



White Paper on Advancing Trusted Research Environments for Healthcare AI

Ivo Emanuilov, Björn Larsson, Andrew Dubber, Michela Magas
Industry Commons Foundation



INDUSTRY COMMONS

Executive summary

This white paper explores the evolution of Trusted Research Environments (TREs) as Europe implements the European Health Data Space (EHDS) Regulation, which mandates that all secondary use of health data must occur within secure processing environments. TREs are now central to health research, ensuring data confidentiality and legal compliance by offering controlled platforms for sensitive data analysis. However, existing TREs fall short when supporting advanced analytics, particularly for AI development, due to limited tools for handling big data, integrating complex workflows, and managing the secure release of AI models. These priorities are situated within a broader expectation that AI in healthcare must be grounded in transparent, accountable and human-centred practices. TREs must operate within a quadruple-helix constellation involving public authorities, industry partners, researchers, clinicians and civil society, including patients whose experiences inform the direction of healthcare AI. To support a level playing field, TREs must give these groups access to shared evidence and consistent insight into how models are developed, validated and monitored.

The paper identifies three critical areas for advancing TRE capabilities: (1) real-time AI oversight and explainability, (2) establishing shared validation toolkits (“Innovation Commons”) to standardise legal, ethical, and technical checks, and (3) enabling secure model transfer and ongoing post-deployment monitoring. Drawing on the Swedish **TRE4HealthAI** project, it highlights the need for enhanced automation, robust risk assessment during model training, and clear frameworks for validation and deployment. These areas must be addressed within institutional processes that recognise the importance of human participation, shared responsibility and documented decision pathways. The emergence of practices such as JUST Data (judicious, unbiased, safe and transparent) offers the procedural discipline needed for high-quality data documentation and annotation, giving researchers, clinicians, regulators and developers a shared foundation for understanding how interpretative choices shape model behaviour. Rather than a ‘humans-in-the-loop’ approach, which involves humans primarily as adjuncts to automated systems, this approach affirms an ‘*AI-in-the-loop*’ model in which human judgement remains central and AI plays a supporting role. This supports clearer reasoning, consistent communication and more reproducible scientific practice.

Addressing these gaps is essential to meet the EHDS’s regulatory requirements and enable trustworthy, data-driven innovation in healthcare. To support these expectations, the paper outlines a federated architecture for next-generation TREs that remains interoperable across HealthData@EU while providing the procedural clarity required for responsible collaboration. This includes the use of Transfer IP frameworks that define the rights and obligations of partners when models, code or datasets move between contributors. It also includes Pre-Flight processes, which establish structured conditions in which partners can test assumptions, examine edge cases and confirm workflow readiness before operational deployment. We contrast Sweden’s federated, decentralised approach with Finland’s centralised model for secure processing, using TRE4HealthAI and Swedish initiatives such as VAI-B as EHDS-aligned demonstrators of

how next-generation TRE capabilities can operate without new state-owned repositories. This comparative lens clarifies the boundary conditions for SPEs/TREs in Sweden and highlights how shared validation, export ‘airlocks’ and real-time oversight can be implemented in such an architecture.

The paper concludes with recommendations to modernise TREs with next-generation features, and to build infrastructures that are technically robust, ethically sound, and fully compliant with European data protection standards. This includes the integration of socio-technical practices that support explainability, reproducibility and informed human oversight. By combining regulatory alignment, technical development and structured collaboration, TREs can provide the basis for a trustworthy environment in which healthcare AI is safe, effective and capable of supporting high-impact clinical and research outcomes.

© Industry Commons Foundation 2025

Authors: Ivo Emanuilov, Björn Larsson, Andrew Dubber, Michela Magas

Table of Contents

Executive summary.....	2
Introduction	5
Background: TREs, security and regulatory drivers.....	7
TRE4HealthAI Use Cases.....	10
UC1 – Real-time monitoring and AI oversight with JUST Data annotation	10
UC2 – Shared validation toolkit of Innovation Commons and Transfer IP	12
UC3 – Deployment Stewardship and “Pre-Flight” validation airlock	13
Use Case Analysis and Synthesis.....	16
Use case 1	16
Background and rationale	16
Current developments & challenges	18
Roadmap to implementation	19
Use case 2	20
Concept and motivation	21
Alignment with regulatory and ethical standards	23
Implementing the Innovation Commons	24
Challenges and mitigations	26
Use case 3	27
Goals and key elements	28
Case example and workflow.....	30
Addressing challenges	31
Best practices and future opportunities	33
Conclusion and recommendations.....	34
Acknowledgements	37
Bibliography	38

Introduction

Secure data environments are becoming indispensable for health research and innovation. In Europe, the European Health Data Space (EHDS) Regulation now requires that any secondary use of health data occurs within a *secure processing environment*, a controlled platform ensuring legal compliance and data confidentiality.¹ Trusted Research Environments (TREs), also known as data safe havens, fulfil this role by allowing researchers to analyse sensitive data inside fortified “walled gardens” rather than by extracting data copies.² Under the Data Governance Act (DGA), a secure processing environment (SPE) is defined as a physical or virtual setting with organisational measures that uphold GDPR, protect intellectual property and confidentiality, and enable the provider to oversee all data processing actions (including display, storage, download, and algorithmic calculations of derived data).³ This model sharply contrasts with the old paradigm of disseminating data extracts, which duplicated privacy risks across many sites⁴. EHDS mandates will effectively make TREs *the default mechanism* for permitted data re-use across Europe, building on the principle already acknowledged in all surveyed EU countries that researchers should analyse health data in secure facilities.⁵

However, **current TRE implementations have limitations when tasked with advanced analytics, including artificial intelligence (AI) development.** A 2022 interview study across TRE operators found that while traditional statistical analyses are well-supported, “*next-generation*” requirements such as *handling diverse big data types, developing AI algorithms and software in-environment, and timely data import/export are not fully met by existing TREs*⁶. The study noted a *lack of tooling and automation* for these cutting-edge needs, with particular challenges around controlling the release of trained AI models from the secure environment⁷. In the UK, the landmark Goldacre Review similarly concluded that TREs are “the only realistic way” to enable greatly expanded health data research safely⁸, but also emphasised that TRE platforms must modernise to support open, reproducible data science and complex workflows

¹ Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and Amending Directive 2011/24/EU and Regulation (EU) 2024/2847 (Text with EEA Relevance) (2025), art. 73, <http://data.europa.eu/eli/reg/2025/327/oj/eng>.

² Kavianpour, S. et al. (2022) ‘Next-Generation Capabilities in Trusted Research Environments’, *Journal of Medical Internet Research*, 24(9), e33720; Goldacre, B. and Morley, J. (2022) *Better, Broader, Safer: Using Health Data for Research and Analysis*, Department of Health and Social Care.

³ Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act), 152 OJ L (2022), art. 2 (20), <http://data.europa.eu/eli/reg/2022/868/oj>.

⁴ Goldacre, B and Morley, J. (2022) ‘Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care’.

⁵ Kessissoglou, I.A. et al. (2024) ‘Are EU Member States Ready for the European Health Data Space? Lessons Learnt on the Secondary Use of Health Data from the TEHDAS Joint Action’, *European Journal of Public Health*, 34(6), pp. 1102–1108.

⁶ Kavianpour et al. (2022) ‘Next-Generation Capabilities in Trusted Research Environments’.

⁷ Jefferson, E. et al. (2022) GRAIMATTER Green Paper: Recommendations for Disclosure Control of Trained Machine Learning Models from Trusted Research Environments, Zenodo, available at: <https://doi.org/10.5281/zenodo.7089491>.

⁸ Goldacre, B and Morley, J. (2022) ‘Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care’.

(e.g., version-controlled code, linkages to external resources) that many early TREs did not accommodate. As a result, significant investments and initiatives are underway – for example, the UK government committed £200 million in 2022 to develop federated TRE infrastructure⁹ – to ensure these environments can meet both *regulatory obligations and research needs*. Recent mapping of 12 EU Member States found a heterogeneous landscape: all had some form of secure analysis facility, but *no country was yet fully ready to meet the then-upcoming EHDS requirements*, underscoring the need for capacity-building¹⁰.

The challenge of capacity building lies in the fractured nature of systems that focus on supporting isolated instances of specific AI applications, and regulating ex-post developers of AI applications for healthcare. As TRE infrastructures evolve, recent research suggests¹¹ that their design and operation must enable knowledge sharing between people throughout the health and research ecosystem, including clinicians, researchers, regulators, industry partners and patients who contribute lived experience and contextual knowledge. Beyond their technical and legal functions, TREs are human-centred environments in which multiple actors work together to guide the responsible use of health data. TREs must do more than satisfy regulatory compliance; they must also support the interpretative, collaborative and judgement-based work carried out by clinicians, researchers, data stewards and regulators, whose contributions shape how health data are understood and applied.

This white paper examines the next-generation capabilities that TREs must develop to support AI-driven health research and align with the EHDS Regulation rules that will gradually become applicable over the course of the next few years. Central to this transition is the recognition that healthcare AI development is fundamentally shaped by human decisions at every stage, from data annotation and curation to model interpretation, validation and regulatory review.

To structure the analysis, the paper examines three use case (UC) scenarios from the Swedish TRE4HealthAI¹² project that highlight the capabilities required of next-generation TREs. It outlines the framework conditions that shape national approaches and examines how these influence the readiness of health systems to support next-generation Trusted Research Environments. We present the regulatory and infrastructural background, after which the three use cases are described in detail, and their implications are synthesised in the use case analysis, which considers how TRE capabilities must evolve to meet both regulatory expectations and the practical needs of researchers, clinicians and industrial partners within a quadruple helix framework, before the final recommendations are set out.

⁹ Goldacre, B and Morley, J. (2022) 'Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care'.

¹⁰ Kessissoglou et al., 'Are EU Member States Ready for the European Health Data Space?'

¹¹ Hager, A., Emanuilov, I. (2025) 'Trusted Research Environments for Healthcare AI: Sweden - Finland Comparison', Industry Commons Foundation

¹² Trusted Research Environment for Health AI Advancement (TRE4HealthAI) is funded as part of the Advanced Digitalisation programme by Vinnova – the Swedish Innovation Agency. Project number 2024-01412. <https://www.vinnova.se/en/p/trusted-research-environment-for-health-ai-advancement/>.

- **UC1** focuses on real-time monitoring, knowledge sharing, collaboration and oversight during AI development, incorporating visibility into how human-led “**JUST Data**” annotation and modelling choices affect performance and risk. It pioneers **the inclusion of regulatory stakeholders** from the earliest stages of development, a novel approach for TREs.
- **UC2** demonstrates the value of a shared “**Innovation Commons**” validation toolkit that provides standardised legal, ethical and technical resources to support reproducible evaluation grounded in human judgement, with AI contributing as a support mechanism, including in the application of “**Transfer IP**” principles to clarify rights and obligations around model use and reuse.
- **UC3** examines secure model transfer, a “**Pre-Flight**” system and “**Deployment Stewardship**”, including controls and mechanisms for continuous performance tracking after deployment.

These use cases together illustrate how TREs can support the entire lifecycle of healthcare AI in an EHDS-aligned, federated infrastructure that prioritises human autonomy in a technology-supported and regulation-enabled framework, rather than a top-down, technocratic model.

The following section outlines the framework conditions that shape national approaches and examines how these influence the readiness of health systems to support next-generation Trusted Research Environments. This is followed by an analysis of the requirements that arise from the project’s use cases, which demonstrate how TRE capabilities must evolve to meet both regulatory expectations and the practical needs of researchers, clinicians, and industrial partners.

Background: TREs, security and regulatory drivers

Before diving into the use cases, it is important to understand the backdrop of TREs and how they intersect with emerging laws. This regulatory backdrop establishes both the constraints and the opportunities for next-generation TREs to evolve into more collaborative, socio-technical spaces rather than remaining purely compliance-driven infrastructures.

