Tsipouras *et al. Rare Dis Orphan Drugs J* 2023;2:17
DOI: 10.20517/rdodj.2023.15

Rare Disease and Orphan Drugs Journal

Opinion

Open Access

Check for updates

Could federated data analysis be the catalyst accelerating the introduction of newborn genome screening for the detection of genetic disease?

Petros Tsipouras[1], Maria Chatzou Dunford[2], Hadley Sheppard[2], Hannah Gaimster[2], Theoklis Zaoutis[3,4]

[1]FirstSteps Greece, Newborn genome screening Initiative, Athens 106 80, Greece.
[2]Lifebit Biotech Ltd., London EC2A 2AP, UK.
[3]National Public Health Organization (EODY), Athens 151 23, Greece.
[4]The 2nd Department of Pediatrics, National and Kapodistrian University of Athens, 'P. & A. Kyriakou' Children's Hospital, Athens 106 80, Greece.

Correspondence to: Dr. Petros Tsipouras, FirstSteps Greece, Newborn genome screening Initiative, Skoufa 64, Athens 106 80, Greece. E-mail: petros.tsipouras@beginnings.gr

How to cite this article: Tsipouras P, Chatzou Dunford M, Sheppard H, Gaimster H, Zaoutis T. Could federated data analysis be the catalyst accelerating the introduction of newborn genome screening for the detection of genetic disease? *Rare Dis Orphan Drugs J* 2023;2:17. https://dx.doi.org/10.20517/rdodj.2023.15

Received: 14 Jun 2023 **First Decision:** 7 Sep 2023 **Revised:** 14 Sep 2023 **Accepted:** 22 Sep 2023 **Published:** 27 Sep 2023

Academic Editor: Virginie Bros-Facer **Copy Editor:** Dan Zhang **Production Editor:** Dan Zhang

Abstract

Data federation intermediated through trusted research environments can help accelerate the adoption and utilization of newborn genome screening worldwide. Data federation will protect individual datasets from unauthorized security breaches, allow analysis *in situ*, and bypass the need for cumbersome data sharing agreements between parties. Finally, data federation could accelerate the adoption of new therapies for rare genetic diseases with the use of synthetic clinical trials.

Keywords: Newborn genome screening, data federation, trusted research environment

INTRODUCTION

Worldwide, millions of children are born with a rare genetic disease[1,2]. Newborn screening (NBS) has been effective in identifying babies who are at risk of developing a genetic disease and initiating a therapeutic intervention. The first genetic disease for which NBS was introduced is phenylketonuria (PKU), where early

dietary intervention prevents serious mental deficiency[3]. In the past fifty years, mandated NBS has expanded to include other, mostly Mendelian, diseases where early therapeutic intervention has been effective in preventing and/or ameliorating irreversible tissue damage. In many countries, states, and regions of the world, public health programs are in place to collect blood specimens from babies soon after birth[4]. Analytes extracted from dried blood spots collected on filter paper are assayed using gas chromatography/mass spectrometry (GC/MS) or tandem MS.

A second layer of screening based on Next-Generation Sequencing (NGS) technology could expand the scope of the existing NBS programs[5,6]. This additional layer of screening will not replace what is currently used, but it will increase the current offering substantially to include a broader spectrum of disorders not detectable by tandem MS.

Newborn genome sequencing could evolve to become the new paradigm for healthcare delivery, where early detection could result in better clinical outcomes. Rapid Whole Genome Sequencing (rWGS) has been shown to be an effective diagnostic test linked to decreased infant mortality and improved outcomes in babies admitted to Neonatal Intensive Care Units (NICU)[7,8].

Extending the use of genome sequencing as a screening test to all newborns is only a matter of time. However, before newborn genome screening is widely adopted, several factors will need to be carefully considered, including:

1. Accurate definition of pathogenic genomic variants in diverse populations.

2. Defined care paths for the follow-up of a screen-positive finding.

3. Evidence that early intervention leads to improved clinical outcomes.

4. Detailed cost analysis.

Persuasive answers to the above will be required by the key stakeholders whose support is essential, i.e., parents, health care providers, public health policymakers, and the pharmaceutical industry.

Several newborn genome screening (including whole genome sequencing and whole exome sequencing) initiatives have been launched, or they will be launched soon[9-11].

