

International Journal of Population Data Science

Journal Website: www.ijpds.org



Enhancing Safety in Distributed Analytics: A Framework for Assessing and Reducing Privacy Risk

Marie-Pier Domingue¹, Jean-Philippe Morissette², Simon Lévesque¹, Anita Burgun³, Félix Camirand Lemyre², Carrie-Anne Whyte⁴, Kim McGrail⁵, and Jean-François Ethier²

¹Groupe de recherche interdisciplinaire en informatique de la santé (GRIIS), Sherbrooke, Canada; Institut Imagine, Université Paris Cité, Paris, France

²Groupe de recherche interdisciplinaire en informatique de la santé (GRIIS), Sherbrooke, Canada

³Institut Imagine, Université Paris Cité, Paris, France; Necker Hospital, AP-HP, Paris, France

⁴Canadian Institute for Health Information, Ottawa, Canada

⁵HDRN Canada, Vancouver, Canada

Distributed analytics enables analyses to be conducted when data is partitioned across multiple sites and cannot be pooled in one location. The primary motivation for using these methods is to protect confidentiality of line-level data. Distributed analytics methods rely on exchanging intermediate numerical outputs generated locally at each participating site. However, the absence of line-level data exchange should not be mistaken for privacy protection. To address this, the GRIIS developed a framework for evaluating privacy, helping users distinguish different levels of privacy preservation when applying distributed methods. This work was especially important, as earlier research demonstrated that an existing distributed method could be reverse-engineered to recover line-level data, which underscores that the distributed nature of an approach does not guarantee data privacy. In the framework, we define

data privacy as guaranteed if line-level data cannot be uniquely recovered from the numerical outputs exchanged or disclosed during the execution of the approach. By eliciting line-level data as a set of unknown variables subject to constraints derived from the procedure, we can, in some settings, mathematically prove that privacy is protected under this definition. Achieving this depends on the nature of the variables included (binary, continuous), the results shared at the end of the procedure and the availability of external information. Ultimately, this work plays a critical role in supporting the responsible use of distributed analytics. By establishing governance mechanisms to evaluate and mitigate privacy risks, it helps ensure that these methods can be adopted appropriately and at scale across the country.