TRE foundations: the “Five Safes” and beyond. Most TREs implement governance based on the “Five Safes” framework – Safe People, Projects, Settings, Data, and Outputs – to minimise misuse of sensitive data¹³. In practice, this means only vetted researchers (Safe People) can access de-identified data (Safe Data) within a secure platform with strict controls (Safe Setting) for clearly approved purposes (Safe Projects), and any results leaving the environment are checked for privacy (Safe Outputs). Traditional TREs provide remote analysis workspaces (often virtual desktops

¹³ DRAGoN, University of the West of England (n.d.) *The Five Safes*, accessed 7 November 2025, <https://fivesafes.org/>.

or Jupyter/RStudio environments) with no internet access and an “airlock” mechanism for importing data or exporting aggregate results via approval processes. These measures have successfully enabled privacy-preserving research on health records, as seen in the UK’s NHS Digital TRE and Finland’s Findata environment, while maintaining public trust.¹⁴

Secure processing under EHDS. The EHDS makes such controlled environments not just best practice but a *legal requirement*. Article 73 of the EHDS Regulation stipulates that access to electronic health data for secondary use shall only be granted through a secure processing environment operated by an authorised health data access body¹⁵. In essence, researchers in the EU will no longer receive raw datasets; instead, they will perform analysis within designated national or cross-border TREs connected via the HealthData@EU network. The Data Governance Act, which has been in force since September 2023, already defines the baseline for these environments and implies public sector oversight of any data use within them. Every operation – viewing, processing, exporting data or *derived outputs* – must be under the control and audit of the environment provider. This regulatory push is spurred by lessons from the TEHDAS (Towards EHDS) initiative, which found uneven readiness in Member States but a clear trend toward centralising data access for privacy reasons.¹⁶ Many countries have either established or are piloting TRE-like platforms for researchers (for example, France’s Health Data Hub or Germany’s planned Forschungsdatenzentrum) to comply with the anticipated EHDS obligations. The goal is a federated network of TREs enabling *transnational research without moving data*, an approach that aligns with approaches such as federated analytics and learning.

Member States’ current TRE implementations reflect their health data governance structures, illustrating the challenge of aligning TRE capabilities across jurisdictions and underscoring the need for flexible, interoperable frameworks under the EHDS. Finland has adopted a centralised model under its Act on the Secondary Use of Health and Social Data (2019), establishing a single health data permit authority (Findata) that operates a nationwide secure environment (the Kapseli platform) for all approved secondary use of health data¹⁷. Sweden, by contrast, is developing a more decentralised, federated approach. In 2022, the Swedish eHealth Agency (E-hälsomyndigheten) was tasked with planning a national health data infrastructure that connects existing regional and sectoral data holders within a federated secure environment model, rather than concentrating all data in one hub¹⁸. A recent

¹⁴ Goldacre, B and Morley, J (2022) ‘Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care’; Kessissoglou et al. (2024) ‘Are EU Member States Ready for the European Health Data Space?’

¹⁵ Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and Amending Directive 2011/24/EU and Regulation (EU) 2024/2847 (Text with EEA Relevance), art. 73.

¹⁶ Kessissoglou, I.A. et al. (2024) ‘Are EU Member States Ready for the European Health Data Space?’

¹⁷ Ministry of Social Affairs and Health (n.d.) Act on the secondary use of health and social data, <https://stm.fi/documents/1271139/1365571/The+Act+on+the+Secondary+Use+of+Health+and+Social+Data/a2bca08c-d067-3e54-45d1-18096de0ed76/The+Act+on+the+Secondary+Use+of+Health+and+Social+Data.pdf?t=1559641328000>.

¹⁸ ‘Färdplan för nationell digital infrastruktur (Interpellation 2024/25:742 av Anna Vikström (S))’ (2025), available at: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/interpellation/fardplan-for-nationell-digital-infrastruktur_hc10742/.

government inquiry explicitly framed health data infrastructure as a national interest and recommended coordinated efforts to link institutions under the EHDS framework¹⁹. Sweden could have taken a very different path had it adopted the E-Health Authority's 2022 proposal for a state-run national data space for medical imaging²⁰. That plan envisioned a centralised data sharing service under government control, combining national storage, federated access, and secure processing for care and research. A mammography pilot would have established state-managed infrastructure, legal mandates for access, and potentially mandatory provider participation, i.e., moving Sweden toward an architecture resembling that of Finland, with unified data handling and nationalised ownership.

At this point, Sweden has not yet created legal mandates for SPEs or TREs, launched a regulatory sandbox for AI in healthcare, or defined national governance or operational frameworks for model pipelines or validation. Next-generation TREs could therefore offer concrete technical, regulatory and human-centred contributions to what Sweden is now structurally preparing, supporting the shared judgement and collaborative responsibility that guide decisions about data use and model development. If positioned strategically, they could serve as demonstrators or pilots for Sweden's EHDS adaptation. In this context, the TRE4HealthAI project serves as a demonstrator for EHDS alignment in Sweden – showcasing next-generation TRE capabilities in a federated setting that can inform national implementation.

Next-generation demands – AI, big data, and real-time collaboration. While compliance and security are paramount, TREs must also evolve to serve modern data science, particularly AI development, which depends on continuous human involvement for interpreting data, diagnosing modelling issues and guiding methodological decisions. Conventional TRE setups were often designed for epidemiological analyses on relatively structured data (e.g., CSV files of clinical records) and static statistical outputs.²¹ In contrast, **AI research introduces new technical requirements.** It involves handling very large and unstructured datasets across imaging, genomics and text, which call for high-performance computing and substantial storage; supporting machine learning libraries (e.g., TensorFlow, PyTorch) and GPU-accelerated training; iterative, code-intensive workflows with software development tools; and the need to evaluate complex model outputs, including cases in which the output is a trained set of neural network weights. A state-of-the-art review revealed that *most current TREs do not yet support these capabilities out of the box*, often due to security-driven constraints (e.g., blocking the internet means no access to external code repositories) or simply because such features were beyond the original scope. For example, analysts have reported difficulty using version control systems inside TREs because connectivity is locked down²², and *no standard process exists to*

¹⁹ Sverige (2024) Delad hälsodata – dubbel nytta: regler för ökad interoperabilitet i hälso- och sjukvården, Regeringskansliet.

²⁰ E-hälsomyndigheten (2022) Förstudie om ett statligt, nationellt datautrymme för bilddiagnostik (S2021/05259 delvis).

²¹ Kavianpour, S. et al. (2022) 'Next-Generation Capabilities in Trusted Research Environments'.

²² Goldacre, B. and Morley, J. (2022) 'Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care'.

safely export machine learning models, which could inadvertently contain sensitive patterns²³.

At the same time, **regulators and data stewards are pushing for TREs to embed greater transparency, auditability, and automation**. The UK’s Goldacre Review recommended that all code used in TREs be shareable and ideally open by default to improve reproducibility and confidence in research outputs²⁴. Projects including GRAIMATTER (Guidelines and Resources for AI Model Access from TREs) have highlighted the risk of AI models acting as a “Trojan horse” for leakage – complex models can memorise data points, so TREs require new *disclosure controls for trained models*, akin to how they scrutinise statistical outputs²⁵. GRAIMATTER delivered a set of draft recommendations in 2022 to guide TRE providers on additional checks (such as scanning model weights for signs of data memorisation, using differential privacy techniques, or requiring model factsheets documenting training data) before allowing models out. Likewise, the FUTURE-AI initiative (an EU-wide consensus on trustworthy AI in medical imaging) calls²⁶ for continuous monitoring of AI performance and robustness as part of the AI lifecycle. This is particularly important as TRE platforms could facilitate such monitoring, especially if they serve as testing environments in future **regulatory sandboxes** for experimental algorithms.

In summary, TREs are at the nexus of *compliance* and *human-centred technological innovation*. They must simultaneously **harden** data protections to satisfy the EHDS requirements (ensuring that no personal data leaks and all usage is supervised) **and broaden capabilities** so that data scientists, clinicians, and even regulators can collaboratively develop and evaluate next-generation AI solutions within them. While the Five Safes provide the structural backbone of TRE governance, their effectiveness relies heavily on the judgment, interpretation and interaction of the people who work within them. TREs are therefore not only technical compliance mechanisms but also social systems in which researchers, clinicians, data stewards, regulators and patients collectively shape how health data are used and understood. The following sections outline the three use cases (UC1, UC2, UC3), analyse how they address these dual priorities and describe the concrete features and practices that are emerging to meet the challenge.

TRE4HealthAI Use Cases

UC1 – Real-time monitoring and AI oversight with JUST Data annotation

This use case addresses the development phase of AI models within a TRE, introducing live oversight and risk management during model training. An integral component of UC1 is the recognition that human interpretation influences not only regulatory

²³ Jefferson, E. et al. (2022) *GRAIMATTER Green Paper*.

²⁴ Goldacre, B. and Morley, J. (2022) ‘Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care’.

²⁵ Jefferson, E. et al. (2022) *GRAIMATTER Green Paper*.

²⁶ Lekadir, K. et al. (2025) ‘FUTURE-AI’.

oversight but also the data on which models are trained. Real-time monitoring in the TRE incorporates visibility into how human annotation and curation choices affect model behaviour over time. This connection between human input and model response allows bias, drift or inconsistency introduced during annotation to be observed and managed in situ. Through integration with the JUST Data framework,²⁷ the TRE can log annotation provenance, trace decision patterns among annotators, and identify where variations in labelling may correlate with downstream model disparities. The goal is to embed tools in the TRE that enable **continuous, real-time monitoring of an AI model's performance, bias, and other risk indicators**. Instead of training models in a black box, researchers and designated overseers (e.g., data protection officers or regulators) would have dashboards and alerts to catch issues early, such as detection of model drift, unstable behaviour, or emerging bias as the model learns. Such insight supports the JUST principles by treating human contribution as a measurable, auditable part of the AI development lifecycle.

This use case also pioneers the involvement of regulatory or ethical stakeholders *during* development, which is novel for TREs. In this white paper, the term “regulators” is used as shorthand for a range of supervisory actors with different mandates, including health data access bodies, data protection authorities, medical device and AI regulators, research ethics committees and, in some cases, hospital or regional governance boards. These actors do not form a single hierarchy, but contribute complementary perspectives on safety, legality and the wider public interest. Within a TRE, the aim is not to give any one group unilateral control, but to allow these supervisory roles to engage with researchers, clinicians and data providers on the basis of a shared evidential record. By giving regulators or ethics boards a window into the model's training process, the TRE enables them to work alongside clinicians, researchers, and developers, while gaining insight into how models evolve without accessing raw, sensitive data. This configuration supports a shared oversight process in which all parties can understand emerging issues simultaneously and contribute their respective knowledge collaboratively. For example, a medical AI regulator could follow training metrics or bias indicators within the TRE's monitoring interface and discuss potential concerns with the development team, helping to shape decisions through dialogue rather than a post-hoc review.