We anticipate that no one project will have the necessary solutions to satisfactorily address all or some of the above-mentioned problems. Thus, aggregation of information collected from different sources could provide part of the solution for critical mass and momentum.

Data aggregation of such magnitude presents significant legal, ethical, and technical challenges related to (i) the security and privacy of sensitive information; (ii) the size and varied nature of stored genomic data; and (iii) legal requirements for data sharing. A viable near- and mid-term solution that can help address these issues will be using trusted research environments (TREs) and data federation for secure storage, access, and analysis of genomic data[12]. A comparison of risks and benefits between existing and federated databases for genomic data is shown in Table 1.

Table 1. The data aggregation challenge. Comparison of risks and benefits between existing and federated databases

	Databases	Federated databases
Security and compliance	Movement and copying of sensitive information increases the risk of data breach	In a TRE and federation environment, data are not moved or copied, reducing security risk
Data size and interoperability	Lack of standardized formats and pipelines limits interoperability, and negatively impacts scalability, cost, and efficiency	Fully standardized data, securely accessible by cloud-based platforms through federation, can be combined with global cohorts and disparate datasets
Collaboration	Data cannot leave jurisdictional borders. Data sharing agreements are frequently difficult to negotiate and implement, hindering collaboration	Federated approaches will eliminate a major barrier across individual datasets, vastly improving the statistical power of research

TRE: trusted research environment.

Federated data analysis platforms, which facilitate secure data access from multiple sources without the need for data movement- where data could be vulnerable to interception, have emerged as a promising part of a solution for safely sharing anonymized genomic data. Here, genomic data remains secure in the TRE, which can then be linked virtually using a set of Application Programming Interfaces (APIs).

Traditional data access methods involve researchers downloading data to an institutional computing cluster. With federated analysis, the analysis is brought to where the distributed data lies, thereby eliminating the risky movement of data and removing many existing barriers to accessibility[13]. Such technology means that data can be made securely accessible but that data controllers (e.g., biobanks and healthcare providers) retain jurisdictional autonomy over data, a key concern in international data sharing.

International initiatives such as the Global Alliance for Genomics and Health (GA4GH)[14] set standards to promote the international sharing of genomic and health-related data, in part by setting interoperability standards and providing open-source APIs.

Common Data Models (CDMs) are crucial to ensuring data is interoperable, with several growing in popularity in the life sciences sector recently, including OMOP (Observational Medical Outcomes Partnership) CDM from the OHDSI (Observational Health Data Sciences and Informatics)-specifically for clinical-genomic data. Examples of health organizations utilizing OMOP as their CDM include the UK Biobank and All of Us from the US National Institutes for Health (NIH)[15,16].

Additionally, extraction, transformation, and loading (ETL) pipelines that can automate this work to process and convert raw data to analysis-ready data help further simplify this process for researchers. Normalizing all data to internationally recognized standards allows researchers to perform joint analyses across distributed datasets, which is key to ensuring diversity and representation of as many populations as possible in studies.

These standardized and interoperable datasets could be combined seamlessly for analysis via federation, enabling researchers to analyze this data collaboratively in conjunction with other complementary datasets. Standardization of data formats and analytical approaches within and even between health systems can bring substantial benefits in terms of comparability of data and contribute to continually improving processes.

Illustrative examples with potential multiplier effects could include:

Sharing pathogenic variants: Defining the frequency and prevalence of a pathogenic variant in diverse populations is essential. Access to the pathogenic variant libraries of the various initiatives will impact the predictive value of a screen positive, and it might help in the reclassification of Variants of Uncertain Significance (VUS).

Sharing care paths: Newborn genome screening is a risk stratification test that places a person in a high- or low-risk group for a particular genetic disease. The accuracy and validity of establishing the presence of disease have an enormous impact on the well-being of the person and the family, the timing of therapeutic intervention, possibly the modality of intervention, and ultimately healthcare cost.

Sharing clinical outcomes: Managing individuals with latent or early-stage disease can potentially increase the burden on health care providers and the health care system. Therapeutic interventions, to the extent possible, will need to be evidence-based. Individual genetic illnesses are often uncommon, and randomized clinical studies are difficult to conduct. Sharing clinical outcomes, on the other hand, may provide an incentive for synthetic clinical studies.

