Defining the best architecture for secure data exchange of diabetes information in Europe: privacy impact assessment in the BIRO project

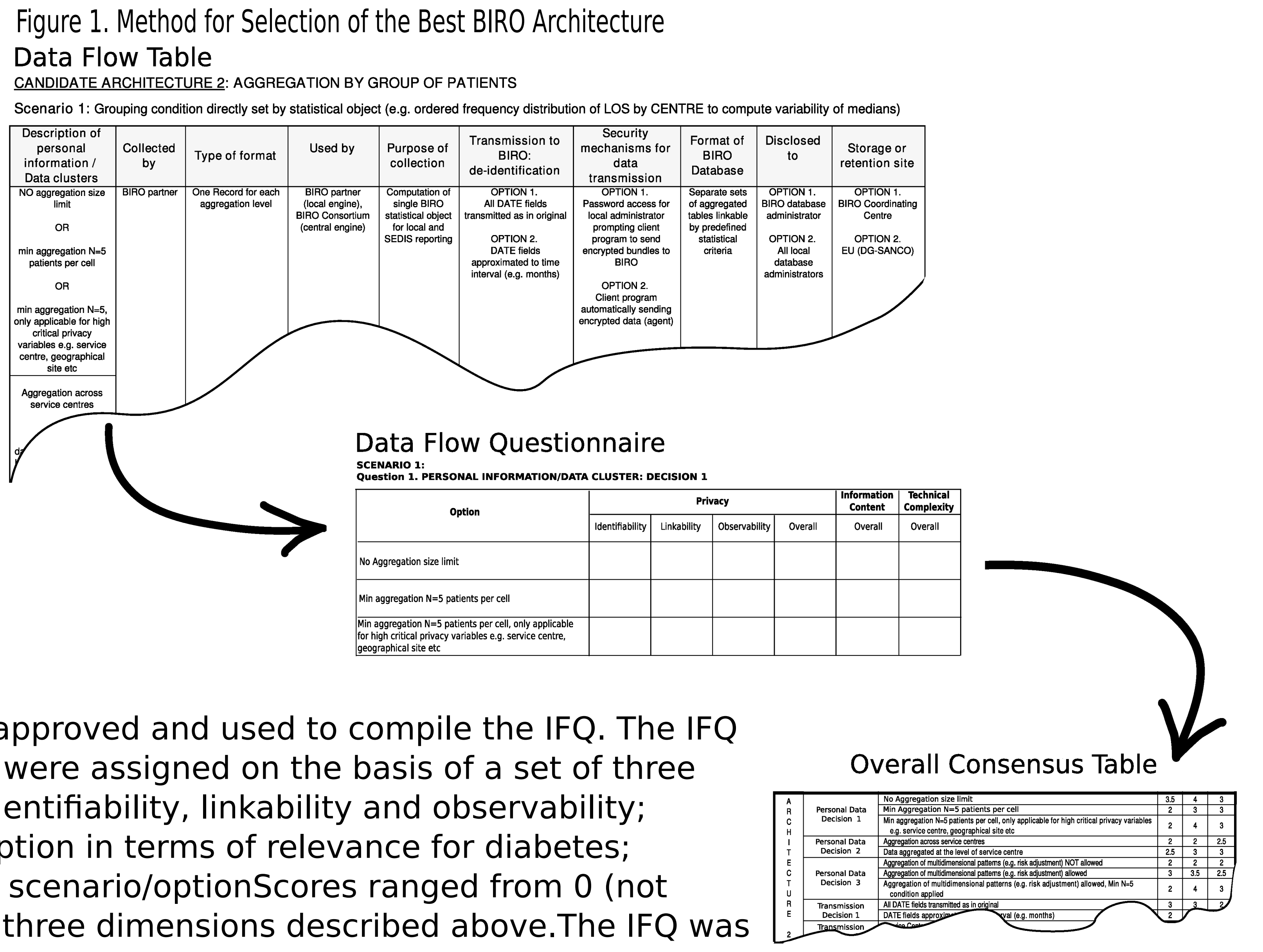
C. Di Iorio [1], F. Carinci [1], J. Azzopardi[2], P. Beck [3], S. Cunningham [4], S. Skeie [5],
G. Olympios [6], S. