The use of synthetic data in clinical trials

From regulatory science to regulatory policy

Summary report
November 2026



Medicines & Healthcare products Regulatory Agency

PHG CAMBRIDGE

Authors

Patricia Cervera de la Cruz, Bhavya Krishnan, Colin Mitchell, Jamie Hearing, Kirsty Wydenbach, Sara Rajendran, Crina Cacou, Puja Myles and Andrea Manfrin.

About this document

This document presents the findings of a sandbox project on the use of synthetic data in clinical trials convened and developed by the Medicines and Healthcare products Regulatory Agency with support from PHG Cambridge.

These considerations do not constitute formal official guidance from the MHRA.

This work was supported by the Regulators' Pioneer Fund.

URLs in this report were correct as of November 2026.

You may re-use this information (excluding logos) free of charge in any format or medium, under the terms of the Open Government Licence. To view this licence, visit http://www.nationalarchives.gov.uk/doc/open-government-licence

or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third party copyright material you will need to obtain permission from the copyright holders concerned.

How to reference this report

Cervera de la Cruz, P. et al. The use of synthetic data in clinical trials: From regulatory science to regulatory policy. MHRA & PHG Cambridge. 2026.

DOI: https://doi.org/10.61599/kenq2

Contents

	Summary		4
1.	Introduction		5
2.	Approach		7
3.	Case studies		9
	Roche: Using synthetic data to address both the data access restrictions and the inherent statistical weaknesses of small subgroups		9
	UCB: Synthetic data for paediatric extrapolation in seizure syndrome trials		11
4.	Outcomes and lessons learnt		13
5.	Conclusions and next steps		16
	Acknowledgements		17
	Project materials		18
	Appendix A	Draft template for the submission of synthetic data to support clinical trial authorisation applications	18
	Appendix B	Draft principles developed and tested with expert group	26
	Appendix C	Glossary	27
	Appendix D	Attendee list for March workshop	31
	Appendix E	Agenda for March workshop	32
	Appendix F	Organisations in attendance at the international workshop (day one)	33
	Appendix G	Agenda for international workshop (day one)	34

Report summary

This report presents the findings of a regulatory sandbox convened by the Medicines and Healthcare products Regulatory Agency (MHRA), in partnership with PHG Cambridge, to examine the application of synthetic data within clinical trials. Through collaboration with two industry partners, Roche and UCB, the project investigated two priority use cases: privacy-preserving data augmentation to strengthen underpowered trial subgroups, and paediatric extrapolation to support dosing decisions in populations where conventional trial data are scarce.

In both instances, synthetic data approaches were shown to offer a scientifically grounded means of addressing longstanding challenges in clinical development, including limited recruitment, rare disease populations, and the ethical constraints of paediatric research.

Adopting a co-design methodology across two sequential phases, the project brought together regulators, sponsors, and international experts to move beyond theoretical discussion toward practical, case-based evaluation. This iterative process yielded two substantive outputs: a draft submission template to guide sponsors in documenting the rationale, methodology, and validation of synthetic data use within regulatory applications, and a set of seven draft principles establishing a foundation for international consensus on responsible governance. A key insight to emerge is that the acceptability of synthetic data rests on transparent, proportionate justification and pre-specified validation criteria.

The project demonstrates the considerable promise of synthetic data to improve trial efficiency, strengthen evidence generation, and accelerate patient access to novel therapies, while also affirming that it should be positioned as a complement to, rather than a replacement for, real-world evidence. More broadly, the sandbox approach itself is shown to be a valuable and replicable model for building regulatory confidence around emerging methodologies. The findings lay a strong foundation for continued international dialogue, with next steps focused on expanding the evidence base, deepening cross-agency alignment, and engaging ethical oversight bodies to support the responsible adoption of synthetic data in clinical research.

1. Introduction

This project established a regulatory sandbox[1] to explore how synthetic data could be incorporated into clinical trial protocols and regulatory decision-making. Clinical development is central to bringing novel therapies to patients, yet it faces challenges across operational, scientific, and ethical fronts. For instance, the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E17 1 guideline highlights that drug sponsors wishing to conduct a single, unified, multi-regional clinical trial (MRCT) that meets the regulatory submission requirements of multiple health authorities can face conflicting expectations.

While an MRCT, in theory, should enable faster global patient access to novel therapies, it may inadvertently lead to differential recruitment into standard-of-care (SOC) arms across geographic regions due to differences in underlying care pathways. Another challenge is ensuring that clinical trial populations adequately represent patients with rare diseases, life-threatening conditions, and vulnerable patient populations. An emerging challenge is also related to the long term follow up required in chronic disease cases, where recruitment and retention challenges for a placebo arm may result in delays to the trial readout or could be ethically challenging to recruit in the first instance.

Working with two industry partners (Roche and UCB), regulators and technical experts, the project sought to address some of these key challenges in clinical trial design and delivery by testing the application of synthetic data within real-world case studies.

The case studies were selected to reflect two priority applications of synthetic data within clinical trials: privacy-preserving data sharing and data augmentation/ extrapolation. Across both use cases, the rationale for using synthetic data is driven by common challenges facing clinical development, including difficulties recruiting patients into trials, the increasing cost and complexity of evidence generation, and the emergence of more targeted therapies for increasingly specific patient subgroups.

Synthetic data approaches may offer opportunities to improve trial efficiency, strengthen evidence generation where real-world observations are limited, and reduce the burden placed on patients, particularly in rare diseases and paediatric settings. It may also support more innovative trial designs, including the use of external controls, extrapolation approaches, and other methodologies that would otherwise be difficult to implement using conventional clinical trial data alone.

[1] A controlled regulatory and technical environment in which new methods, products or data approaches can be tested and evaluated, often with closer regulatory engagement before wider adoption. See Appendix C Glossary of terms used in the report.

At the same time, these potential benefits must be considered alongside important technical and ethical questions regarding the appropriate use, validation, and governance of synthetic data, including how to ensure its reliability and mitigate bias. Addressing these questions will help determine when synthetic data appraoches can appropriately complement real-world observations in clinical research.

The project's objectives were to assess how synthetic data approaches could support clinical trials involving human participants, inform the development of the first regulatory guidance on the use of synthetic data in clinical trials, and facilitate engagement with international regulatory authorities to explore opportunities for alignment and collaboration.

