation tooling exists, at times leaving researchers to develop *ad hoc* rules of thumb.

All of these issues are largely a feature and not a bug of the TRE model, reflecting the reality of cloud resource allocation and secure, governance-locked environments. And though seemingly minor in isolation, taken together they can hinder the kind of 'financially inefficient', exploratory, spur-of-the-moment, trial-and-error experimentation using personalised setups that is a key part of statistical training and at the heart of scientific research[9].

Crucially, such exploratory work can be essential for the detection of artifacts that inevitably exist within real-world datasets, such as EHRs, and, more consequentially, is often the precursor to serendipitous discoveries[10,11]. The design of digital interfaces may ultimately affect the potential of such discoveries[12]. Moreover, swift adoption of TREs that fails to fully consider these realities can sometimes lead to pushback from researchers, particularly in data-intensive areas like genetics and imaging[13].

Practical ways to improve TRE usability

We offer several general recommendations for platform providers and researchers to maximise the usability, scientific potential, and ease of getting started with TREs (**Box 1**). These build on our experience using OFH, alongside the UKB (which currently relies on the same TRE platform provider, DNAnexus), AoU, and other biobanks. Our recommendations are intended to be practical and relevant to biobank-hosted TREs in general, including OFH but also the UKB ahead of its RAP 2.0 development and other data providers transitioning to a TRE model, including those outside the UK.

We focus on three intersecting priorities: (1) enabling trial-and-error, exploratory discovery workflows, which can subsequently also reduce computing costs, (2) reducing the at-times steep learning curve in moving between common existing software practices and working in a TRE; and (3) improving data intelligibility and researcher-facing, cost-friendly infrastructure. Importantly, many data providers have already taken steps in a number of these directions; the UKB, for instance, pioneered providing students and those from lower income countries with financial support in the form of lower access fees and credit programmes, an approach that other TREs have also adopted.

A concern that cuts across all three priorities, and that we wish to highlight explicitly, is the shift to pay-as-you-go compute pricing. For research groups, this represents a fundamentally different cost model from the institutional computing arrangements many researchers have historically relied upon: costs that were once absorbed centrally must now be anticipated, itemised, and justified in advance. This can be particularly acute for PhD students and early-career researchers, for whom working with large, population-scale biobank data[2] is itself part of the learning process, and for whom the cost of exploration is by definition unpredictable.

Beyond training, anecdotal evidence suggests that pay-as-you-go pricing may increasingly be affecting the integrity of scientific review: peer reviewers asked to test newly published software, or independent groups seeking to replicate high-profile findings, face real financial disincentives to do so within a TRE. As complex, data-hungry methods, including complex, agentic AI[9] approaches (difficult to evaluate without running on real data) become more prevalent in biomedical TRE-based research, this problem might worsen.

Accordingly, we recommend that platform providers explore the introduction of subsidised or zero-cost compute allocations designated for peer review and independent replication (**Box 1**). In practice, this may require creating a new kind of ‘replication environment’ in a TRE, in which researchers have time-limited access to a specific dataset and model linked to a particular publication, or ensuring researchers are able to share their TRE project with a TRE-registered peer reviewer for independent inspection. Importantly, these are initial

suggestions; our aim here is not to not offer the ultimate solution, rather we contend that these issues need explicit acknowledgement and discussion within the community.

Introducing Phenofhy for processing OFH phenotypes

If we are to make health data more collectively accessible, useable, and shareable[5], while also preserving privacy, security, and trust, increased attention arguably needs to be paid towards researcher usability in the context of TREs, especially when it comes to costing and tooling. To this end, stemming from our research using the OFH TRE as part of their Early Adopters Programme[2], we have developed the community tool Phenofhy, an open-source Python package available to all approved OFH researchers via the TRE (researchers can apply for access to OFH and the OFH by first becoming 'registered researchers' at https://research.ourfuturehealth.org.uk/apply-to-access-the-data/).

Phenofhy is designed to process OFH phenotypic data in an efficient and user-friendly manner by offering phenotypic preprocessing and quality control. The package is currently in active development and is being beta released to invite input from the research community to enable improvements as the number of studies and researchers using OFH continues to grow.

Phenofhy aims to sit alongside other recent TRE-focused packages, specifically, the allofus R package[14] for the AoU Researcher Workbench. Both these tools aim to reduce the technical burden, programming knowledge, and compute costs required to get started with conducting research using TREs. Inspired by the allofus package, we built Phenofhy with a number of design principles in mind that aim to maximise TRE usability (**Fig. 1**), which we hope future TRE tooling options and those under development can take inspiration from.