Concretely, UC1 implements features such as: integrated performance metrics (accuracy curves, etc.), automatic bias detection and fairness auditing tools, and explainability modules that can analyse why the model is making certain predictions. Embedding these in the TRE means issues can be identified and addressed **before** the model is finalised or deployed. For example, if a model training on hospital data starts to underperform on a subgroup of patients, the TRE could flag this to researchers and log it for regulators' review. Alongside these technical components, UC1 incorporates the JUST Data approach to enable all participants to contribute contextual information

²⁷ Emanuilov, I. and Magas, M. (2024) White Paper on Data Documentation for JUST Data Practices, Industry Commons Foundation, available at: <https://doi.org/10.5281/zenodo.14228289>. JUST Data Annotation was funded as part of the Advanced Digitalisation programme by Vinnova, project reference 2023-03231, available at: <https://www.vinnova.se/en/p/just-data-annotation/>. See also JUST Data in Magas, M. and Dubber, A. (2020) 'Expanding EOSC: Engagement of the Wider Public Sector and Private Sectors in EOSC', Zenodo, <https://doi.org/10.5281/zenodo.4463437>.

that clarifies how data are interpreted and annotated. By enabling annotators, clinicians, developers and regulators to read, document and reflect on the assumptions and decisions that inform each stage of the training process, the TRE supports behaviour that is judicious, unbiased, safe and transparent. This ensures that performance signals can be understood not only in computational terms but also in light of the human choices that shaped the underlying data. The emphasis is on **safety and transparency during AI development**, aligning with emerging regulatory expectations for continuous risk management in AI.²⁸ and the FUTURE-AI initiative²⁹ recommends continuous performance monitoring of medical AI models. UC1 directly contributes to those aims by equipping the TRE with the capability to perform such oversight within a secure technology sandbox.

By recommending the implementation of UC1, TRE4HealthAI aims to highlight that real-time oversight is feasible and beneficial. Success for this use case means researchers can develop models with greater confidence (knowing that any compliance or performance issues will be caught early), and regulators gain a new level of transparency into the AI development process. In short, UC1 turns the TRE into **not just a safe data environment**, but a proactive **“AI watchdog”** that guards against model risks during creation.

UC2 – Shared validation toolkit of Innovation Commons and Transfer IP

This use case focuses on the validation phase – ensuring that once an AI model (or dataset) is developed, it is thoroughly evaluated for performance, fairness, and compliance using standardised methods. Today, validation is often siloed: each team cobbles together its own test datasets, evaluation code, and legal reviews, leading to duplicated effort and inconsistent quality. UC2 proposes an **“Innovation Commons”**³⁰ within the TRE: a shared suite of validation resources and tools accessible to all projects. This Commons includes legal, ethical, and technical components to streamline how models are validated and prepared for sharing or deployment.

In practice, the **Innovation Commons** would offer reusable test datasets (including synthetic data for privacy-safe testing), standardised validation protocols and metrics, documentation templates (e.g., model fact sheets, datasheets or audit checklists), and template legal agreements to govern data and IP usage and generation. For example, if a team validates an AI model for diabetic retinopathy with a certain imaging dataset and produces a results report, they could contribute a **“validation bundle”** to the Innovation Commons so that others can reuse that methodology. Another research

²⁸ Notably, the EU AI Act calls for risk and bias monitoring in high-risk AI systems Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024/1689 (2024), art. 10, <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.

²⁹ Lekadir, K. et al. (2025) ‘FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare’, BMJ, e081554.

³⁰ Svensberg, L., Lindvall, J., Danielsson, P., Nilsson, A.G. et al. (2024) Vinnova Policy Paper for Commons for Development of More Efficient Innovation Ecosystems, available at: https://digitalwellarena.se/wp-content/uploads/2024/05/policypaper_commons_eng.pdf.

group building a similar model could then take this bundle (perhaps including code for evaluation and a summary of results) and apply it to their own model, allowing like-for-like comparison without reinventing the wheel. This greatly lowers the barrier for new teams to rigorously test their AI, and promotes **standardisation** in how things are validated across the TRE. Importantly, if everyone validates using common reference measures, **regulators can more easily compare and trust the results** across different AI models.

A novel aspect of UC2 is addressing the **legal and IP clarity** around sharing and reusing outputs in the TRE. The Innovation Commons will embed template agreements covering questions like: “*Who owns a model trained on Hospital X’s data?*”, “*Can this model be used commercially and under what licence?*”, or “*What restrictions apply to data-derived assets?*”. For instance, a standard agreement could grant joint IP ownership to the data-providing hospital and the model developer, or require that any exported model carry certain use restrictions, such as behavioural use conditions.³¹ By having these templates built in, users of the TRE can easily apply them as they validate and export models, ensuring **everyone knows the rules of the game upfront**. We call this the “**Transferable IP**” (**Transfer IP**³²) concept – explicitly managing intellectual property and usage rights as part of the validation/export workflow. This clarity protects data contributors (e.g., a hospital can mandate that a model can only be used for non-commercial research unless otherwise licenced) and gives researchers confidence about how they can later use or commercialise their work.

To summarise, UC2 transforms the TRE into a **collaborative testing ground**. Instead of each project working in isolation, the TRE provides a common set of tools and knowledge that grows over time to become a library of validated models, benchmark datasets and related resources. This not only makes individual projects more efficient, but also drives **higher standards** – if everyone uses a shared **validation toolkit that aligns with regulatory and ethical norms**, then all AI outputs from the TRE are more likely to meet approval criteria. In fact, the Innovation Commons can act as a “*pre-certification*” mechanism: if a model passes all the standardised checks in the TRE, it could indicate readiness for regulatory approval, thereby potentially shortening its time to market. UC2’s benefits are long-term as well, creating a foundation for ongoing innovation where each new project can build on the last. It strongly supports **responsible AI reuse** and continuous improvement, while respecting confidentiality and IP through built-in governance.

UC3 – Deployment Stewardship and “Pre-Flight” validation airlock

This use case tackles the end of the pipeline: moving a validated model out of the TRE into the real world (such as a hospital IT system or a cloud service) *and* ensuring the model is monitored over time after deployment. This is crucial because currently, even

³¹ Danish Contractor et al. (2022) ‘Behavioral Use Licensing for Responsible AI’, *2022 ACM Conference on Fairness, Accountability, and Transparency*, 21 June 2022, 778–88, <https://doi.org/10.1145/3531146.3533143>.

³² Magas, M., Radziwon, A., Altosaar, A., Wretblad, L., Emanuilov, I., Bertels, N. (2022) White Paper: IP and Industry Agreements towards Industry Commons. Industry Commons Foundation. Available at: <https://zenodo.org/records/14566936>

if a model is developed in a TRE, there is a gap when it leaves that safe environment – how do we ensure it carries no sensitive data, is properly documented, and stays accountable during actual use? UC3 introduces a “**Pre-Flight**³³ **validation airlock**” for model export and a framework for **post-deployment monitoring**, making the transition secure and compliant.

In practical terms, when a team is ready to deploy an AI model developed in the TRE, they will use an Export function that packages the model with all necessary documentation and approvals. In practice, this goes far beyond just downloading a file. The TRE’s export pipeline will perform a series of checks and attach a comprehensive model release bundle. Key elements of this bundle include:

- **The model artefact** (e.g., the model weights or a container image) in a standardised, portable format (such as ONNX or a Docker container) to ensure it can run in any common environment. The system can auto-convert models if needed to these formats for interoperability.
- **Detailed documentation**, akin to a “qualification dossier” or model factsheet, that describes how the model was developed, on what data (analysed with JUST Data tools), with what performance, etc.³⁴ This likely compiles information from the training and validation phases (UC1’s monitoring logs and UC2’s validation results) into a human-readable report. Regulators reviewing the model will expect this evidence.
- **Personal data safety check** confirming that the model does not contain any disclosive information. For example, the TRE could scan the model weights to ensure it hasn’t memorised any patient records or includes any direct identifiers (e.g., using membership inference attacks). Only if the model is deemed “clean” (no GDPR-covered data embedded) can it be exported. This satisfies the European Health Data Space Regulation’s mandate that only non-sensitive outputs leave the secure environment.
- **Legal/IP clearance** forms that encapsulate the **Transfer IP** agreements from UC2. Essentially, an export certificate would be generated, stating who owns the model, under what licence it can be used, and confirming that deployment has been authorised under the data sharing terms. For example, it might say: “*This model was trained on Hospital X’s data and is co-owned by Hospital X and University Y. It is approved for use in Hospital X’s clinics. Commercial use requires separate licensing.*” Having this in writing prevents later disputes and ensures all stakeholders (developers, hospitals, regulators) are on the same page about usage rights.
- **Provenance and integrity audit** artefacts, such as cryptographic hashes or signatures, to prove that the model leaving is exactly what was built inside the TRE. This provides a trust mechanism, so if someone outside wants to verify the

³³ Björling, S-E. (2024) ‘Proposal for Pre-Flight for Health’, Industry Commons Foundation, presented at ‘Synthetic and Health Data Day’, JUST Data Industry Week, Linköping Science Park: <https://mtflabs.net/just/linkoping-programme/>

³⁴ Emanuilov, I. and Magas, M., (2024) *White Paper on Data Documentation for JUST Data Practices*.

model hasn't been tampered with, they can validate the signature. It also proves the model came from the controlled environment, which can be important for liability and trust.

This “airlock” process ensures that by the time an AI model is deployed in a live setting, it carries with it a full compliance and performance pedigree. It addresses regulatory requirements such as those under the EU Medical Device Regulation (MDR), which increasingly expects manufacturers to provide comprehensive technical documentation and post-market surveillance plans for AI tools. In fact, part of UC3 is to integrate **Deployment Stewardship**: once the model is in use (say at a hospital), the TRE framework doesn't simply let it go entirely. It sets up a channel (if the deployer consents) for ongoing performance data to feed back into the TRE or to a regulator's dashboard. This can be achieved by deploying a lightweight monitoring agent with the model that collects non-sensitive telemetry – e.g. how often the model is used, summary of outputs, accuracy on new cases (if outcomes are known) – and sends periodic reports. No patient data is sent back, only aggregate metrics or drift indicators. The TRE or relevant oversight body can then track if the model's performance degrades over time or if it starts being used in scenarios outside its original scope. This aligns strongly with the concept of “**continuous monitoring**” in AI governance; for instance, the EU AI Act and MDR both require monitoring AI systems in real-world usage and managing any emerging risks.

By extending the TRE's involvement to the deployment phase, UC3 closes the loop in the AI lifecycle. It provides confidence to end-users (for instance, clinicians) and regulators that an AI model coming out of the TRE isn't a wild card – it's a vetted, transparent, and trackable product. A hospital adopting a model via this process can be given a “**health report**” of the model periodically and know that if something goes wrong, there is a clear line of traceability back to how and when it was developed. Essentially, UC3 positions the TRE as not only a **development sandbox** but part of a continuous **quality-assurance system for AI in healthcare**. This significantly elevates trust in AI models: stakeholders know that the model was developed under rigorous and continuous shared scrutiny (UC1), validated to high standards (UC2), exported with all safeguards (UC3's airlock), and will be monitored continuously – a cradle-to-grave approach for responsible AI.

Together, these use cases address the **full lifecycle of an AI model in a TRE** – from development and validation to deployment and continuous oversight – expanding the TRE's role into a *trusted research and regulatory environment*. We review the state-of-the-art for each scenario, discuss challenges and relevant work (drawing on recent European projects and literature), and recommend best practices to operationalise these capabilities within EHDS-compliant infrastructures. By strengthening TREs with such next-generation features, Europe can facilitate data-driven innovation, including clinical AI, in a manner that is technically robust, ethically sound, and legally compliant with data protection and forthcoming EHDS requirements.

Use Case Analysis and Synthesis

This section synthesises the implications across the three use cases and identifies the cross-cutting capabilities that next-generation TREs will require. While each use case addresses a distinct phase of the AI lifecycle, together they outline a coherent model of how TREs can integrate technical safeguards, human judgment and regulatory expectations into a single workflow. The analysis reflects the principle that trustworthy healthcare AI emerges not from isolated procedures, but from an environment in which development, validation and deployment are shaped by shared evidence, consistent documentation and collaborative oversight. The tools for real-time oversight in UC1 lay the evidential groundwork for the shared validation practices in UC2, and both are prerequisites for the structured pathways for deployment and stewardship in UC3. What follows is a synthesis that shows how these elements form a continuous socio-technical foundation for EHDS-aligned trusted research environments in which models are developed, evaluated, exported and monitored within a coherent governance framework that reflects an ‘AI-in-the-loop’ approach, where AI supports and strengthens human judgement, rather than the reverse.

Use case 1

Modern AI models are not static artefacts – their development is an iterative process with significant risks that can emerge *during* training or testing. The first use case (UC1) envisions a TRE with built-in capabilities for **real-time monitoring, risk assessment, and explainability** to support safe and transparent AI model development. The aim is to detect problems early (e.g., a model becoming biased or unstable) and to provide *continuous oversight* that aligns model development with safety and accountability standards. UC1 also involves **regulatory or supervisory stakeholders**, which is novel for any TRE setup. By embedding oversight tools in the TRE, this diverse set of regulators can remotely observe and collaborate on AI development, rather than only assessing models after they are fully developed. This supports a *proactive compliance approach* and builds an evidential foundation for eventual regulatory decisions.

Background and rationale

The challenge of model drift and bias. In healthcare AI, models trained on a given dataset can perform unpredictably when data or conditions change – a phenomenon known as *concept drift*. For example, a diagnostic model could gradually become less accurate if new patient populations or different medical imaging devices appear over time. Traditionally, such issues would only be discovered post-deployment, potentially causing harm or necessitating costly recalls. TREs can help shift this detection to an earlier stage. By introducing JUST Data annotation for imported datasets and monitoring training in real time, the TRE can flag warning signs, e.g., if a model’s validation accuracy suddenly drops or if error rates differ markedly across subgroups (indicating potential bias). Crucially, these signals are intended to support human interpretation rather than replace it; annotators, clinicians, developers and regulators can use the contextual information recorded through JUST Data to understand why performance is changing and how to respond. The concept is analogous to live “telemetry” for model development, but understood as a set of cues for people to assess rather than an automated decision process. This is increasingly recommended

in AI risk management. For instance, the FUTURE-AI guidelines emphasise continuous evaluation of AI for **robustness, fairness, and transparency** throughout the lifecycle³⁵. A real-time dashboard in the TRE would compile relevant metrics, such as performance by cohort, data drift measures, and algorithmic explanations, giving researchers and regulators immediate information that they can collectively interpret and discuss when judging whether a model is behaving as expected.

Explainability and accountability built in. Another motivation is to ensure that models being built are not “black boxes” to regulators. A TRE would integrate explainable AI (XAI) tools directly into the TRE environment. For example, if researchers are training a neural network on medical images, the TRE could generate saliency maps or feature attributions on the fly to show which image regions influence the model’s predictions. By logging these as part of JUST Data annotation processes and making them visible, the TRE turns model training into a transparent process – accessible not just to the data scientists but also to compliance officers who may not be AI experts. This aligns with emerging regulatory expectations (e.g., the EU AI Act requires transparency and human oversight for high-risk AI systems). In practice, it means the TRE would host libraries for JUST Data fairness assessment, bias detection (e.g., checking output distributions across genders or ethnic groups), and error analysis as first-class features.

Regulator involvement towards building a “sandbox”. A key innovation is to treat the TRE as not only a research space but also a testing environment for a **regulatory sandbox**. This means regulators and competent authorities are given *read-only* access to monitor ongoing projects within the TRE. In the context of EHDS, a national health data access body or a medical device regulator could use this to supervise that an AI system being developed on patient data respects the approved use and meets safety thresholds. By doing so, compliance issues can be caught early, innovation can be de-risked quickly, and regulators themselves can gain familiarity with the model’s properties, making final evaluation more evidence-based. The UK’s NHSX has advocated for such models of co-development, and the TRE4HealthAI project explicitly aims to support “an AI regulatory sandbox” via these oversight features. Embedding so-called regulatory users transforms the TRE into a *Trusted Research and Regulatory Environment*. It is worth noting that Recital 15 of the DGA already suggests that analyses in secure environments “should be supervised by the public sector body”³⁶; we suggest that development efforts should focus on providing the tooling to make such supervision practical and real-time.

From a policy standpoint, the notion of embedding regulators in the TRE must adapt to different national governance models. In a centralised scheme like Finland’s, oversight is inherently channelled through the single data access body (Findata), which supervises all secondary use projects and their secure analytics environment. In Sweden’s more decentralised landscape, regulatory oversight may need to be shared or coordinated across multiple TRE nodes (for example, regional health authorities or domain-specific agencies) under an overarching framework. This makes the TRE4HealthAI sandbox approach – where regulators participate in a federated TRE environment – a valuable test case for establishing real-time oversight in a distributed

³⁵ Lekadir, K. et al. (2025) ‘FUTURE-AI’.

³⁶ Regulation (EU) 2022/868 (Data Governance Act) (2022).

setting. By involving supervisory authorities at the platform level, even in a federated model, TREs can ensure continuous compliance despite a fragmented data governance structure.³⁷

Current developments & challenges

Implementing these capabilities faces both technical and policy challenges.

In terms of **technical feasibility**, the main question is: *How to perform continuous monitoring without compromising security or performance?* TREs traditionally isolate researchers heavily – no internet, limited compute. Yet next-generation TREs would likely require heavy computations (e.g., calculating drift statistics after each training epoch, or generating XAI visualisations) and possible streaming updates to an external dashboard. One approach is to keep a dashboard within the secure environment (accessible to users via remote desktop or a web portal) so that data never leaves. Regulators would log into the TRE’s portal to view the monitoring panels. Ensuring that this does not degrade the user experience is key – the environment must be equipped with sufficient computational resources to handle both model training and monitoring tasks in parallel. Cloud-based TRE implementations are addressing this by leveraging scalable infrastructure. For instance, a GPU cluster can be provisioned for model training, while a separate container continuously reads logs and model outputs to update metrics charts.

In terms of **data protection** concerns, adding more monitoring means generating more *metadata* about the sensitive data and models. Could these reveal patient information? For example, a drift detection could involve comparing current data statistics to original data stats; if handled improperly, one might inadvertently display personally identifying distribution changes. The solution is to focus on aggregate, non-identifying metrics (means, performance percentages) and treat them as *non-sensitive outputs* allowed for internal use. All monitoring outputs would still be subject to output checking if they were to be exported outside the TRE. The GRAIMATTER recommendations are again relevant, as they note that even summary statistics used to evaluate models should be scrutinised for disclosure risk³⁸. In a TRE context, since all monitoring stays inside, it’s acceptable, but a robust audit trail is needed so that any findings or plots that eventually go into a publication would be reviewed.

In terms of **human factors and workflow**, having regulators peer into a live research project is a new way of working. Both researchers and regulators will need clear protocols. For example, how should a regulator intervene if they spot a major issue during model training? Do they communicate with the researchers via the TRE’s communication tools? Do they have the authority to pause an experiment? These governance questions must be settled in a sandbox agreement. The involvement is not framed as supervision from above, but as a shared process in which regulators, researchers, clinicians and annotators learn from one another’s reasoning. The inclusion of JUST Data annotation practices strengthens this by giving all participants

³⁷ See recommendations for a federated JUST Data infrastructure as part of the European Open Science Cloud. EOSC Strategic and Research and Innovation Agenda (SRIA) 2021–2024. <https://eosc.eu/eosc-about/sria-mar/> The recommendations were based on research performed in Magas, M. and Dubber, A. (2020) Expanding EOSC, Zenodo.

³⁸ Jefferson, E. et al. (2022) *GRAIMATTER Green Paper*.

access to the same contextual information about annotation choices, decision assumptions and uncertainties. This allows regulators to understand how the data have been shaped and interpreted, and enables researchers to understand how regulatory concerns arise, creating a more symmetric dialogue. Early trials suggest a *light-touch oversight* is preferable where regulators primarily observe and only intervene if something clearly conflicts with an ethical or legal requirement. In other cases, their reflections, informed by the contextual material documented through JUST Data, can be discussed with the researchers as part of an ongoing interpretative dialogue rather than a unilateral judgement.

Despite challenges, some **best practices** could include the use of open source monitoring frameworks (so they are inspectable for security), ensure all logged metrics are non-disclosive, involve end-users (clinicians, patients) in defining what “safety” metrics to track, and integrate with risk management standards (e.g. ISO/IEC 23894 on risk management for AI).

Roadmap to implementation

To realise these new requirements, a TRE would need to incorporate several features and processes:

- **Real-time dashboard.** This could be a web-based dashboard within the TRE that displays ongoing experiment metrics. For example, in TRE4HealthAI we call for a design of a real-time monitoring dashboard that allows researchers and regulators to observe and track the performance of AI models. This includes charts of accuracy vs. time, confusion matrices updated with each batch, data drift indices (e.g. population stability index), and system metrics (GPU utilisation, etc.). Off-the-shelf solutions such as TensorBoard (commonly used for monitoring deep learning) can potentially be embedded in a secure mode, showing model training curves without exposing raw data. To support meaningful interpretation, the dashboard can incorporate continuous annotation following JUST Data guidelines, allowing participants to record contextual information about training decisions, unexpected behaviours or emerging concerns. These annotations give others a shared basis for understanding the metrics they see and help create a collaborative space where regulators and researchers can discuss observations on equal footing rather than treating the dashboard as an automated reporting tool.
- **Risk analytics & alerts.** The TRE should implement an automated risk scoring system. For instance, it could use predefined rules, such as: if validation performance drops by more than $X\%$ or if bias metric exceeds Y , flag it. These triggers can prompt notifications in the JUST Data system. In practice, an email or TRE-internal message could be sent to project owners and oversight persons when such conditions occur, prompting them to investigate.
- **Integrated documentation.** Every run of an experiment in the TRE should generate an audit record. This includes parameters used, dataset versions, and any notes by the researcher explaining that run. Through the JUST Data annotation system, researchers, clinicians, developers, regulators and, in

appropriate circumstances, patients can contribute commentary that records the contextual reasoning behind particular decisions or observations, ensuring that documentation reflects the perspectives of all participants rather than a single authoritative view. This creates a rich provenance trail in which technical metrics sit alongside human interpretation and experience. Such entries are stored as part of the shared annotation record and can be revisited, compared or discussed as the model evolves. By the end of development, this trail can be compiled as evidence for regulatory submission – essentially a living *Model Development Report*. It should align with what regulators will ask for in approval, e.g., under the EU Medical Device Regulation (MDR) or the AI Act; both require showing how the model was developed and tested.

- **Privacy-preserving XAI.** If explainability tools such as partial dependence plots or example outputs are shown, the TRE provider must ensure they are aggregated. One approach is to use *representative synthetic examples* rather than actual patient data to illustrate model behaviour. These synthetic examples should be reviewed using JUST Data guidelines so that clinicians, researchers, developers and regulators can interpret whether any patterns suggest emerging bias or drift. A future TRE could explore generating synthetic test data under controlled conditions for bias and drift within the secure environment for visualisation purposes, providing shared reference cases that allow meaningful discussion without exposing real patient records.

Overall, the purpose is to move TREs from passive data vaults to *active safety enablers*. By instrumenting the AI development process with oversight capabilities, TREs not only help researchers build better models (faster feedback, error catching) but also create the trust and evidence needed for those models to be responsibly deployed. In the next section, the second use case complements this by focusing on the tools and rules that ensure validation and re-use of AI models can happen efficiently and lawfully within the TRE.

Use case 2

Where UC1 focuses on how models are created and monitored within a TRE, UC2 addresses what must happen once that development work stabilises. The insights, annotations and oversight gathered during UC1 become the inputs for UC2's shared validation workflows, enabling the TRE community to test, compare and document models using common resources.