Through this work, the project aimed to lay the foundations for a new regulatory pathway for clinical trials that could improve efficiency and reduce development costs while accelerating patient access to safe innovative treatments. The sandbox also sought to generate practical learning for regulators and industry, supporting the development and dissemination of best practice for the use of synthetic data in clinical research both within the UK and internationally.

2. Approach

This project adopted a regulatory sandbox and co-design methodology, bringing together MHRA staff, including assessors and data scientists, with pharmaceutical sponsors to explore the use of synthetic data approaches in clinical trials collaboratively. Drawing on principles of multidisciplinary regulatory innovation and iterative problem-solving, the approach created a structured environment in which regulators and sponsors could jointly examine real-world challenges, test potential solutions, and develop a shared understanding of the evidentiary requirements needed to support regulatory decision-making.

The sandbox was designed to move beyond theoretical discussion and focus on practical application. Rather than evaluating synthetic data approaches against predefined criteria alone, project participants worked through specific use cases to identify challenges, assess methodological considerations, and explore how existing regulatory frameworks might apply or require adaptation. Through a series of workshops, technical reviews, and follow-up engagements, sponsors and MHRA experts co-developed a position on potential approaches for generating, validating, and presenting synthetic data within clinical trial submissions. This iterative process enabled emerging issues to be identified early and recommendations to be refined as new insights were generated.

The study was designed in two sequential phases to reflect a collaborative approach between regulatory and industry partners. The first phase deployed a regulatory sandbox to co-create learnings for the generation and use of synthetic data brought forward by use cases from industry partners (Roche and UCB) in collaboration with MHRA. These included a paediatric seizure disorder programme and a multiple myeloma programme, providing opportunities to explore both methodological and practical considerations across different therapeutic areas and development challenges. The case studies were deliberately chosen to reflect areas where conventional clinical trial approaches may face challenges, including small patient populations, difficulties generating robust comparator data, and the need to balance timely patient access with the generation of high-quality evidence.

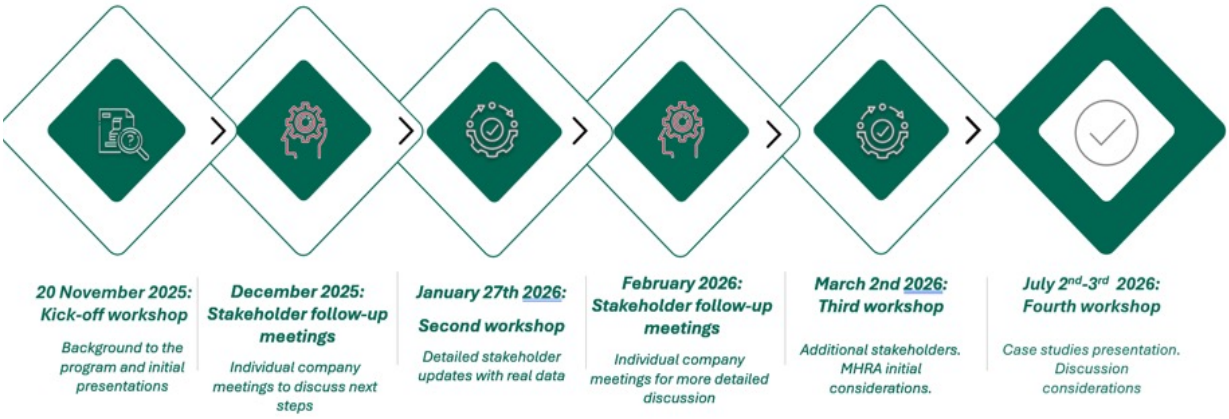
In the second phase, to build upon these foundational learnings and achieve wider alignment, a multi-stakeholder expert group was convened. This group comprised regulatory experts from the MHRA and other jurisdictions, Health Technology Assessment (HTA) experts, clinicians, academics, data scientists, industry representatives, and health service AI implementation experts. Through a series of roundtable workshops hosted by the MHRA (Appendix E and G), the expert group evaluated the use cases and lessons learned from the sandbox to build consensus on whether and how these synthetic data applications could be accepted for regulatory purposes. The final workshop in July 2026 convened international regulators to conduct regulatory discussions and seek opportunities for alignment across regulatory agencies.

Through the sandbox process, participants co-developed and refined a draft assessment template (see Appendix A) and questionnaire intended to help sponsors clearly describe the rationale, generation methods, validation approaches, and proposed use of synthetic data within regulatory submissions. Feedback on the evolving template was actively sought from participating sponsors throughout the project to ensure that it was both scientifically rigorous and practical for use in real-world development programmes.

While the project was intentionally focused on supporting the development of a UK regulatory perspective through the MHRA, it was recognised from the outset that synthetic data approaches are a global area of innovation. Consequently, input and perspectives from international regulatory stakeholders and experts were also sought to help ensure that the resulting recommendations would align, where possible, with broader international regulatory thinking and remain relevant to sponsors operating across multiple jurisdictions.

By combining co-design, iterative learning, and real-world case study evaluation, the sandbox approach provided a practical mechanism for translating emerging regulatory science into policy-relevant guidance while strengthening capability and knowledge within both industry and the MHRA.

Figure 1 Project journey



3. Case studies

Roche: Using synthetic data to address both the data access restrictions and the inherent statistical weaknesses of small subgroups

Case outline

This hypothetical case study describes a Phase 3 randomised controlled trial (RCT) evaluating a new investigational treatment (referred to as RO) in patients with relapsed or refractory multiple myeloma (RRMM). The trial compares RO against two existing standard-of-care (SoC) treatment regimens, SoC A and SoC B.

Multiple myeloma affects approximately 4 in every 100,000 people, making it a relatively rare condition. Running a large, adequately powered trial in this population is inherently challenging due to the size of the patient population.

Synthetic data approaches are being proposed as a proactive mitigation plan to address limitations in data, due to imbalances in trial recruitment, before it affects the trial outcome and patient access to medicines.

The challenge

There is currently no single, universally agreed standard-of-care (SoC) for patients with RRMM. Treatment decisions vary considerably by region, driven by local guidelines, availability of drugs, and clinical practice patterns. Specifically:

- SoC B is used more commonly in Europe than in the United States.
- The frequency with which SoC B is used also varies depending on how many prior lines of therapy and specific drugs a patient has already received.