For Phenofhy, our specific goals were to create a tool that would: (1) simplify the extraction and preprocessing of phenotypic variables from OFH datasets; (2) facilitate Quality Control through automated phenotype profiling prior to downstream analyses; (3) support local simulation of OFH-structured data so that researchers can prototype and test analysis pipelines without needing any data access and before incurring cloud compute costs; and (4) lower barriers to entry for early-career and other researchers engaging with TRE-based research for the first time. The package is designed as a complement to existing OFH tooling rather than a replacement, and its scope is expected to grow in tandem with the expanding OFH researcher community.

The Phenofhy package is currently distributed as a beta release via GitHub, where it can be downloaded and imported into a researcher's TRE project space via the OFH Airlock process. Full documentation, a quick start guide, key concepts, tutorials, and an API reference are available, along with all source code. OFH study data are available to approved researchers via the OFH TRE and an embedded community discussion thread provides a channel for feature requests, bug reports, and user contributions.

Together, Phenofhy and our recommendations (**Box 1**) aim to support reproducible, secure research and biomedical discovery; we now invite data providers and platforms to work with researchers to build on this momentum to ensure the TRE model continues to serve both the security needs of data owners and the practical demands of modern biomedical research.

Competing interests

V.J.S. is a research scholar for OFH. M.C.M. is a research ambassador for OFH, a trustee and on the ethics advisory board of the UK Biobank, previously on the scientific and currently on the ethics advisory boards of OFH, and on the advisory boards of the Netherlands Lifelines Biobank, US Health and Retirement Survey and UK CLS Cohort Studies.

References

1. Gallagher, C.S., Ginsburg, G.S. & Musick, A. *Nat. Rev. Genet.* **26**, 191–202 (2025).
2. Straub, V.J., Benonisdottir, S., Kong, A. *et al. Nat. Genet.* **57**, 2341–2348 (2025).
3. Deflaux, N., Selvaraj, M.S., Condon, H.R. et al. *Nat. Commun.* 14, 5419 (2023).
4. D'Altri, T., Freeberg, M.A., Curwin, A.J. et al. *Nat. Genet.* 57, 481–485 (2025).
5. Stark, Z., Glazer, D., Hofmann, O. *et al. Nat. Rev. Genet.* **26**, 141–147 (2025).
6. O'Donovan, C., Coleman, S., Kerr, D., Cole, C., Li, S., Sarmiento, D., & Sood, H. (2023). DARE UK (Data and Analytics Research Environments UK). https://doi.org/10.5281/zenodo.10066800.
7. Langmead, B., Nellore, A. *Nat. Rev. Genet.* **19**, 208–219 (2018).
8. Shiffrin, R.M., Börner, K. & Stigle., S. M. *Proc. Natl. Acad. Sci. U.S.A.* **115** (11) 2632-2639. (2018).
9. Li, B., Saini, A.K., Hernandez, J.G. *et al. Nat. Biotechnol.* (2026).
10. Hargrave-Thomas, E., Yu, B. & Reynisson, J. *World J. Clin. Oncol.* **3**, 1–6 (2012).
11. Ross, W., Copeland, S. & Firestein, S. *J. Trial Error* (2024). https://doi.org/10.36850/v91j-7541
12. McMurray, C. *The Transmitter.* (2024). https://doi.org/10.53053/WOSF6165.

13. Beaulieu, A. Interfaces for New Relations. in *Revealing Relations* 129–149 (Bristol University Press, 2026).
14. Smith, L. H. & Cavanaugh, R. *J. Am. Med. Inform. Assoc.* **31**, 3013–3021 (2024).

Box 1 | Recommendations for improving usability of Trusted Research Environments (TREs) for exploratory, cost effective and reproducible science

1. Recommendations for data controllers and platform providers

Enable exploratory research workflows

- Provide high-fidelity synthetic datasets or sandbox training environments that mirror real data structures, enabling local prototyping before large-scale cloud execution
- Offer predictable cost structures via an interactive cost estimator that provides guidance on the costs of running various types of analyses (e.g., GWAS)

Support researcher onboarding and scientific reproducibility

- Provide onboarding documentation tailored to researchers migrating from other major TRE platforms, mapping equivalent concepts, tools and terminology across environments to reduce platform-specific learning overhead
- Introduce subsidised or zero-cost compute allocations specifically designated for peer review and independent replication, recognising that pay-as-you-go pricing currently creates a financial barrier to verifying published results—a problem that will intensify as complex AI methods become more prevalent in biobank research.