Validating AI models, i.e., assessing their performance, fairness, and compliance, is often an arduous and siloed effort. UC2 proposes a solution: an “**Innovation Commons**”³⁹ inside the TRE that provides a *shared toolkit of validation resources* for legal, ethical, and technical checks. The Innovation Commons bridges proprietary and public interests by lowering barriers to access standardised validation methods, while clearly delineating intellectual property (IP) and data protection guardrails for secondary use. This responds to the fragmented and duplicative efforts often seen in AI

³⁹ Potts, Jason, *Innovation Commons: The Origin of Economic Growth* (New York, 2019; online edn, Oxford Academic, 22 Aug 2019), <https://doi.org/10.1093/oso/9780190937492.001.0001>, accessed 13 Nov. 2025.

development, where each team reinvents validation processes or struggles to access suitable benchmark data. In this use case, the term *Innovation Commons* refers to shared technical resources as well as a cooperative governance space in which contributors pool distributed knowledge to develop shared interpretations that guide validation. The overarching goal is to support responsible AI use and re-use: when one team develops a model or technique, others (including regulators or external innovators) can confidently validate and build on it using common tools in the TRE, without constantly reinventing the wheel or breaching legal boundaries. In essence, UC2 turns the TRE into a collaborative testing ground where best-practice validation protocols, synthetic test datasets, documentation templates, and legal agreements are readily available to all participants.

At its core, the Innovation Commons is about creating an **ecosystem of shared knowledge and tools** inside the TRE. This includes:

- **Reference datasets** (e.g., curated subsets of health data, synthetic data, or anonymised benchmarks) that developers can use to test their models' performance or bias, in addition to their project-specific training data.
- **Reusable software tools** for validation, such as scripts for statistical evaluation, libraries for fairness and robustness checks, and notebooks demonstrating best-practice analytics.
- **Templates and guidelines** covering legal/ethical aspects – for instance, standard operating procedures for privacy assessment, template data sharing agreements, model “factsheets” or documentation templates, and checklists for regulatory compliance.

Concept and motivation

Breaking silos in validation. Today, each research group often must assemble its own validation approach for an AI model – finding appropriate datasets for external testing, devising evaluation criteria, consulting legal teams about IP or patient consent limitations, etc. This duplication is inefficient and leads to *inconsistent standards*. This can be addressed by creating a *common resource pool* within the TRE, curated with contributions from multiple stakeholders (academia, healthcare providers, industry, regulators). The aim is not only to share tools but also to share the reasoning, contextual judgements and annotation practices that underpin them, so that users can understand *why* certain datasets or metrics were selected and how ethical or legal considerations shaped the validation. For example, consider the AI model for diabetic retinopathy detection described earlier. A group validating such a model would not only release their performance results, but also document the decisions that informed their process, such as how specific image categories were handled, what ambiguities they encountered and how they resolved them, and which legal or consent boundaries influenced dataset selection. Publishing this as a “validation bundle” allows others to understand the underlying interpretative work rather than simply the numeric output. A second team working on a related eye-screening system could then build on this foundation, using the bundle to benchmark their model while seeing precisely how earlier choices were made. In this way, standardisation becomes less about duplicating

procedures and more about sharing a common interpretative frame. It enables researchers, clinicians and regulators to read results in comparable terms, identify where divergences arise and ensure *regulatory alignment*. When teams validate against shared benchmarks, regulatory bodies can assess results with greater clarity.

Legal and IP clarity – the Transferable IP (“Transfer IP”) concept. One of the novel ideas within this use case is establishing clear rules for how outputs from the TRE can be shared or commercialised, without compromising the interests of data providers or original creators. Often, when a model is developed on sensitive data, questions arise: Who owns the model? Can it be licenced out, and under what terms? Are there data usage restrictions embedded (e.g., a hospital might only allow use for non-commercial research)? The Innovation Commons would include template **legal agreements** to handle these questions. For instance, the TRE could have a standard joint IP agreement that says any model trained on Hospital X’s data gives Hospital X and the researcher institution joint ownership or a licence. This template could be stored and easily applied as part of the validation/export workflow. Similarly, *licensing models for external use* (open source, proprietary, dual licensing, etc.) can be standardised. The Commons could provide recommended licences (for instance, an AI model open source licence or a specific data-sharing agreement clause) that ensure compliance with GDPR and EHDS. The notion of “Transfer IP” refers to explicitly managing intellectual property when transferring outputs in a way that lowers the entry barrier to the results of research.⁴⁰ By clarifying these upfront (as part of the Commons resources), all parties know the ground rules. The benefit is twofold: researchers know how they can later use their models, and data contributors (e.g. a hospital) know that their data won’t be misused because any derived model will carry forward the agreed restrictions (for example, requiring anonymisation or usage only in certain domains). In this context, Transfer IP operates not merely as a licensing mechanism but as a shared governance tool that helps participants articulate rights, responsibilities and expectations, ensuring that collaboration in the Commons is grounded in clear and equitable relationships.

Reusability and sustained innovation. The vision is for the Innovation Commons to serve as a *knowledge base* and tool repository that grows over time. Each project’s outputs – if deemed generally useful – can be contributed back. Over years, this could mean the TRE holds a library of validated algorithms, reference data, and methodologies. For example, a collection of **synthetic datasets** could be part of the Commons. Synthetic data (i.e., artificially generated patient-like data) is useful for testing algorithms without risking privacy. If such data are available in the TRE Commons, any user can check them for bias and drift and test their AI system on them to get an initial sense of performance or to demonstrate methods without needing real data. Another example is **documentation templates**: the Innovation Commons can include a template for a *Model Factsheet* (also called a “model card”), which captures all important info about an AI model (intended use, training data summary, performance, biases, etc.). This information can be derived from the JUST Data annotation tool, which allows participants to collectively build knowledge about the data and models by testing them in different use case scenarios. In this sense, the

⁴⁰ Magas, M. et al. (2022) *White Paper: IP and Industry Agreements towards Industry Commons*, Zenodo.

Model Factsheet can be dynamically updated, building new knowledge about the data and models used. Google and others have pushed for static model cards that provide a snapshot at a point in time; a TRE Innovation Commons can adapt that idea to a dynamic system where every validated model has an updated factsheet, making it easier to review the suitability of specific models and to share knowledge responsibly.

The *Innovation Commons* thus acts as a **conduit between proprietary efforts and public benefit**. It allows knowledge and tools to flow across projects (and even organisations), while respecting confidentiality and rights. An example of this concept in action can be seen in Sweden’s **VAI-B** initiative, a multi-centre platform for validation of AI algorithms in breast cancer imaging⁴¹. In this project, researchers from several Swedish regions pooled their efforts to create a shared validation framework: AI models from different vendors were all evaluated on the same curated set of mammography images drawn from participating hospitals. The VAI-B platform enabled objective comparison of algorithm performance in a way that no single hospital could have easily done alone. Importantly, the data never left the hospitals; instead, the platform brought the code to each data source, embodying a federated approach to validation. This is analogous to a TRE Commons providing a *distributed* test bed where each site contributes data or resources under a common protocol, and the results (e.g., performance metrics, bias findings) are shared back for collective learning. Such a federated validation approach aligns with Sweden’s decentralised policy ethos and demonstrates how an Innovation Commons can function even when data cannot be centralised. By contrast, in Finland’s more centralised model, the national TRE operated by Findata could host a single repository of validation tools and datasets within one environment, simplifying access at the cost of requiring all participants to use that central platform. Both approaches benefit from the Commons concept – one through networked sharing across nodes, the other through a one-stop hub – underscoring the versatility of this use case.

Alignment with regulatory and ethical standards

The use case approach strongly aligns with emerging policies on secondary data use and trustworthy AI:

- **EHDS and data quality.** EHDS will require that data made available for secondary use is described in data catalogues and conforms to certain quality and interoperability standards. A Commons can host **metadata and schema standards** for datasets, ensuring that when data custodians contribute data to the TRE, they attach the needed info for others to find and use it properly⁴². The Commons could integrate with national data catalogues. For example, if Finland’s Findata TRE publishes a dataset of EHR records, the Commons could list a descriptor so that researchers in Sweden’s TRE know it exists (if cross-border data sharing is allowed). TEHDAS findings showed that less than half of

⁴¹ Cossío, F. et al. (2023) ‘VAI-B: A Multicenter Platform for the External Validation of Artificial Intelligence Algorithms in Breast Imaging’, *Journal of Medical Imaging*, 10(6), 061404; Karolinska Institutet (2025) ‘Now Local Hospitals Can Determine How AI Systems Would Detect Breast Cancer’, available at: <https://news.ki.se/now-local-hospitals-can-determine-how-ai-systems-would-detect-breast-cancer>.

⁴² Kessissoglou, I.A. et al. (2024) ‘Are EU Member States Ready for the European Health Data Space?’

the Member States have established a national dataset catalogue⁴³. The Commons could be a step toward such cataloguing within the TRE context.

- **AI Act compliance.** The EU AI Act will enforce data governance and risk management for AI, including documentation of training data and performance (Article 10 deals with data governance and data management). The validation toolkit can offer *automated compliance checks*, e.g. a tool to evaluate if a training dataset meets representativeness requirements or if bias mitigation steps were logged. By packaging these in the TRE, developers can essentially “pre-certify” their models. Think of it as a *linting* tool for AI compliance: it might output a report indicating, say, *coverage of demographic groups in test data* or *presence of explanation methods*, mirroring what AI Act conformity assessments will look for. This directly helps both developers (to fix issues early) and regulators (who can see these reports).
- **Ethical oversight and Patient and Public Involvement (PPI).** Many grants and ethics bodies now ask for patient and public involvement (PPI) and transparent governance in health data projects. An Innovation Commons fosters this because it inherently encourages sharing and scrutiny. The Commons may include contributions from patient organisations, e.g. a checklist of ethical considerations for algorithm deployment. By standardising validation, it also ensures *ethical consistency*: if every AI system is, say, tested for fairness across key protected attributes, it becomes an ethical norm. Additionally, by having an open repository (within the secure enclave) of what validations have been done, it is easier for ethics committees to audit projects.
- **Pre-certification and sandboxing.** Regulators, including the EMA, are exploring “pre-certification” programs that use known good practices to expedite approvals. A mature TRE Commons could serve as a proving ground that if a model passes all the Commons-provided validation steps, it is more likely to meet regulatory expectations. In other words, the Innovation Commons encapsulates regulatory science know-how (e.g., how to conduct robust clinical validation) and thus can become part of a regulatory sandbox’s toolkit. Notably, one of the benefits listed for this use case is “*long-term regulatory alignment*” and expanding to regionally relevant data. The Innovation Commons should help incorporate data or standards from multiple regions (important if, say, a model trained in Sweden should also be validated on Spanish or Czech data – the Innovation Commons could contain representative datasets or connections to other TREs housing such data, subject to agreements).

Implementing the Innovation Commons

Establishing a functional Innovation Commons in a TRE involves both technical components and governance structures, but its purpose goes beyond resource pooling: it provides a cooperative space in which distributed expertise can be assembled,

⁴³ Kessissoglou, I.A. et al. (2024) ‘Are EU Member States Ready for the European Health Data Space?’

allowing stakeholders to interpret evidence, refine assumptions and build shared understanding before formal structures take shape.

Commons repository and access control. Technically, the Innovation Commons can be a set of storage areas in the TRE (e.g., shared folders or databases) where approved content is stored. All TRE users could have read access, but write/contribution access would be controlled by a curation team. New contributions (a model, dataset, tool) can go through a review or vetting process, overseen by a committee including data providers and legal experts to ensure quality and that sharing it doesn't violate any agreements. Each item in the Innovation Commons should have clear tags (e.g., "Synthetic Data – Cardiology" or "Model Validation Script – Python") and documentation. Modern data platforms use **metadata catalogues** to organise such assets, so integrating one could be useful.

Legal templates and data licences. The Innovation Commons should include a library of legal documents. These would likely be prepared by legal experts in the project. Templates may include a data-sharing agreement that permits an external party to use a model, a model licence for situations in which the model is released as open source, or a contributor agreement confirming that anyone adding to the Innovation Commons has the rights required to share what they contribute. When a researcher exports a model via the TRE, the system could prompt them to pick one of these templates or automatically attach it. In the Innovation Commons, any model or dataset entry should list its *usage rights*: e.g., "*This test dataset can only be used for non-commercial validation*" or "*This model may be reused under CC-BY licence.*" By standardising these, the TRE avoids one-off negotiations each time. A core component of this library should be the Transfer IP framework, which provides a structured way to document the provenance of models, clarify ownership of jointly developed outputs and specify permitted uses when code, models or derived datasets move between partners. Transfer IP ensures that rights and obligations are clear at the point of exchange, reducing uncertainty and enabling contributors to share their work without fear of losing control over future applications. Agreements should also carry contextual notes explaining how rights attach to specific stages of development, which assumptions shaped the licensing choice, and where contributors anticipate uncertainty or future negotiation. This helps create a shared interpretative baseline rather than a purely procedural set of documents.

Reusable test datasets and benchmarks. A particularly valuable part of the Innovation Commons is curated test datasets that do not contain personal data (or are authorised for broad use). Many countries are developing such datasets; for example, the UK has the "National COVID Chest Imaging Database", which, if permitted, could be loaded into a TRE Commons for any COVID-related model validation. Another example is analytics on *cancer data* that leverages the EHDEN/EMIF catalogue of synthetic cohort data. The Commons maintainers should liaise with external data initiatives to bring in useful data (with permission). These datasets can be versioned and come with baseline benchmark results (so a new model's performance can be contextualised against known ones).

Toolkit integration. Tools for bias assessment and model explainability, adversarial robustness toolkits, etc., can be pre-installed in the TRE environment and documented in the Innovation Commons. For instance, the Innovation Commons could contain a

Jupyter notebook template that uses a bias assessment tool to produce a bias report for any given model – this template is a resource anyone can copy. Additionally, standardised “computational workflows” (using WfMS or pipelines) could be set up, e.g. a pipeline for k-fold cross-validation or for generating differential privacy metrics. The user could then run these with minimal tweaking on their own data.

Governance and curation. An Innovation Commons likely needs a governing body to decide what enters and how it is managed. A committee from the TRE project with members from different sectors could be formed to cover technical, legal, and ethical angles. They would periodically update tools (e.g., add a new version of a library or new templates if regulations change) and review contributions. For durability, one could formalise this structure, possibly as part of the national data governance mechanism under EHDS.

Example scenario. A regional health authority and a university collaborate on an AI for early cancer detection. They use the TRE Commons at the project’s start to select appropriate Transfer IP licensing terms (e.g., double use for public and commercial deployment), that clarify ownership of any jointly developed model, specify how the model may be reused or licenced and document the rights retained by each contributor. This early choice helps reduce uncertainty by giving all parties a shared understanding of what future deployment or commercialisation may involve. They then confirm that the data inputs meet JUST principles (Judicious, Unbiased, Safe, Transparent)⁴⁴ and apply a standardised validation protocol. All these steps are facilitated by Innovation Commons resources: a licensing template is chosen, a dataset checklist (JUST) is run, and a validation pipeline is executed. Alongside these procedural tools, the team adds interpretative notes explaining how clinical experts judged data quality, how annotators resolved ambiguous cases and how developers selected evaluation metrics. Once the model is validated, they contribute their protocol and findings back to the Commons, allowing others to learn from it. Meanwhile, because they used standard agreements, they can publish the model or share it with another hospital under the predetermined conditions.

This use case seeks to enable a virtuous circle that **lowers entry barriers**, **standardises processes**, and **enables trust and reuse** so that, over time, the cost (in time and risk) of validating an AI model in health drops significantly. That encourages more institutions (including smaller ones or startups) to innovate using health data, because the TRE Innovation Commons provides a supportive ecosystem.

Challenges and mitigations

Implementing this use case is not without its challenges.

Incentivising contribution. The Innovation Commons relies on people sharing their tools and data. Some may be hesitant (e.g., a company might feel its validation pipeline is proprietary). To address this, policies can mandate that if one uses the TRE funded by public money, one should contribute back non-sensitive outputs (similar to open science mandates). Also, demonstrating the *mutual benefit* is key: when everyone shares, everyone saves effort. There could be recognition or citation mechanisms – e.g.,

⁴⁴ Magas, M. and Dubber, A. (2020) Expanding EOSC; Emanuilov, I. and Magas, M. (2024) White Paper on Data Documentation for JUST Data Practices.

Commons contributions are citable in the same manner as publications, giving academic credit.

Quality control. Poor quality or outdated resources in the Innovation Commons could mislead others, and continuous JUST Data annotation is needed. This could include archiving older versions and clearly marking them, using user feedback to allow researchers to rate or comment on Commons tools (“this dataset has these quirks...” etc.). Over time, a robust, community-driven knowledge building can emerge, analogous to how open-source software is reused on platforms such as GitHub through community issues and pull requests (though within the confines of the TRE, this would be internal).

Interoperability across TREs. Ideally, Innovation Commons resources could be shared across multiple TRE instances (with agreements). If each country or region has its TRE, how can duplication of the Commons in each be avoided? This could be solved via the future HealthData@EU network – a federation where data and perhaps resources can be linked. The Commons concept can potentially scale up to a pan-European “AI validation toolkit” shared by all Health Data Access Bodies. Early cooperation via projects like TEHDAS or Joint Actions could pave the way. The Gaeta et al. (2025) publication on the GATEKEEPER platform for EHDS suggests building secure processing environments with *common services* for secondary use, which likely aligns with having shared validation modules.⁴⁵

Maintaining legal relevance. Law evolves, e.g., new IP cases or new guidelines for AI arise. The Innovation Commons’ Transfer IP legal templates need updating to remain enforceable and effective. It will be important to involve legal counsel on an ongoing basis.

By systematically addressing these, the Innovation Commons can become a cornerstone of a sustainable TRE ecosystem. It operationalises the often-mentioned but seldom implemented idea of *learning health systems* in AI – where each experiment informs the next, and knowledge accumulates in a repository for the community (within the secure boundaries needed for privacy).

Use case 3

UC3 builds on the outputs of UC2. Once a model has been validated, documented, and legally framed through the Commons, the next challenge is to move it safely and accountably into real-world use. UC3 therefore takes the results of UC2 – the model card, the JUST-aligned annotation record, and the Transfer IP terms – and turns them into a deployment-ready package with mechanisms for post-deployment learning. UC3 is about creating structured pathways for the **secure and compliant transition of AI models from the research TRE to production settings**, whether in a hospital clinical workflow or an external operational setting. This includes the technical and procedural means to export the model with all necessary documentation (a “Pre-Flight validation airlock”), handling of intellectual property and licensing for deployment, and establishing ongoing channels for monitoring model performance post-deployment. UC3 essentially extends the TRE’s influence into the deployment phase, ensuring that

⁴⁵ Gaeta, E. et al. (2025) ‘GATEKEEPER Platform: Secure Processing Environment for European Health Data Space’, IEEE Access, 13, pp. 34627–34638.

the model remains under appropriate oversight so that any emerging issues can be traced and addressed. The central question is how to move from a model that exists as a technical artefact inside a TRE to a model that is understood, trusted and governed by the people who will use, maintain and oversee it in clinical and organisational settings.

Goals and key elements

Seamless transfer with regulatory readiness. When researchers have finished building an AI model in a TRE, that model often needs to be moved to a different environment – for example, to integrate into a hospital’s IT system or to be packaged as a product by a company. UC3 aims to make that transfer as **seamless, secure, and well-documented** as possible. The TRE would provide an “*Export Model*” function that is much more than just downloading a file. The export process should include technical, legal, and IP packaging, a Pre-Flight validation airlock to assess model safety and explainability, predefined export formats (e.g. ONNX, Docker container format) with standardised documentation, legal/IP templates for deployment agreements, data protection assessment, and logging of provenance.

In practice, this means the TRE will output a bundle containing:

- The model artefacts themselves (e.g., weights file, code or container image).
- Accompanying updated documentation (model factsheet or “qualification dossier” detailing how it was developed and validated – likely assembled from UC1/UC2 logs).
- Confirmation that the model has been checked for personal data (for instance, scanned to ensure it does not include any sensitive remnants of the training data or memorised patterns that may reveal real patient data points).
- Legal clearance forms (for instance, an export certificate summarising IP ownership and licence).
- An audit trail or cryptographic signature to prove provenance (so that if any question arises, one can verify this model indeed came from the TRE unaltered).

This comprehensive packaging is akin to preparing a medical product for market approval. It addresses regulators’ needs (documentation and assurance of safety measures) and developers’ needs (clear licensing to use the model).

Clarity on IP and licensing. In many cases, multiple parties have rights to an AI model trained on sensitive data. For example, if a model was trained on hospital patient data by a university team with some algorithm from a company, then the hospital, university, and company might all have stakes. UC3 formalises how those rights are sorted *before* export. In the TRE, using templates from UC2, an **IP assessment** is done: key questions such as ‘*who owns the model weights?*’, ‘*under what conditions can it be deployed?*’, ‘*does it count as a medical device and need CE marking?*’ are addressed. A likely outcome is a *joint ownership or licensing agreement* that travels with the model. In the diabetic retinopathy example given, the IP assessment assigns joint ownership to the

research institution and the healthcare provider, and a Transfer IP licensing agreement allows commercial deployment under set conditions by clarifying ownership, permitted uses and shared responsibilities at the moment a model leaves the TRE, reinforcing that deployment is not only a technical transition but also a negotiated handover among the people who developed, validated and will ultimately use the system. Having this spelt out prevents disputes later and ensures that, for example, the hospital can share in benefits or that the model can legally be integrated into a vendor's software.

Technical portability. Another crucial aspect is making sure the model can actually run outside the TRE environment. TREs might use specific hardware or configurations; the exported model should be in a standard format that any common platform can run (ONNX is an open neural network format, Docker containerises everything needed). The TRE should automate the conversion of the model to such formats if needed. For instance, if a model is in a Python pickle (an inadvisable format for deployment), the TRE could export it as an ONNX file, which is more interoperable. Providing a Docker means the entire inference environment (with libraries, versions, etc.) can be preserved, reducing "it works on my machine" issues. This is a best practice in machine learning ops being directly adopted as a TRE feature. Essentially, the TRE acts as a DevOps pipeline that, at the end of research, produces a production-ready artefact.

Pre-Flight validation airlock checks. Before greenlighting the export, the researcher, clinician, developer or team working within a TRE should perform one last evaluation round – akin to a pre-flight checklist. A Pre-Flight validation airlock assesses model safety, explainability, and regulatory readiness. In practice, this means running a final set of automation-supported checks, such as running the model on a hold-out test set one more time to ensure it meets a certain accuracy threshold or does not exhibit drift since training, or an explainability check to confirm it's not using inappropriate features. Possibly even a "red teaming" security test (some AI can embed data or be tricked). These could be integrated from the UC2 toolkit. Only if users verify that the model passes these checks (or waivers are signed) does the TRE package it for export. This ensures that what leaves the TRE is genuinely believed to be safe and effective.

Post-deployment monitoring integration. A standout feature of UC3 is extending monitoring *beyond* the TRE. Once the model is deployed in, say, a hospital's system, the TRE project doesn't wash its hands of it – instead, it sets up mechanisms to continue receiving information. This could be achieved by deploying a *monitoring agent* along with the model that collects performance statistics without any personal data, ideally limited to aggregate outcomes or drift metrics, and periodically returns these in aggregated form to the TRE JUST Data annotation system or a regulator's dashboard. By doing so, if the model's performance in the real world starts to degrade or certain rare errors pop up, those can be seen by the original developers or oversight entities. The benefit is twofold: model developers can update or retrain models proactively (perhaps pushing an update through the TRE export process again), and regulators have ongoing evidence rather than a one-time check. This aligns with how medical device regulation is moving – requiring *post-market surveillance* of AI as part of compliance (the EU MDR requires manufacturers to continuously collect and analyse performance data of deployed devices). The TRE could provide a structured way to do this by serving as the hub where new data from deployment flows in under controlled conditions for analysis.