As a result, enrolment into the SoC B arm is expected to be substantially lower than enrolment into SoC A. The overall trial is powered to compare RO against the combined SoC population (SoC A and SoC B together), but it will be heavily dominated by SoC A as the number of patients in the SoC B group is simply too small to measure its unique contribution with enough precision. That is why advanced statistical methods are needed to decrease uncertainty enough to allow the comparison of RO vs SoC B.

This is not a minor limitation. Regulatory agencies and clinicians may need to understand how RO performs against each SoC regimen individually, because prescribing decisions are made with consideration of the specific regimen a patient would otherwise receive. An inconclusive result for the RO versus SoC B comparison could affect the scope of any approved drug label, resulting in those patients missing the opportunity to receive a beneficial drug.

The solution that is being tested

The goal is not to fabricate outcomes, but to generate plausible, representative data points that fill the statistical gap created by low enrolment, allowing the SoC B subgroup analysis to be conducted with adequate power.

To address this limited sample size in the SoC B subgroup, a proactive plan was proposed to supplement the trial data, which includes a four-stage data-augmentation pipeline:

- First, data from high-quality historical Phase 3 trials are identified and the suitability of this source data is assessed to ensure it is fit for purpose.
- Second, to overcome the limitation of data access, synthetic data versions of the historical trial data are generated using well-established algorithms such as MIIC-SDG or SynthPop, to computationally produce high-fidelity, privacy-preserving synthetic patient profiles that mirror the underlying multivariate associations of the original data.
- Third, the resulting dataset undergoes a rigorous quality scorecard evaluation that verifies marginal and multivariate distribution congruence alongside an outcome-based Cox proportional hazards validation.
- Finally, this synthetic data is integrated into the randomised controlled trial via a Bayesian dynamic borrowing framework, that dynamically calibrates and discounts the weight of the borrowed external information, based on its statistical consistency with the concurrent trial data.

What synthetic data approaches add

Rather than waiting for the final analysis to reveal an inconclusive SoC B subgroup result, at which point it would be too late to act, this plan is being developed and validated in advance. Proactive development of the synthetic data methodology allows for:

- Pre-specified validation of the synthetic data approach before unblinding to ensure the weighting given to synthetic data is determined in advance on the basis of its statistical consistency with real trial data.
- Regulatory engagement on the methodology before results are known
- A stronger evidentiary basis for the SoC B subgroup if the trial proceeds as anticipated.

UCB: Synthetic data for paediatric extrapolation in seizure syndromes trials

Case outline

UCB is exploring whether synthetic data approaches can help improve understanding of how seizure symdromes treatments could be dosed in children with rare seizure disorders. Studying these conditions in paediatric cases can be particularly challenging because there are very few affected children, and ethical considerations often limit the use of traditional placebo-controlled clinical trials.

To address these evidence gaps, UCB is investigating whether synthetic data generated from existing paediatric studies can support the process of paediatric extrapolation for an anti-seizure medicine. Paediatric extrapolation is the process of using evidence from one population, such as adults, together with other supporting data to predict how a treatment is likely to perform in children.

The challenge

A key challenge is understanding how the medicine under investigation works in patient groups that are poorly represented in available clinical data.

UCB has already used evidence from adult studies, alongside pharmacokinetic (PK) models that describe how the body absorbs, distributes and clears a drug, to estimate appropriate dosing and safety in children. However, important evidence gaps remain for some smaller patient groups based on challenges relating to variation in weight distribution among different demographics. For example, in world regions with few cases of childhood obesity clinical data for such patients may be limited or unavailable, making it difficult to directly confirm whether model predictions are accurate.

As a result, there is uncertainty about how well existing models perform in these groups, creating a need for additional approaches like the use of synthetic data to augment existing datasets that can help explore these unanswered questions.

The solution that is being tested

UCB is not using synthetic data as a replacement for real clinical data. Instead, it is being used to strengthen and test the evidence that already exists.

The aim is to assess how sensitive current conclusions are to different assumptions within the models (e.g. PK structure and covariate relationships) and to explore how the treatment may perform in patient groups where limited real-world evidence is available.

The approach combines two complementary methods:

- Biological and pharmacological models, which incorporate established knowledge about how the drug behaves in the body and how factors such as age and body weight influence drug exposure.
- Data-driven techniques, including machine learning, which can generate realistic patient profiles and identify patterns within existing data.

At the centre of the process is a PK model that has previously been validated and accepted by regulators. Although it cannot be fully verified in populations where little or no data exist, it provides a scientifically grounded starting point based on known biological relationships.

Additional statistical and machine-learning methods are then used to generate realistic virtual patient populations. These methods recreate relationships between characteristics such as age, weight and ethnicity, helping researchers explore how different patient groups may respond to treatment. Copula models capture how key characteristics like age, weight, and ethnicity occur together, while machine learning models help account for additional patterns in how drug levels change over time. Model complexity is carefully controlled to ensure results can be generalised beyond the training data.

By combining mechanistic models with machine learning, UCB aims to make uncertainty transparent and measurable. This hybrid approach may provide a safer, more efficient pathway for delivering treatments to children with rare conditions.

What synthetic data approaches add

The goal is to move beyond static data and stress-test the PK model through various simulated scenarios. By shifting variables like age and weight, UCB can define the boundaries of where the model remains robust. This can allow regulators to see if exposure equivalence holds across subgroups, potentially reducing the need for further invasive clinical trials.

Synthetic data are explicitly positioned as complementary, not a replacement for real (e.g. paediatric) data. Where little or no real world evidence is available, confidence in the results relies on a combination of scientific understanding, expert judgement and consistency with existing evidence. Some uncertainty inevitably remains, particularly for populations that cannot be directly studied. However, the approach can make these uncertainties more transparent, allow a wider range of scenarios to be explored, and could support more informed decision-making about treatments for children with rare conditions.

4. Outcomes and lessons learnt

The project generated learning at two levels. First, it identified key considerations regarding the use of synthetic data in clinical trial applications, including the circumstances in which synthetic data may add value, the evidence required to support the use of synthetic data, and the conditions necessary for regulatory confidence. Second, the project generated broader lessons regarding how regulators, sponsors, and other stakeholders can work together to assess technically complex innovations whose regulatory acceptability remains uncertain.