Prioritise data intelligibility and researcher-facing infrastructure

- Develop searchable metadata accessible outside the TRE that enable researchers to understand variable definitions and relationships across all available data sources (e.g., EHRs, clinical measurements, surveys) before initiating analyses, to increase usability, feasibility, and discovery, but also reduce costs.
- Provide "run-it-yourself" examples of quality control (QC) pipelines implemented on any released data
- Establish and actively moderate community forums (as implemented by UKB Community Forum) to support peer-to-peer troubleshooting and surface recurrent usability barriers.

2. Best practices for researchers

Develop platform and governance literacy

- Understand the computational, financial and governance constraints of the TRE before scaling analyses
- Use available guidance to reduce avoidable inefficiencies; for convenience, we provide a curated list of practical resources for getting up to speed with DNAnexus-hosted TREs used by the UKB and OFH here.

Design analyses iteratively and share reusable artefacts

- Prototype locally or on synthetic data where appropriate before large-scale execution
- Share reusable code, QC pipelines, phenotype definitions and workflow templates through community channels where individual data is not shared and governance allows it

Advocate collectively

- Provide structured feedback to data providers and call for the tools and infrastructure needed to conduct high-quality science. As TREs become a dominant access model for sensitive health data, researchers have an important role in shaping systems that maximise scientific utility while maintaining security

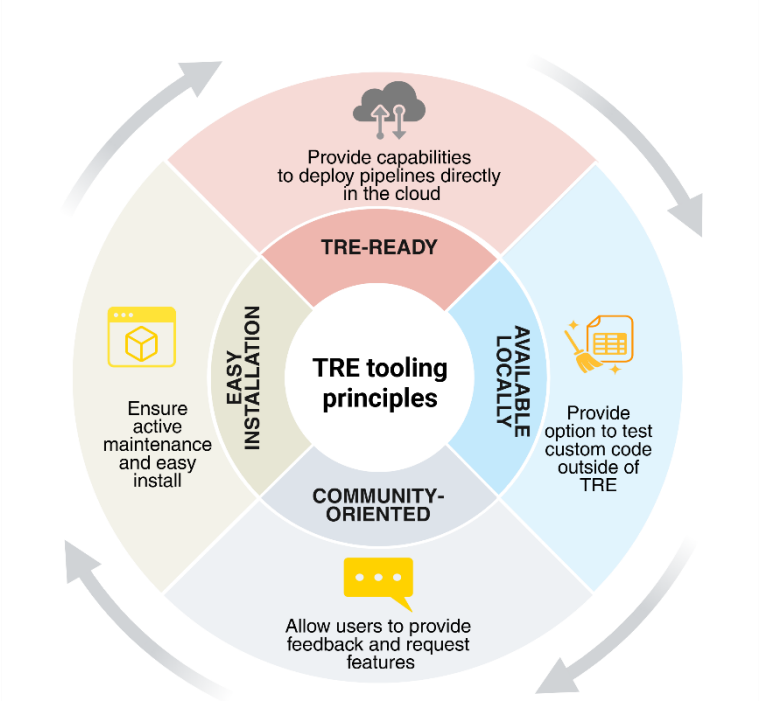

Fig. 1: Principles for designing TRE tooling. Phenofhy is designed to abide by four design principles, which other tooling options can adopt to maximise usability of TREs. These include (1) being built to be TRE-ready, that is, easily deployable directly in the cloud; (2) available locally for testing without compute costs; (3) community-oriented to enable user feedback; and (4) easily installable, versioned, and maintained. TRE, trusted research environment.

Acknowledgements

V.J.S. and M.C.M. are supported by ESSGN (HORIZON-MSCA-DN-2021 [101073237]). M.C.M. is supported by ESRC/UKRI Connecting Generations (ES/W002116/1), an ERC advanced grant (835079), and the Einstein Foundation Berlin (EZ-2019-555-2). Both are supported by the Leverhulme Trust Large Centre Grant LCDS (RC-2018-003). The development of Phenofhy required access to restricted OFH data, approved under the OFH study ID 240228. Ethical approval for OFH was granted by the Cambridge East Research Ethics Committee (REC reference: 21/EE/0016). We would like to acknowledge all of the research participants who have donated their data to the OFH research program. We thank Stefania Benonisdottir and Augustine Kong for useful comments on an initial draft.

Author information

Authors and affiliations

Leverhulme Centre for Demographic Science, Nuffield Department of Population Health, University of Oxford and Nuffield College, Oxford, UK

Vincent J. Straub & Melinda C. Mills

Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands

Melinda C. Mills

Department of Economics, Econometrics and Finance, University of Groningen, Groningen, The Netherlands

Melinda C. Mills

Author ORCID IDs

Vincent J. Straub, https://orcid.org/0000-0003-3393-6027

Melinda C. Mills, https://orcid.org/0000-0003-1704-0001

Corresponding authors

Vincent J. Straub

Melinda C. Mills