Sharing the analytic modality used to generate a variant: The validity of a variant is frequently linked to the analytical platform used to generate the information, i.e., panel, short-read NGS, and long-read NGS. Providing a barcode record could be helpful in assessing a variant and its possible value as a biomarker.

Recently, a pioneering example of a multi-party federation between Genomics England and Cambridge Biomedical Research Centre (BRC) was demonstrated. This allowed secure data analysis across TREs in the UK's first known demonstration of genomic data federation. This highlights that the technology now facilitates secure data access via federation for authorized researchers to perform joint secure data analysis on global cohorts. Data sharing via federated databases also decreases the danger of unauthorized access and encourages the adoption of advanced privacy-preserving encryption methods when analyzing data[17]. It is conceivable to imagine that this technology could be used in an undiagnosed disease program targeting newborns to help the integration of clinical, genomic, therapeutic, and outcome inputs residing in different datasets. A summary of how this could work is provided in Figure 1.

The Federated European Genome-Phenome Archive (EGA) is another program that uses federation to provide global discovery and access to human data for research while still adhering to jurisdictional data protection rules. The Federated EGA promotes data reuse, facilitates reproducibility, and accelerates biomedical research by providing a solution to increasing issues in the safe and efficient handling of human omics and related data[18].

While there are significant advantages to moving towards a genomic approach to newborn screening, these programs also have challenges. These include considerations of the ethical, legal, and social implications (ELSI) of newborn genomic screening - these can include concerns surrounding sensitive data sharing, patient autonomy, and consent. A detailed discussion of these issues is out of the scope of this piece, but we refer the reader to other references that have discussed these issues more completely[19,20].

As federation is an emerging technology, careful consideration must be given to scaling up federation across different TREs, particularly surrounding governance and assurances in particular across different jurisdictions. Additionally, genomic data federation could potentially have risks, which may include improper use of data, hacking, and identification of incidental findings such as detection of variants associated with pathologies not immediately treatable or relevant to the newborn[21]. It is important that

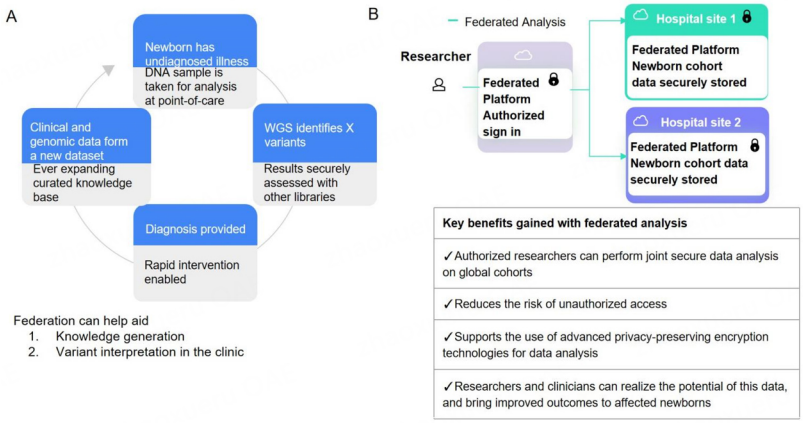

Figure 1. An overview of how federated data analysis can be incorporated into an undiagnosed disease program targeting newborns to help enable secure data access across research laboratories and clinics worldwide. (A) The steps involved in diagnosing a rare disease in an affected newborn; (B) A summary of how federated data analysis is performed and the benefits that can be gained.

these all be considered and addressed as federated approaches continue to be developed.

CONCLUSION

Newborn genome screening is a promising approach to early disease detection with considerable advantages compared to traditional approaches, but the integration into clinical care comes with complex technical challenges, which must be meaningfully explored to ensure effective and equitable impact. Standardized data federation could provide part of a crucial solution as a collaboration framework for the various newborn genome screening initiatives underway worldwide. Such efforts to facilitate secure joint data access and analysis to information among relevant stakeholders will accelerate the existing momentum of collaboration between global newborn sequencing initiatives, ultimately improving outcomes for patients.