Grounded by the case studies, the sandbox provided a structured environment in which emerging methods could be examined through real-world applications, enabling technical, regulatory, ethical, and governance questions to be considered simultaneously and iteratively. These learnings and outcomes are summarised below, and they will be disseminated more widely via forthcoming academic publications, including an initial overview paper currently under review.

Acceptability of and justification for use of synthetic data in clinical trial applications

A key lesson from the project was that the acceptability of synthetic data approaches in clinical trial applications depends on transparent justification of their relevance and appropriateness in the specific context in which they are proposed. Across sandbox and workshop discussions, participants consistently emphasised the importance of clearly articulating why synthetic data approaches are being used, what evidentiary gap they intend to address, and how they contribute to the overall evidence package.

Acceptability was seen as requiring a proportionate assessment of benefits and risks, supported by transparency regarding assumptions, limitations, uncertainty and the intended context of use. Sponsors were also encouraged to adopt a reflexive approach, considering not whether a particular method is inherently “right” or “wrong”, but whether it is appropriate and fit for purpose for the regulatory question being addressed. Furthermore, discussions highlighted the importance of prospective evaluation, including pre-specifying the criteria that will be used to determine whether synthetic data approaches are fit for purpose and defining in advance what would constitute acceptable performance or indicate that an approach has failed.

A major innovation of the project was the translation of this emerging consensus and learning into a practical submission template for the use of synthetic data in clinical trial applications (Appendix A). Rather than being produced at the conclusion of the project, the template evolved iteratively alongside the stakeholder discussions and workshops, allowing learning to be incorporated as it emerged. The template operationalises many of the project's key lessons, encouraging transparent justification for synthetic data use, clear documentation of source data, methodology and limitations, explicit consideration of evidentiary reliance, and prospective specification of validation and success criteria. More broadly, it provides a structured mechanism for supporting consistent and transparent dialogue between sponsors and regulators when considering the use of synthetic data in clinical trial applications.

Principles

While the template provided a practical reflexive framework for assessing specific synthetic data applications, discussions during the July international workshop highlighted an opportunity to identify areas of emerging consensus amongst international regulators. Although regulatory approaches and evidentiary requirements may continue to vary across jurisdictions, participants identified a number of common considerations that were considered fundamental to the responsible use of synthetic data in clinical trials.

A key outcome of the project was therefore the development of a preliminary set of draft principles (Appendix B) that could serve as a starting point for future international alignment on the governance and evaluation of synthetic data in clinical trials. These include the need for a clear value justification for the use of synthetic data; demonstration that the proposed approach is fit for purpose; assurance of data quality and representativeness; recognition of synthetic data as a supplement to, rather than a replacement for, empirical evidence; protection of patient safety and trial integrity; transparency regarding methods, assumptions and limitations; and pre-specification of validation and success criteria.

While the principles require further refinement and elaboration, they represent an important first step towards identifying areas of common ground amongst regulators and provide an initial foundation for future international alignment on the governance, evaluation and acceptable use of synthetic data in clinical trials.

Discussions during the international regulators' workshop revealed a strong appetite among participants to continue working towards greater convergence in the standards, evidence requirements and governance approaches applied to synthetic data, recognising that some degree of alignment will be important to enable clinical trials across multiple jurisdictions.

Deliberative engagement as an approach

The project generated broader lessons regarding how regulators, sponsors and other stakeholders can work together to address challenges associated with emerging technologies for which there is a need to build regulatory confidence. A key lesson was the value of the regulatory sandbox approach itself. By creating a structured environment for collaboration and iterative problem-solving, the sandbox enabled participants to explore novel applications of synthetic data through real-world use cases, supporting a more practical and nuanced understanding of the evidentiary, regulatory and governance questions involved than would have been possible through theoretical discussion alone.

The project also demonstrated the importance of creating a trusted environment for collaborative learning. Initial discussions were often constrained by understandable commercial sensitivities surrounding emerging methodologies and development strategies. However, repeated engagement and the iterative nature of the sandbox helped build trust between participants, enabling increasingly open discussion of technical challenges, limitations and lessons learned. This richer exchange of perspectives was an important contributor to the collective learning generated through the programme.

These experiences also highlighted the value of deliberative engagement as a mechanism for addressing technically complex innovations. Through successive workshops, sandbox activities and international regulatory discussions, stakeholders were able to move beyond technical questions alone to consider wider issues relating to acceptability, governance, trust, evidence standards and decision-making.

Overall, the project demonstrated that the sandbox provides a practical mechanism for supporting innovation while helping regulators, sponsors and other stakeholders collectively explore the boundaries of acceptable use.

5. Conclusions and next steps

The project has made considerable tangible progress towards the objectives of assessing how synthetic data could support clinical trials involving human participants, informing the development of the first regulatory guidance on the use of synthetic data in clinical trials, and facilitating engagement with international regulatory authorities to explore opportunities for alignment and collaboration.

Through the case studies, submission template and draft principles, the project has generated practical insights into the conditions under which synthetic data may be used credibly, safely and transparently within clinical trials, while also establishing a foundation for future regulatory and international discussions on its governance and acceptable use.

In addition, it has surfaced the potential of the regulatory sandbox as an innovative deliberative approach to policy development, which could play a significant role in addressing other areas of regulatory uncertainty around adoption of innovative technologies in the development of medicines and medical devices. However, further work is needed to routinise the use of synthetic data in clinical trial applications and to develop consensus at an international level. Further steps would build on the foundations established through this project and support the further development, implementation and alignment of approaches to synthetic data in clinical trials.

What is still needed to advance regulatory confidence and certainty?

