MD-Paedigree and CARDIOPROOF General Policy on Privacy and Regulatory Issues:

Properly addressing privacy and regulatory guidelines is a fundamental issue, rightly emphasized by the EU as a necessary goal for "strengthening individuals' trust and confidence in the digital environment and enhancing legal certainty".

The legal framework of data protection in the European Union is built on several general principles. These principles are, at least at a general level, also recognized in other international sources, such as the Council of Europe Convention for the Protection of Individuals with regards to Automatic Processing of Personal Data, from 1980, on which the EU Data Protection Directive is partially built, and the OECD Guidelines governing the Protection of Privacy and Transborder Flows of Personal Data from 2013. Ali Gholami et al. have identified the following principles: lawfulness, informed consent, purpose binding, data minimization, data accuracy, transparency, data security, and accountability[1]. The European Commission's Data Protection Directive (95/46/EC) also defines health data as a special category of data to which a higher level of data protection applies. Therefore when dealing with health data the legal, ethical and regulatory aspects have to be carefully considered. Nevertheless, the Data Protection Directive gives no clear definition about what identifying information actually is, and each Member country is free to implement the Directive with its own rules, so-called "safeguards". Twenty years on, a (probably) final draft of a much more binding General Data Protection Regulation (GDPR), to be directly applied to all Member states (and passing no more through national reception legislation), is currently being discussed at the EU level, but is only expected to be issued during the course of 2016.

In the absence of a comprehensive legal framework to refer to, and waiting for assessing the likely impact of this forthcoming GDPR on clinical data management in particular, MD-Paedigree and CARDIOPROOF aim nonetheless at guaranteeing higher privacy standards in collection, storage, use and re-use of patient data, as provided for by the following criteria.

1. **Patient data stay and remain within hospitals**, its management is subject to applicable national rules and ethical committees approval. In general, the following principles are to be respected:

    - Consent: informed consent is the most likely derogation from the general prohibition of the processing of personal health data settled in Article 8 (1) of the Data Protection Directive;

    - Transparency: clear information about the usage of data will be provided to the patients.

    - **Purpose limitation:** in order to ensure the precise limitation of further processing of these datasets for different purposes, by defining "clear compatible and legitimate purposes of the data processing" and thus minimizing the risk of misuse of the data.

[1] Ali Gholami, Anna-Sara Lind, Jane Reichel, Jan-Eric Litton, Ake Edlund, Erwin Laure, Privacy Threat Modeling for Emerging BiobankClouds, Journal of Procedia Computer Science, Volume 37, Pages 489–496, 2014, p. 493.

- **Data minimization:** it means to only collect, process and store the personal data that is absolutely necessary for the purpose of the collection.

In accordance with the abovementioned principles, a still detailed but wider approach to "**enhanced consent**", meant to allow also for research purposes going beyond the strict duration of the MD-Paedigree and CARDIOPROOF project, and bringing back to the patient further useful health information, may be considered and applied wherever possible.
In particular, as highlighted by MD-Paedigree's Ethical and Legal Committee, "clarity and transparency towards the participants involved in the research (i.e., patients/parents) about ways that their data are, and may be used in the future, in dynamic unforeseeable scenarios, along with a realistic acknowledgement that no system can *a priori* guarantee in a complete way security, privacy and confidentiality in all circumstances" is recommended.

MD-Paedigree's Ethical and Legal Committee also highlighted specific ethical issues to be taken into account:

- redesign of consent forms to include information on data collection, data access, retention and security measures to be adopted;
- assent of children in older ages groups to be obtained;
- information about what happens once children reach the age of majority (will their consent be sought to continue to retain/use collected biological material and/or data?);
- consent for storage and future use must be obtained, if biological material to be collected is to be retained in a biobank;
- confirmation that use of previously collected data from Health-e-Child and Sim-e-Child, in the current study was covered by the original consent forms.

2. **Data is made available**, including for internal project purposes, **only in de-identified forms and following the "privacy-by-design" principle**. **De-identification** is comprised of two levels:

   A. A **first level of de-identification** is requested from clinicians before they submit patient data to the repository.

- According to HIPAA, which still is, as yet, the strongest worldwide de-identifying constraint expression, the Privacy Rule allows a covered entity to de-identify data by removing all 18 elements that could be used to identify the individual or the individual's relatives, employers, or household members; these elements are enumerated in the Privacy Rule[2]. Also the covered entity

[2] HIPAA http://privacyruleandresearch.nih.gov/pr_08.asp

must have no actual knowledge that the remaining information could be used alone or in combination with other information to identify the individual who is the subject of the information.

- Proper anonymisation or pseudonymisation can thus be reached:
  - either straightforwardly, by using a specially developed tool provided by gnúbila (Anonymizer), namely the data protection procedures addressing privacy issues, security issues and storage, provided for by gnúbila's FedEHR product[3]. This product is made up from 10 software modules that each address functional requirements in the field of data protection and confidentiality. That same software is being used by the Human Brain Project[4] for securing the data collected and shared in its Medical Informatics Platform.
  - or by following the step-by-step technical guidelines, based on HIPAA. Under this method, identifiers that must be removed are the following:

    1. Names.
    2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
       - The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people.
       - The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
    3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older. (*)
    4. Telephone numbers.
    5. Facsimile numbers.
    6. Electronic mail addresses.
    7. Social security numbers.
    8. Medical record numbers.
    9. Health plan beneficiary numbers.
    10. Account numbers.
    11. Certificate/license numbers.
    12. Vehicle identifiers and serial numbers, including license plate numbers.

[3] Federated Electronic Health Records (FedEHR) www.fedehr.com
[4] The Human Brain Project https://www.humanbrainproject.eu/

13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints. (**)
17. Full-face photographic images and any comparable images. (**)
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification. (**)

- For MD-Paedigree and CARDIOPROOF purposes, these 18 elements have been refined, adding research constraints for young children data:
  3bis. Dates can not be reduced to year when patients can be under 3 years old, it is not enough, months are important:
  (*) We decided to reduce all dates to Month instead of year.

- Some issues still require further clarification: (**)
  16. How X-nomics data have to be treated, they are by definition identifying data.
  16. GAIT is known (or assumed) to be identifying information.
  17. GAIT labs take videos of patients, a reliable way to hide/scramble faces on videos still needs to be found.
  18. Physicians prefer using identification numbers and don't feel comfortable in using unmemorable codes generated by computers.
  It is important to take care of this need and find a way to get back to these unmemorable codes and show patient ID when we are in a clinical centre where re-identification of data is possible.

In accordance with these criteria, in DICOM anonymisation, the following rules will apply:
- Metadata is modified to keep month information for dates (see: HIPAA validated norm DICOM PS3-15 [5])
- Acquisition to be performed by acquisition tool, the zones to blank on images will be defined and applied.

B. A "**quarantine area" is provided for de-identified data quality control**

Once the data has been de-identified according to one of the above-mentioned approaches, it is deposited in a quarantine area only accessible to local data managers for final quality control. The data manager can thus check if anonymisation was properly applied and make the decision whether to share the resulting anonymous data or not.

[5] http://medical.nema.org/dicom/2013/output/chtml/part15/PS3.15.html