International Journal of Population Data Science

Journal Website: www.ijpds.org

How to safely RELEASE-AI models: a lifecycle framework for Trusted Research Environments

Alba Crespi-Boixader[1], Simon Li[1], James Liley[2], Laura Ward[1], Christian Cole[1], and Jim Smith[3]

[1]University of Dundee, Dundee, United Kingdom
[2]University of Durham, Durham, United Kingdom
[3]University of the West of England, Bristol, United Kingdom

Trusted Research Environments (TREs) enable approved researchers to securely analyse personal data, including Electronic Health Records (EHRs), under strict governance. The accelerating use of Artificial Intelligence and Machine Learning in research with sensitive data, however, introduces novel privacy and disclosure risks that traditional statistical disclosure control methods cannot adequately address. This work presents a comprehensive framework for managing these risks throughout the full AI/ML project lifecycle within TREs. Organised across six phases—design, governance, development, evaluation, disclosure control, and release—the framework provides clear, phase-specific guidance and assigns explicit responsibilities to all involved parties: researchers, project teams, output checkers, data controllers, and TRE staff. The RELEASE-AI contains 28 practical statements, each allocated to a responsible role (e.g., researcher, TRE operator) and prioritised using a modified MoSCoW model (Must, Should, Could). This prioritisation enables proportionate implementation according to risk and resource constraints. The framework promotes early risk identification, proportionate mitigations, and good AI/ML practices, including robust code and documentation standards, privacy-enhancing techniques (e.g., differential privacy), restricted model access via secure query systems, licensing agreements, and pre-release adversarial testing. It emphasises role-specific training to ensure effective implementation. A novel tiering system for disclosure control is proposed, categorising AI projects based on the likelihood of attack and the severity of potential sensitive data leakage. By integrating a lifecycle-focused risk management process with this scalable disclosure control tiering system, the framework enables innovative AI research while maintaining rigorous data protection standards and sustaining public trust in the use of sensitive personal information.

https://doi.org/10.23889/ijpds.v11i5.3646