DECLARATIONS

Authors' contributions

Wrote the paper: Tsipouras P, Sheppard H, Gaimster H
Reviewed the paper: Chatzou Dunford M, Zaoutis T

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Hageman IC, van Rooij IALM, de Blaauw I, Trajanovska M, King SK. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. *Orphanet J Rare Dis* 2023;18:106. DOI PubMed PMC
2. Pogue RE, Cavalcanti DP, Shanker S, et al. Rare genetic diseases: update on diagnosis, treatment and online resources. *Drug Discov Today* 2018;23:187-95. DOI
3. Guthrie R. Blood screening for phenylketonuria. *JAMA* 1961;178:863. DOI
4. Fidan Ç, Örün H, Alper AB, et al. Expanded newborn bloodspot screening: developed country examples and what can be done in Turkey. *Intractable Rare Dis Res* 2022;11:63-9. DOI PubMed PMC
5. Bick D, Ahmed A, Deen D, et al. Newborn screening by genomic sequencing: opportunities and challenges. *Int J Neonatal Screen* 2022;8:40. DOI PubMed PMC
6. Ceyhan-Birsoy O, Machini K, Lebo MS, et al. A curated gene list for reporting results of newborn genomic sequencing. *Genet Med* 2017;19:809-18. DOI PubMed PMC
7. Kingsmore SF, Smith LD, Kunard CM, et al. A genome sequencing system for universal newborn screening, diagnosis, and precision medicine for severe genetic diseases. *Am J Hum Genet* 2022;109:1605-19. DOI
8. Kingsmore SF; BeginNGS Consortium. Dispatches from biotech beginning BeginNGS: rapid newborn genome sequencing to end the diagnostic and therapeutic odyssey. *Am J Med Genet C Semin Med Genet* 2022;190:243-56. DOI PubMed
9. Gaff CL, M Winship I, M Forrest S, et al. Preparing for genomic medicine: a real world demonstration of health system change. *NPJ Genom Med* 2017;2:16. DOI PubMed PMC
10. Holm IA, Agrawal PB, Ceyhan-Birsoy O, et al; BabySeq Project Team. The BabySeq project: implementing genomic sequencing in newborns. *BMC Pediatr* 2018;18:225. DOI PubMed PMC
11. Pichini A, Ahmed A, Patch C, et al. Developing a national newborn genomes program: an approach driven by ethics, engagement and co-design. *Front Genet* 2022;13:866168. DOI PubMed PMC
12. Alvarellos M, Sheppard HE, Knarston I, et al. Democratizing clinical-genomic data: how federated platforms can promote benefits sharing in genomics. *Front Genet* 2022;13:1045450. DOI PubMed PMC
13. Chaterji S, Koo J, Li N, Meyer F, Grama A, Bagchi S. Federation in genomics pipelines: techniques and challenges. *Brief Bioinform* 2019;20:235-44. DOI PubMed PMC
14. Rehm HL, Page AJH, Smith L, et al. GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genom* 2021;1:100029. DOI PubMed PMC
15. Papez V, Moinat M, Voss EA, et al. Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. *J Am Med Inform Assoc* 2022;30:103-11. DOI PubMed PMC
16. Mayo KR, Basford MA, Carroll RJ, et al. The all of Us data and research center: creating a secure, scalable, and sustainable ecosystem for biomedical research. *Annu Rev Biomed Data Sci* 2023;6:443-64. DOI
17. Nik-Zainal S, Seeger T, Fennessy R, et al. Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets. *Zenodo* ;2022:online ahead of print. DOI
18. Rueda M, Ariosa R, Moldes M, Rambla J. Beacon v2 reference implementation: a toolkit to enable federated sharing of genomic and phenotypic data. *Bioinformatics* 2022;38:4656-7. DOI PubMed
19. Hunter A, Lewis C, Hill M, et al. Public and patient involvement in research to support genome services development in the UK. *J Transl Genet Genom* 2023;7:17-26. DOI
20. Goldenberg AJ, Lloyd-Puryear M, Brosco JP, et al; Bioethics and Legal Workgroup of the Newborn Screening Translational Research Network. Including ELSI research questions in newborn screening pilot studies. *Genet Med* 2019;21:525-33. DOI
21. Richards S, Aziz N, Bale S, et al; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405-24. DOI PubMed PMC