- **Further use cases and typologies**: The template provides a strong foundation for discussion between regulator and sponsors, and initial use cases provide important tangible examples, but identification of further use cases and typologies would aid sponsors in developing their proposals for other potentially beneficial uses of synthetic data in clinical trials.
- **Continued organisational learning and alignment within the MHRA**: Further dialogue and collaboration is required within the MHRA to ensure alignment across relevant teams on this topic, including clinical investigations and trials, licensing, and safety and surveillance.
- **Further international regulatory dialogue and collaboration:** Further international dialogue among regulatory authorities to progress alignment and start a process that begins with general principles and may advance through to specific and technical guidance.
- **Engagement with ethical oversight organisations:** Further delibertion including with ethical oversight organisations (in particular the NHS Health Research Authority) will ensure development of an ethically robust approach which is aligned across the regulatory pathway.
- **Extension to other emerging technologies:** More generally, rolling out this collaborative and deliberative approach to adjacent challenges which require close dialogue and consensus building. The model developed in this project complements those of similar MHRA initiatives such as the AI Airlock and AI sandbox, and provides further evidence of the progress that can be achieved through such approaches.

Acknowledgements

This work was informed by a previously funded Regulators' Pioneer Fund programme, which provided the foundation for the development of this regulatory sandbox. The authors are grateful to the Expert Group for their advice and support, and to all stakeholders who participated in the roundtables and workshops, both in person and online. The authors also extend their sincere thanks to Sara Rajendran for managing delivery of the research programme and contributing to its successful completion.

Project materials

Appendix A: Draft template for the submission of synthetic data to support clinical trial authorisation applications

Template

- The template must be completed and be as detailed as possible. It must be submitted as part of the clinical trial authorisation (CTA) application documentation.
- Reference to other documents that have been submitted is possible, for example, the protocol, but a summary must still be included in this document.
- If reference is made to the statistical analysis plan, then this must be included in the clinical trial document package.
- The cover letter must indicate that the trial includes the use of synthetic data.
- Inclusion of a cover letter statement and this document will be reviewed as part of the initial validation for a CTA application.
- The overarching requirement is that the proposed use of synthetic data is justified transparently and clearly. The questions and considerations in this document encourage reflection on whether the proposed approach is appropriate and proportionate, given the benefits and risks involved in a specific context.
- This won't lead to an answer about whether the method is right or wrong but encourages the trial sponsor to adopt a reflexive approach in determining what method is most appropriate and to provide transparent justifications.
- Transparency about limitations of the proposed approach is also crucial.
- Sponsors should pre-specify the criteria that will be used to determine whether the synthetic data are fit for purpose in this use case. This should include:
  - what constitutes an acceptable level of agreement or performance (e.g. fidelity, validation metrics, or impact on analysis outcomes)
  - what outcomes would indicate that the approach is not adequate or has failed

These criteria should be defined in advance of synthetic data generation and analysis.

Description of the clinical study, synthetic data value proposition, risks and justification

Sponsors should provide a brief, high-level summary of the clinical study, although they may instead refer to the full protocol when sufficient detail is already provided.

Provide a clear rationale for why you are proposing to use synthetic data in your study, considering how synthetic data fits within the overall evidence package.

This includes whether the output is to inform other clinical trials in future and/or a marketing authorisation application.

The context of use is important; for example, is it to provide support to pivotal data, or is it aiming to de-risk an issue?

Consider outlining the problem - the challenge – how you plan to address it – what is the consequence.

This section should include a discussion and conclusion on the level(s) of reliance in accordance with the table below. Applicants should assess reliance at the level of each relevant population, endpoint, or evidentiary component, considering the availability, quality, and limitations of real-world data. Where reliance* varies, multiple levels may be assigned; a single overall level is not required:

Level 1	Are you using synthetic data to supplement an already large real-world dataset?
Level 2	Are you incorporating synthetic data alongside expert knowledge or mechanistic modelling to compensate for gaps in real data?
Level 3	Are you relying on interpolation or boosting techniques to generate additional synthetic data where real data is limited?
Level 4	Are you relying primarily on synthetic data, without additional real-world evidence or expert-driven adjustments?

* “Reliance” refers to how much weight synthetic data carries in informing regulatory conclusions, relative to clinical trial data, real-world data, and other evidence sources.

Notes

a. What new evidence or insights may synthetic data provide?

b. What gap is synthetic data filling?

c. How is it different to a more traditional methodology?

d. How do you anticipate that the synthetic data will support regulatory decision making? For example, if there is potential to impact a future Marketing Authorisation Application (MAA), such as data that may support the final Summary of Product Characteristics (SmPC), this should be clarified and specified how it relates to the clinical trial objectives.

e. Can you justify your approach, considering:
   i. What would you otherwise have to do without using synthetic data? Speed and cost are not considered appropriate reasons for the use of synthetic data. Scientific justification should be the key consideration. Discussion of the validity and known regulatory acceptance of other methodologies should be considered, for example, as identified in regulatory guidance.
   ii. Whether use of synthetic data may exacerbate existing risks or introduce new risks? This may include statistical risks or direct risks to participants in the trial, for example.
   iii. What is the consequence of making a wrong decision based on synthetic data?

f. Why, on the balance of potential harms and benefits, do you think it is appropriate to use synthetic data in this case?

   The discussion above should also address elements covered by existing guidelines and how the proposals align with those guidelines.

Source data provenance and suitability (note may not be relevant to mechanistic model-based approaches)

The credibility of synthetic data depends on the provenance of the source data. A non-exhaustive set of considerations are set out below. Sponsors may provide a data flow diagram describing the synthetic data generation process, data handling, and transfer, as supplementary information to support transparency on source data provenance.

When discussing any bias mitigation, it is important to clarify whether the mitigation is applied before or after the synthetic data is generated.

Notes

a. What is your source or ground truth dataset for synthetic data generation? It is important to be fully transparent about the source data.

b. How do you justify its selection? In scenarios where the source data is to bematched to real data (for example, for eligibility criteria) it may be helpful to ask questions such as to what extent are they similar? Where do they differ, and why? Do you weigh certain criteria as critical and others less critical? In many ways, this is no different to matching RWD to real trial data, because any differences are then carried into the synthetic data. Will you be selective about which patients in the other data you use? Or is it all of it, with the trial selected as a whole? The impact of this needs discussion in terms of risk.

c. What criteria have you used to assess suitability of the source data for this use case? It is important that there are boundaries for the selected source data and that these relate to the data gap the synthetic data is expected to fill. For example, there may be boundaries for age, disease severity or other baseline patient characteristics, in the same way RWD can be 'matched' to real patients in a clinical trial.

d. Have you ensured that the source data are sufficiently representative of the intended use population? This is not just about ensuring representative source data but ensuring it is fit for purpose, for example ask yourself 'how old is the data (is it out of date?)', 'how much original data is there?', 'is it sufficient?'.

e. If data has been modified prior to use, can you describe all modifications (including privacy-enhancing techniques such as suppression, anonymisation, aggregation, as well as other alterations, such as pre-processing)? What are the likely implications for the fidelity and diversity of the resulting synthetic data?