Building confidence for end-users. UC3 *builds* confidence for regulators, developers, and end users by providing structured pathways from experimental to operational use. Healthcare providers (end users) will trust an AI more if they know it came through a rigorous pipeline – especially if they see that it continues to be monitored and supported. For example, if a hospital deploys an AI via this TRE process, they could receive periodic “health reports” on the model from the TRE team. And if a regulator approved it, the hospital knows it’s not just an unregulated novelty.

Case example and workflow

To illustrate, consider the *Research-to-Deployment* journey of the diabetic retinopathy model described earlier:

1. **Development in TRE.** A deep learning model is trained on de-identified eye images in the TRE, achieving good accuracy. Throughout training, researchers and clinicians annotate contextual information using JUST Data practices, documenting interpretative choices and data nuances that may influence the model’s behaviour. This shared record forms part of the model’s provenance trail.
2. **Final validation & export prep.** Researchers trigger the export process. The TRE supports them by assembling the necessary evidence for review. It confirms the model file contains no patient IDs (perhaps comparing some of its internal values to ensure none match individual data points), flagging points that may require human attention, such as unusual parameter patterns or inconsistencies with previous validation runs, and generates a *model card* document. The legal review inside TRE performs JUST Data checks and confirms that under the data permit, this model can be exported (because it’s sufficiently transformative, containing no personal data – as required by GDPR’s test for anonymous output).
3. **IP & licence setup.** The system knows the data came from Hospital X (with an agreement that the hospital expects co-ownership) and that the algorithm originated from University Y. It uses the pre-agreed Transfer IP template to state: *“Hospital X and University Y are co-owners of the model. The model can be used for patient care within Hospital X freely. Commercial use outside requires a licence.”* A Transfer IP licensing agreement is automatically included that, say, grants a certain company the right to deploy the model in its diagnostic software for a fee or for free under specified conditions.
4. **Packaging.** The TRE exports the model as an ONNX file and also provides a Docker container with a simple API to run the model on new images, plus all needed libraries (say PyTorch, etc.). It also exports log files and the documentation PDF.
5. **Approval & deployment.** The hospital’s regulatory team and perhaps the national regulator review the export bundle. Seeing that all documentation is in order, they approve using it on patients. The model is then loaded into the

hospital's system from the Docker.

6. **Post-deployment link.** The hospital system has a monitoring component. Each time the model processes images, it logs whether its prediction was likely correct (later confirmed by clinical review) and some statistics such as image quality. These logs, aggregated weekly, are sent back to a secure endpoint at the TRE or oversight body in accordance with data protection compliance. Only summary information is transmitted, or the transfer occurs through a secure channel, and the material is integrated into continuous monitoring and annotation.
7. **Ongoing analysis.** Back at the TRE, those incoming logs are ingested. The original researchers establish an ongoing project to analyse performance drift and record it in the JUST Data annotation system. Over six months, they observe that the model begins to miss certain lesion types or shifts in imaging characteristics it previously identified. This could stem from new imaging devices producing slightly different data. They prepare an improved model version addressing that drift.
8. **Re-iteration.** The improved model is retrained in the TRE (perhaps incorporating some new data if allowed), goes through the Pre-Flight export airlock again, and version 2.0 is deployed. All the while, regulators are in the loop and can acquire new knowledge from this continuous improvement loop functioning under the umbrella of the TRE sandbox rather than missing out on information that is acquired from testing ad hoc in the wild. The emphasis is on a collaborative decision-making process supported by technical tools, rather than an automated pass/fail mechanism.

This scenario shows how the three use cases actually tie together: UC1 ensured drift was detectable; UC2 ensured documentation and legal frames; UC3 executes the actual transition and monitoring. Taken together, these stages form a socio-technical process in which human judgement, shared interpretation and negotiated responsibility remain central, with the TRE providing the secure infrastructure that enables these forms of collaboration to take place.

Addressing challenges

Several challenges arise in UC3.

Ensuring no personal data escapes. There is a theoretical risk that a model, especially a complex one, could encode aspects of the training data. For instance, if a nearest-neighbour style model memorised one patient's data, or a generative model could recreate images. To mitigate this, the export check should include techniques such as *membership inference tests* (to see if the model can distinguish its training data from other data)⁴⁶. Research in machine learning security is ongoing to create automated tools for this. In the meantime, policy might demand that a human review the model's

⁴⁶ Jefferson, E. et al. (2022) GRAIMATTER Green Paper, Zenodo.

outputs on some test inputs to ensure, for example, detection of any residual identifying signal embedded within the model parameters or introduction of harmful bias into the system.

Data protection compliance. Under GDPR, once a model is fully trained and provably does not contain personal data,⁴⁷ it can be considered anonymous and move freely. The TRE data protection assessment (likely a part of the export procedure) will document the rationale that the model is anonymised (or pseudonymised, etc.). Recital 15 DGA suggests data should only be transmitted out of an SPE if a legal basis allows or the data is non-personal. So the TRE as an access body would use that logic: if the model was trained on personal data, they ensure it is sufficiently aggregated and transformed to no longer be personal. This can involve a DPIA (Data Protection Impact Assessment) on the model export. UC3's structured approach ensures this is not an afterthought but an integral step.

Integration with external systems. Once the model is out, how can it be monitored? This requires integration agreements. The hospital has to agree to send data back (which can itself be considered secondary use). This may be managed by treating those performance logs as a new data contribution to the TRE under EHDS (the hospital is a data holder providing data about model performance). Because EHDS explicitly encourages feedback loops for improving algorithms, this can most likely be arranged legally. Technically, a simple approach is an API where the deployed model container sends statistics to the TRE's secure API endpoint. The data sent should be aggregated or at least pseudonymised sufficiently not to re-identify individuals. For example, "out of 100 cases this week, the model flagged 5, of which 4 were confirmed true positives, 1 false positive". These are statistics with no personal details included.

Operational responsibility. Who monitors the monitors? In practice, the original researchers or the party that took the model to deployment, such as a company, will be responsible for reviewing the incoming data on a regular basis. The TRE could support this by providing suitable infrastructure, including the option of an automated alert if a significant change occurs. A question remains about the duration of the TRE's involvement, which may be limited to the defined period of a regulatory sandbox or pilot. After that period, the responsibility could shift entirely to the deployer. If the TRE forms part of a broader continuous improvement ecosystem, models could continue to pass through it for updates.

Scaling to multiple deployments. If a model is deployed in dozens of hospitals, the monitoring data could become a large-scale stream. TREs would need to accommodate this increase in scale and the accompanying complexity, which starts to look like a federated network of models "calling home". One approach to this issue is to condense the data at each site and transmit only the summaries. Another is to use a form of federated evaluation in which the TRE periodically issues a test query to each site's model to assess its current behaviour.

⁴⁷ European Data Protection Board (2024) Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models, available at: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

Best practices and future opportunities

UC3 essentially sets up a **closed-loop system** for AI lifecycle management. Some best practices that emerge include:

Version control and provenance. Each exported model should be versioned, and the TRE should keep a hash of it. If that model is later involved in any incident, one can verify if it was altered. The Commons could also keep an archived copy if allowed, or at least its evaluation results.

User training and documentation. The end-users (doctors, etc.) who deploy the model should receive documentation (which the TRE export provides) on the model's intended use and limitations from the JUST Data system. That reduces misuse. It can be a requirement that the deploying organisation signs off that they understand and monitor updates of the model card.

Feedback into research. Data coming back could potentially be used to further improve models within the TRE (as new training data or at least test data). This needs careful consent and legal basis (using real patient outcomes as new training data may require fresh permissions unless covered under research/policy mandate). But EHDS aims to make such cross-uses easier via data permit systems. So, if the hospital agrees, those performance data will become new entries in the TRE's dataset catalogue for future research to analyse algorithm effectiveness.

Cross-border deployment. If a model from a Swedish TRE gets deployed in, say, Finland, how do the Finnish data feed back? This could route through the HealthData@EU. Possibly, the model developer would apply for a data permit to get performance data from Finland's health data via its access body, meaning the oversight goes through official channels. In a (cross-border) regulatory sandbox context, special arrangements could simplify that.

UC3 completes the picture by ensuring that the TRE approach does not stop when the research paper is written, but carries through to real patient impact, creating a **learning loop**. By doing so, it effectively imports the concept of *pharmacovigilance* (continuous drug safety monitoring) into the AI realm – we might call it “*model-vigilance*.” In sum, the TRE is not just a one-way street where data goes in and models come out; it becomes a two-way street where models go out with careful shepherding and feedback comes in to inform safe use and next iterations. This makes TREs invaluable to both innovation and regulation: they serve as the place where cutting-edge AI meets the harsh realities of clinical practice, under watchful yet supportive eyes.

Viewed together, the three use cases form a complete progression: UC1 focuses on how models are built and understood, UC2 shows how that work is tested and documented so others can reuse them, and UC3 carries the model into real-world use while ensuring continued learning and oversight.

Conclusion and recommendations

TREs are poised to become the backbone of safe and legal health data innovation in Europe, especially as the European Health Data Space comes into effect. The three next-generation capability areas explored – real-time AI oversight (UC1), shared validation commons (UC2), and seamless model deployment with ongoing monitoring (UC3) – together chart a path toward TREs that are not only secure by design, but also *innovation-friendly and compliance-facilitating*. These capabilities directly address the gaps identified in current TRE operations: handling complex AI workflows, ensuring consistent validation practices, and managing the full lifecycle of data-driven solutions from research to real-world impact. Their value ultimately lies in the way they allow diverse participants to see the same evidence, record their reasoning and share responsibility for how health data and AI systems are developed, assessed and used.

By implementing UC1, TREs actively mitigate risks during AI development rather than catching problems late. This meets emerging regulatory expectations for continuous risk management and creates a collaborative space for regulators and researchers to jointly navigate AI development. It demonstrates that oversight is most effective when it is embedded within the development process and when regulators, clinicians, data providers and developers share a common view of how a model evolves in real time. We recommend that TRE operators invest in developing integrated monitoring dashboards and analytic logging tailored to machine learning workflows. Pilot projects should be undertaken with regulatory bodies (for example, involving medical device regulators in sandbox trials) to refine what metrics and interventions are most useful. Clear, co-designed protocols should also establish the respective roles and limits of participants within the TRE, ensuring that supervision enhances learning rather than introducing a hierarchical gatekeeping dynamic. In short, build oversight into the process, not just as an afterthought.

A central component of this governance is the use of Transfer IP, which clarifies rights, responsibilities and expectations as models move between contributors, ensuring that collaborative development is matched by equally clear and traceable accountability. Through UC2, TREs can dramatically improve efficiency and trust by sharing validation resources. We recommend establishing a governance board for the *Innovation Commons* with representatives from data providers, AI developers, and legal experts to curate high-quality content. Priority should be given to populating the Commons with: a) reference datasets (especially synthetic or openly sharable ones) for popular validation tasks; b) standard evaluation pipelines such as JUST Data annotation (for bias, robustness, accuracy) that align with guidelines such as FUTURE-AI and upcoming AI Act requirements; c) a legal and Transfer IP template library covering common secondary use scenarios. Crucially, the Commons should not be treated as a passive repository of automated tools, but as a shared interpretative space in which contributors articulate uncertainties, add contextual knowledge and document how decisions are made. To operationalise this, funding bodies and regulators might require that projects using a TRE contribute certain outputs (e.g. a validated model or protocol) to the Commons, fostering a culture of open innovation. In essence, the **TRE becomes an “open lab”** where validated building blocks for AI in health accumulate, and where collaborative knowledge replaces isolated technical efforts, lowering barriers for

newcomers and increasing overall quality and reproducibility of research. We also suggest exploring cross-TRE federations of Commons – perhaps via the HealthData@EU infrastructure – so that resources can be shared internationally, promoting harmonisation of standards across Member States.