Synthetic data generation methodological detail

The structure, assumptions, and inputs of the synthetic data generation (SDG) model influence the quality, fidelity, and representativeness of the synthetic data it produces.

Notes

a. What methods will you use to generate the synthetic data, and which model(s) will be applied? Is the approach data-driven, mechanistic (process-driven), or hybrid?
   i. If using a mechanistic or model-based approach: Describe the model origin, validation status, and basis for suitability.
   ii. Specify any software libraries used.

b. What are the key assumptions underlying the SDG model? And how do you justify those assumptions?

c. What parameters were used in the generation process, and how were they selected or tuned? Did they need adjusting during the data generation? If so, why?

d. Has the synthetic data been generated specifically for this context (e.g. clinical trial design), or has it been repurposed from another use?

e. Has the SDG model been previously validated? If so, in what context? (Note: This is particularly important when synthetic data is used to augment or boost real-world data, especially for underrepresented groups or edge cases)

f. Is the SDG model likely to reinforce or amplify biases present in the source data?

g. Have you incorporated any external information—beyond the training data-set—to mitigate bias and ensure sufficient diversity, including representation of edge cases?

Synthetic data quality and suitability

Notes

a. Have you assessed whether synthetic data introduces, preserves, or mitigates biases from the source data? If synthetic data is used for bias mitigation, how have you evaluated its suitability for this purpose?

b. Have you ensured sufficient fidelity to the data?
   i. State whether low, medium, or high fidelity is needed.
   ii. Explain how you decide what differences between synthetic data and real data are acceptable.

c. Where both statistical fidelity assessments and mechanistic model–based evaluations are used, describe how consistency between these approaches has been assessed. In cases where results differ, explain:
   i. which approach takes precedence and why
   ii. the criteria used to resolve discrepancies
   iii. how this impacts confidence in the synthetic data for the intended use.

d. Does your synthetic data generation account for potential shifts in medical practice, patient demographics, or disease trends that may impact its representativeness?

e. Have you guarded against overfitting? Does the synthetic data support generalisability, or does it risk overfitting?

f. How does the synthetic data compare to real data on a univariate level?

g. Have you assessed multivariate relationships to ensure that interactions between variables in the synthetic data align with those in the real data?

h. Considerations for boosting:
   i. Boosting to augment real-world data: Have you externally validated the boosted synthetic data using available real-world data and human expert review? What oversight measures are in place? Have you planned future data collection to further assess generalisability?
   ii. Boosting when real data is unavailable for validation: If real data is inaccessible, what alternative validation methods have you employed to assess appropriateness and minimise the risk of bias?
   iii. Boosting and generalisation boundaries: How do you define the boundaries between acceptable extrapolation and unsafe synthetic generalisation? What safeguards are in place to avoid generating synthetic data distributions that are not clinically plausible? How is clinical relevance ensured in such cases?

i. How is uncertainty within the synthetic dataset assessed (either qualitatively or quantitatively), in terms of data completeness, representativeness, or unknown biases?
   i. Based on your current evidence package, what level of uncertainty remains in your system, and how does the quality of the synthetic data help reduce it? This may not be relevant at the clinical trial stage but could have an impact on a future clinical trial or MAA, and uncertainty can be discussed in that capacity.
   ii. What is the target level of certainty required for your intended use, and what role does synthetic data play in reaching that level?
   iii. Have you anticipated potential worst-case scenarios due to uncertainty linked to the synthetic data?

Limitations
Please describe any limitations associated with your use case which have not been discussed above.

Notes

a. How does the use of synthetic data for this use case affect your risk management approach?

Other considerations when using synthetic data

Notes

a. In order to comply with Good Clinical Practice (GCP), explain how you will document and keep track of the use of the generative model so it can be checked or reused. Outline briefly which aspects would be available for inspection if required.

b. If applicable: Have you implemented procedural safeguards to prevent data leakage or inappropriate reuse of synthetic datasets across training, tuning, or validation phases that could lead to overfitting?

Appendix B: Draft principles developed and tested with expert group

Draft principles
Principle 1 (Value justification) The purpose of using synthetic data in a clinical trial should be clear, specific and scientifically justified. Its use should address an identified unmet need, limitation, or evidence gap that cannot be adequately addressed through conventional trial design approaches. Applicants should clearly demonstrate the added value of the synthetic data in this specific context, its relevance to the intended purpose, and why its use is appropriate for the specific context. This justification should be explicitly described as part of the clinical trial application. Synthetic data approaches may, for example, support the generation of evidence in settings where recruitment is challenging, enhance the efficiency of a trial while maintaining scientific validity, or facilitate the development of medicines for conditions with unmet medical need and thereby contribute to faster more timely patient access.
Principle 2 (Approach and fitness for purpose) The approach used to generate, validate, incorporate and manage synthetic data should be chosen based on the intended use and specific objectives of the trial. No single approach is universally appropriate and the chosen approach should be scientifically justified, proportionate to the context of use, and consistent with applicable regulatory guidance and legislative requirements.
Principle 3 (Data quality and representativeness) The source data, synthetic data, and associated generation processes should be demonstrated to be of sufficient quality for the intended use. Synthetic data cannot compensate for deficiencies in the underlying data source; therefore, the suitability, provenance, and representativeness of the source data should be established and justified.
Principle 4 (Supplementary, not replacement) Synthetic data should generally be used to supplement evidence derived from observed participant data and should not ordinarily constitute the primary source of pivotal evidence.
Principle 5 (Patient safety and integrity) The use of synthetic data must preserve participant safety, scientific validity, data integrity, and the interpretability of evidence used to support regulatory decision-making. Participant safety should always take precedence and should not be compromised by the use of synthetic data. Risks associated with the use of synthetic data should be identified, appropriately mitigated where possible, and monitored throughout the lifecycle of the clinical trial.
Principle 6 (Transparency) The use of synthetic data should be transparent, traceable and reproducible. Sponsors should clearly document the source data, generation approach, assumptions, validation activities, intended use within the clinical trial, and any limitations or uncertainties associated with the synthetic data. Sufficient documentation should be maintained to enable independent assessment and reproduction of key outputs and the decisions informed by the use of synthetic data.
Principle 7 (Pre-specification) The intended use, or likely use, of synthetic data should be prospectively specified and justified in the clinical trial protocol. The protocol should clearly describe the role of synthetic data within the trial and any associated decision-making processes. Where supporting technical information is held elsewhere, it should be readily accessible to enable transparent review by all relevant stakeholders.

Appendix C: Glossary

Bayesian (dynamic) borrowing	A statistical approach in which data from an external or prior source, such as a historical trial, a related patient population, or a previous study arm, is combined with data from a current trial to strengthen the evidence base. The amount of "borrowed" information is adjusted dynamically depending on how similar or consistent the external data are with the data being collected, so that less weight is given to external data when those data conflict with current results.
Clinical trial authorisation	The formal regulatory approval required before a clinical trial of a medicine can begin in the UK. The MHRA reviews an application covering the trial protocol, investigational medicinal product information, and supporting safety and quality data. Authorisation confirms that the trial may proceed in line with the agreed design and safeguards.
Control arm	The group of trial participants in a clinical trial who do not receive the experimental treatment under investigation, but instead receive a comparator such as a placebo or the standard of care. Outcomes in the control arm provide the benchmark against which the experimental treatment's effects are measured.
Copula model	A statistical technique used to describe and simulate the relationship between two or more variables independently of their individual distributions. In synthetic data generation, copula models help preserve realistic correlations between variables, such as the relationship between age and a particular biomarker, in the synthetic dataset.
Cox proportional hazards model	A statistical method used to assess how different factors influence the time until an event occurs, such as disease progression or death.
Cox proportional hazards validation	The process of checking that a Cox proportional hazards model meets its underlying assumptions and produces reliable estimates. Validation typically involves confirming the proportional hazards assumption holds over time and testing the model's performance against independent or held-out data.
Distribution variate congruence	The degree to which the statistical distribution of values (variates) in a synthetic or modelled dataset matches that of the real source data for a given variable. Strong congruence indicates that the synthetic data reproduces the shape, spread, and key statistical properties of the original variable, supporting its credibility for downstream analysis.

Fidelity	In the context of synthetic data, fidelity refers to how closely the statistical properties, structure, and patterns of a synthetic dataset resemble those of the real data it was generated from. High fidelity means the synthetic data behave similarly to the original for analytical purposes, without necessarily containing any real patient records.
First-in-human trials	Early-phase clinical trials in which an investigational medicine is administered to people for the first time, typically in a small group of healthy volunteers or patients. These trials primarily assess safety, tolerability, and pharmacokinetics, and they inform dosing and design for later-phase studies.
Machine learning (ML)	A branch of artificial intelligence in which algorithms identify patterns in data and improve their performance on a task through exposure to data, rather than through explicit hand written rules. In clinical research, machine learning models are increasingly used for tasks ranging from summarisation of clinical notes, to risk prediction models, to diagnostic tools.
Model informed drug development (MIDD)	Model-Informed Drug Development (MIDD), sometimes referred to as Model-Based Drug Development (MBDD), is a process whereby key program decisions are supported by mathematical models and simulations that predict the likelihood of success for the drug. By using mathematical and statistical methods, one is able to build models of drug concentrations (pharmacokinetic) and/or drug responses (pharmacodynamic) over a time course.
Paediatric extrapolation	A scientific approach that allows efficacy, safety, and/or dosing data generated in adults (or older children) to be used in part or full to support conclusions about a medicine in a younger paediatric population, reducing the need to repeat full clinical trials in children where the disease and treatment response are expected to be sufficiently similar.
Pharmacokinetic models	Mathematical models that describe how a drug is absorbed, distributed, metabolised and eliminated by the body over time (pharmacokinetics). These models are used to predict drug concentrations under different dosing regimens and populations, and they inform decisions on dose selection.
Randomised control trial (RCT)	Studies in which participants are randomly assigned to receive either the intervention being tested or a comparator, such as a placebo or standard of care. Randomisation helps ensure that the groups are comparable at the outset, so that differences in outcomes can be attributed to the intervention rather than to other factors.

Rare disease	A condition that affects a small proportion of the population, commonly defined as affecting fewer than 1 in 2,000 people. Rare diseases present particular challenges for clinical trial design, including small and geographically dispersed patient populations, which has driven interest in approaches such as extrapolation and synthetic control data.
Real world data (RWD)	Data relating to patient health status or delivery of health care collected outside of a clinical study. There are vast amounts of data being collected on patients, for example, in electronic health records (EHR), and disease and patient registries. Such data are commonly called RWD, reflecting that they are collected while patients go about their regular lives, as opposed to being specifically collected in a clinical study. When such data are analysed, the information produced may be referred to as real-world evidence (RWE).
Sandbox	A controlled regulatory or technical environment in which new methods, products or data approaches can be tested and evaluated, often with closer regulatory engagement, before wider adoption.
Source data	The original records from which clinical trial data are derived, such as hospital records, laboratory reports, or patient charts. Source data provide the evidence base for the data entered into a trial's case report forms and are central to data verification and audit.
Standard of care	The treatment or set of treatments that is widely accepted by clinicians as appropriate for a given condition, based on current evidence and clinical guidelines. In a clinical trial, the standard of care is often used as the comparator in the control arm when withholding all treatment would be inappropriate.
Synthetic data	Data that have been created artificially (e.g., through statistical modeling, computer simulation) so that new values and/or data elements are generated. Generally, synthetic data are intended to represent the structure, properties and relationships seen in actual patient data, except that they do not contain any real or specific information about individuals. For example, in healthcare, synthetic data are artificial data that are intended to mimic the properties and relationships seen in real patient data. Synthetic data are examples that have been partially or fully generated using computational techniques rather than acquired from a human subject by a physical system. (FDA Digital Health and AI Glossary)

Training data	The dataset used to teach a machine learning (ML) model or algorithm to recognise patterns and make predictions. The quality, size and representativeness of training data strongly influence how well a model performs.
Validation	The process of checking that data, a model, or a method performs as intended and produces accurate, reliable, and reproducible results. In the context of a data-driven approach to synthetic data generation, validation involves comparing the synthetic dataset against the source data to confirm it preserves the statistical relationships and utility needed for its intended analytical use.
Virtual patient population	A virtual patient is a hypothetical model built using extensive population data rather than representing a specific individual. Researchers use these models to predict how different demographic groups might respond to new treatments, helping to optimise safety and dosages before conducting physical clinical trials. They are primarily used to answer “what if?” questions and to generate broader evidence for drug development and licensing.

Appendix D: Attendee list March workshop

MHRA

- Crina Cacou, Head of Clinical Investigations and Trials Operations
- Kingyin Lee, Head of Clinical Trials
- Andrea Manfrin, Deputy Director, Clinical Investigations and Trials (CIT)
- Puja Myles, Director, Clinical Practice Research Datalink
- Sara Rajendran, Policy Specialist, CIT
- Khadija Rantell, Lead Senior Statistical Assessor, HQA
- Nisha Kanwar, Clinical Pharmacology Assessor, HQA
- Susan Cole, Clinical Pharmacology Assessor, HQA
- Maria Malamatari, Clinical Pharmacology Assessor, MHRA
- Shiva Ramroop, Medical Assessor at the MHRA and paediatric specialist
- Ibrahim Ivan, Medical Assessor in Clinical Trials

NICE

- Michael Bell, Senior Analyst
- Stephen Duffield, Associate Director for Real-World Evidence

Independent

- Kirsty Wydenbach, Head of Regulatory Strategy and Drug Development Clinician, Weatherden; former expert medical assessor at the MHRA within CIT

PHG

- Colin Mitchell, Head of Humanities
- Patricia Cervera de la Cruz, Policy Analyst
- Jamie Hearing, Senior Policy Analyst

Roche

- Francesco Brizzi, Senior Scientist (Predictive Modelling)
- Felipe Castro, Epidemiologist and Principal Data Scientist
- Gracy Crane, Regulatory Science Policy Lead
- Federico Mattiello, Biostatistician
- Laurent-Philippe Albou, Scientific Project Manager
- Ken Wang, Senior Principal Scientist

UCB

- Helena Andres-Terre, Associate Director
- Daniel Manela, Machine Learning R&D and Innovation Lead
- Martine Dehlinger-Kremer, Senior Development Strategy Lead Pediatrics
- Walter Krauwinkel, Director Quantitative Clinical Pharmacology / PKPD Lead

External stakeholders

- Venet Osmani, Professor of Clinical AI and Machine Learning at Queen Mary, University of London
- Allan Tucker, Professor of Artificial Intelligence in the Department of Computer Science at Brunel University
- Gillian Raab, Research Fellow at University of Edinburgh
- Sofia Villar, Research Professor at MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge

Appendix E: Agenda for March workshop

Time	Activity	Speaker / Participants
10.00 - 10.15	Welcome, introductions	All
10.15 - 10.45	Progress & temperature check	Andrea Manfrin
10.45 - 11.45	Identifying key considerations from a regulatory perspective	Kirsty Wydenbach
11.45 - 12.45	UCB case study presentation & discussion	UCB / All
12.45 - 13.30	Lunch	
13.30 - 14.30	Roche case study presentation & discussion	Roche / All
14.30 - 15.15	Developing guidance and supportive information	PHG
15.15 - 15.45	Building international consensus	Puja Myles / PHG
15.45 - 16.00	Wrap up & next steps	Puja Myles

Appendix F: Organisations in attendance at the international workshop (day 1)

Ministerial department:

- Department for Science, Innovation and Technology (DSIT, UK)

Regulators:

- Medicines and Healthcare products Regulatory Agency (MHRA, UK)
- Health Research Authority (HRA, UK)
- Health Science Authority (HSA, Singapore)
- Health Canada/Sante' Canada (HC/SC, Canada)
- Swissmedic (Switzerland)
- Therapeutic Goods Administration (TGA, Australia)
- Swedish Medical Product Agency (Läkemedelsverket) (MPA, Sweden).

Executive non-departmental public body:

- National Institute for Health and Care Excellence (NICE, UK)

Academia:

- University of Cambridge
- University of Manchester
- Queen Mary, University of London (QMUL)
- Napier University Edinburgh
- University of Ottawa (Canada)
- University of Queensland (Australia)

Pharmaceutical companies:

- Roche
- UCB
- GSK

Other organisations:

- Association of British Pharmaceutical Industries (ABPI)
- Pistoia Alliance
- Flagship Pioneering
- PHG Cambridge (formerly PHG Foundation)
- Weatherden

Appendix G: Agenda for international workshop (day 1)

Time	Activity	Speaker / Participants
09.30 - 10.00	Arrival and registration	All
10.00 - 10.15	Chair's welcome and introduction	Kirsty Wydenbach
10.15 - 10.35	MHRA introduction to the project and UK regulatory context	Andrea Manfrin
10.35 - 10.55	Clarification questions and digital feedback	All
10.55 - 11.15	Regulators' perspectives	Regulators
11.15 - 11.30	Break	
11.30 - 12.30	Sandbox case studies	Presentations by Roche and UCB, followd by questions and feedback
12.30 - 13.15	Lunch	
13.15 - 13.45	Key considerations for synthetics data in clinical trials	Kirsty Wydenbach
13.45 - 14.45	Working session: exploring the sandbox learnings	Small group discussions miro board for online attendees
14.45 - 15.00	Break	
15.00 - 15.50	Feedback from the working session	Rapporteurs followed by plenary discussion
15.50 - 16.00	Conclusion remarks and close	Kirsty Wydenbach, Andrea Manfrin



Medicines & Healthcare products Regulatory Agency

The Medicines and Healthcare products Regulatory Agency (MHRA) regulates medicines, medical devices and blood components for transfusion in the UK. The MHRA is responsible for making sure these products meet set standards for safety, quality and effectiveness. The MHRA is an Executive Agency of the Department of Health and Social Care.
www.gov.uk/mhra

PHG CAMBRIDGE

PHG Cambridge is a not-for-profit consultancy and linked exempt charity of the University of Cambridge. We work to achieve better health through the responsible and evidence-based application of biomedical science. We are a registered company, no. 5823194.
www.phgcambridge.org