Implementing UC3 will ensure that TREs fulfil the EHDS vision of a *complete pipeline from data to societal benefit*. We urge TRE initiatives to design their architecture with **deployment in mind from the start**. This means defining an “export contract” for models: what formats, documents, and clearances are required. An important recommendation is to **formalise the Pre-Flight validation airlock** as a mandatory step for releasing any results beyond simple statistical tables, extending the existing output-checking process to cover complex outputs such as ML models. TRE operators should collaborate with standards bodies to define how to certify that an exported model is free of personal data (e.g. developing audit scripts or guidelines, potentially contributing to ISO or CEN standards on anonymous outputs). Equally, exported models must carry a clear and shared narrative of their provenance, captured through JUST Data annotation, UC1 monitoring and UC2 validation, so that deployment settings inherit not only the artefact but the knowledge that produced it. We also recommend that health institutions and companies planning to consume models from TREs adapt their systems to accept the packaged models and provide feedback. One possible approach is establishing **data contracts for model feedback**, such that when a hospital obtains a model via the TRE, it agrees to supply back de-identified performance data for a certain period. This will operationalise the learning loop in practice.

From a policy perspective, **ensuring legal compatibility** is vital. The EHDS Regulation should be interpreted flexibly to allow the kind of iterative data flows described. For example, data returned for model monitoring may need a fresh permit – data authorities should streamline this and, where appropriate, integrate it into the original project’s permit when planned in advance. Additionally, guidelines under EHDS could explicitly endorse these TRE capabilities: e.g., recommending that Health Data Access Bodies (HDABs) provide model export and monitoring services. The **Data Governance Act’s secure processing framework is broad enough** to encompass these features, but clear communication is needed so that stakeholders realise TREs can do more than static analysis; they can be incubation and testing environments under regulatory oversight.

In summary, aligning TREs with the EHDS requirements and maximising their value requires a shift from viewing TREs as purely technical infrastructures to understanding them as socio-technical environments that depend on human judgement, shared interpretation and cooperative governance. It is necessary to:

- **Augment TRE platforms** with AI-specific tools (monitoring interfaces, validation pipelines, export packaging) to support advanced use cases that traditional safe havens could not handle. This requires technical development and close stakeholder collaboration, but examples in this report show it is feasible.
- **Embed multi-stakeholder governance** where regulators, developers, data subjects (patients) and data providers should all have roles in the TRE process –

not only to supervise, but to learn from one another and create shared standards of practice, from co-monitoring development to contributing to the commons to overseeing deployment. This multi-stakeholder approach transforms TREs into true *Trusted Research and Regulatory* Environments supporting a future of evidence-based regulation (regulators base decisions on rich evidence generated in TRE sandboxes rather than abstract submissions).

- **Promote interoperability and standards.** To avoid each TRE becoming an island, work on common standards for model documentation, audit logs, JUST Data annotation, Transfer IP licensing frameworks, Pre-Flight airlocks and even APIs for cross-TRE model sharing. Align these with international efforts (for instance, ISO/AAMI work on AI in healthcare). An EU-wide initiative under the EHDS could facilitate a federated network of TREs where resources and computational tasks can be shared securely across borders (building on projects such as TEHDAS outcomes, which highlighted the heterogeneity but willingness to align).

By following these recommendations, Europe can ensure that its move to secure data environments does not stifle innovation, but, on the contrary, that it **actively enables more ambitious, accountable and trustworthy research**. An AI developed under this TRE paradigm comes with built-in proof of safety, fairness, and effectiveness – thus enabling more timely and well-supported routes to clinical use. The TRE becomes a place where “*Better, Broader, Safer*” isn’t just an aspirational slogan but a daily practice. In doing so, it supports not only compliance with EHDS and GDPR, but the fundamental outcomes of new knowledge, improved healthcare solutions, and protected patient rights.

A next-generation TRE is therefore more than a secure computational space. It is an environment in which distributed human reasoning can take place, where clinicians, researchers, patients, regulators and technical developers share access to the same evidence and contribute to the interpretative and evaluative work that shapes an AI system. This collaborative arrangement strengthens collective accountability because decisions about model behaviour, model limits and model deployment are made in view of others who understand their implications. It also supports participatory governance, enabling those affected by AI systems – including data providers and patient representatives – to contribute substantively to their evaluation and oversight. These are precisely the types of human-centred practices envisaged by the EHDS, which positions secure processing environments not only as compliance mechanisms but as spaces in which trustworthy, evidence-based use of health data can be achieved.

The TRE is not solely an instrument of compliance, but also an institutional setting for long-term responsibility. The tools described across UC1, UC2 and UC3 allow the entire lifecycle of an AI model – from annotation and training through validation, deployment and post-deployment monitoring – to be documented, revisited, discussed and scrutinised by the people who rely on it. This strengthening of socio-technical accountability helps ensure that healthcare AI does not proceed as a sequence of

isolated technical steps, but as a sustained process of shared judgement, evidence-building and stewardship. In doing so, next-generation TREs directly support the EHDS ambition to create an ecosystem in which innovation and fundamental rights reinforce one another, enabling data-driven advances in healthcare without compromising patient autonomy or public trust. The future of health data science will be *secure by default and collaborative by design*, and next-generation TRE capabilities are the vehicle to get there.

Acknowledgements

This work is funded as part of the Advanced Digitalisation programme by Vinnova, the Swedish Innovation Agency, project reference 2024-01412.

The authors wish to thank Daniel Lundqvist, Andreas Hager, Ulf Petrusson and Christoffer Hermansson for their valuable insights and contributions.

Bibliography

- Cossío, F., Schurz, H., Engström, M., et al. (2023) 'VAI-B: A Multicenter Platform for the External Validation of Artificial Intelligence Algorithms in Breast Imaging'. *Journal of Medical Imaging*, 10(6), p. 061404. Available at: <https://doi.org/10.1117/1.JMI.10.6.061404>
- Danish Contractor, McDuff, D., Haines, J., et al. (2022) 'Behavioral Use Licensing for Responsible AI'. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 21 June, pp. 778–788. Available at: <https://doi.org/10.1145/3531146.3533143>
- DRAGoN, University of the West of England (n.d.) *The Five Safes*. Accessed 7 November 2025. Available at: <https://fivesafes.org/>
- E-hälsomyndigheten (2022) Förstudie om ett statligt, nationellt datautrymme för bilddiagnostik (S2021/05259 delvis). E-hälsomyndigheten.
- Emanuilov, I. and Magas, M. (2024) *White Paper on Data Documentation for JUST Data Practices*. Industry Commons Foundation. Available at: <https://doi.org/10.5281/zenodo.14228289>
- Emanuilov, I., Larsson, B., Magas, M. and Dubber, A. (2025) 'Trusted Research Environments for Healthcare AI: State of the Art Global Landscape Report'. Industry Commons Foundation. Available at: <https://doi.org/10.5281/zenodo.17670445>
- Emanuilov, I., Larsson, B., Dubber, A. and Magas, M. (2025) 'Requirements Specification for a Swedish Trusted Research Environment for Healthcare AI'. Industry Commons Foundation. Submitted for review.
- European Data Protection Board (2024) Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models. European Data Protection Board. Available at: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf
- 'Färdplan för nationell digital infrastruktur (Interpellation 2024/25:742 av Anna Vikström (S))' (2025) Swedish Parliament, 12 September. Available at: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/interpellation/fardplan-for-nationell-digital-infrastruktur_hc10742/
- Gaeta, E., Haleem, M.S., Lopez-Perez, L., et al. (2025) 'GATEKEEPER Platform: Secure Processing Environment for European Health Data Space'. *IEEE Access*, 13, pp. 34627–34638. Available at: <https://doi.org/10.1109/access.2025.3539559>
- Goldacre, B. and Morley, J. (2022) *Better, Broader, Safer: Using Health Data for Research and Analysis. A Review Commissioned by the Secretary of State for Health and Social Care*. Department of Health and Social Care.
- Hager, A., Emanuilov, I. (2025) 'Trusted Research Environments for Healthcare AI: Sweden - Finland Comparison', Industry Commons Foundation. Submitted for Review.

- Jefferson, E., Liley, J., Malone, M., et al. (2022) GRAIMATTER Green Paper: Recommendations for Disclosure Control of Trained Machine Learning (ML) Models from Trusted Research Environments (TREs). Zenodo. Available at: <https://doi.org/10.5281/zenodo.7089491>
- Karolinska Institutet (2025) 'Now Local Hospitals Can Determine How AI Systems Would Detect Breast Cancer'. 3 November. Available at: <https://news.ki.se/now-local-hospitals-can-determine-how-ai-systems-would-detect-breast-cancer>
- Kavianpour, S., Sutherland, J., Mansouri-Benssassi, E., Coull, N. and Jefferson, E. (2022) 'Next-Generation Capabilities in Trusted Research Environments: Interview Study'. *Journal of Medical Internet Research*, 24(9), e33720. Available at: <https://doi.org/10.2196/33720>
- Kessissoglou, I.A., Cosgrove, S.M., Abboud, L.A., et al. (2024) 'Are EU Member States Ready for the European Health Data Space? Lessons Learnt on the Secondary Use of Health Data from the TEHDAS Joint Action'. *European Journal of Public Health*, 34(6), pp. 1102–1108. Available at: <https://doi.org/10.1093/eurpub/ckae160>
- Lekadir, K., Frangi, A.F., Porras, A.R., et al. (2025) 'FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare'. *BMJ*, e081554. Available at: <https://doi.org/10.1136/bmj-2024-081554>
- Magas, M. and Dubber, A. (2020) 'Expanding EOSC: Engagement of the Wider Public Sector and Private Sectors in EOSC'. Zenodo. Available at: <https://doi.org/10.5281/ZENODO.4463437>
- Magas, M., Radziwon, A., Altosaar, A., Wretblad, L., Emanuilov, I. and Bertels, N. (2022) *White Paper: IP and Industry Agreements towards Industry Commons*. Industry Commons Foundation. Available at: <https://zenodo.org/records/14566936>
- Ministry of Social Affairs and Health (n.d.) *Act on the Secondary Use of Health and Social Data*. Available at: <https://stm.fi/documents/1271139/1365571/The+Act+on+the+Secondary+Use+of+Health+and+Social+Data/a2bca08c-d067-3e54-45d1-18096de0ed76/The+Act+on+the+Secondary+Use+of+Health+and+Social+Data.pdf?t=1559641328000>
- Potts, J., (2019) 'Innovation Commons: The Origin of Economic Growth' (New York, 2019; online edn, Oxford Academic, 22 Aug 2019), available at <https://doi.org/10.1093/oso/9780190937492.001.0001>
- Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act) (2022) *Official Journal of the European Union*. Available at: <http://data.europa.eu/eli/reg/2022/868/oj>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence... (Artificial Intelligence Act) (2024). Available at: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space... (2025). Available at:
<http://data.europa.eu/eli/reg/2025/327/oj/eng>

Svensberg, L., Lindvall, J., Danielsson, P., Nilsson, A.G., et al, Vinnova *Policy Paper for Commons for Development of More Efficient Innovation Ecosystems*.
https://digitalwellarena.se/wp-content/uploads/2024/05/policypaper_commons_eng.pdf

Sverige (2024) Delad hälsodata – dubbel nytta: regler för ökad interoperabilitet i hälso- och sjukvården. Regeringskansliet.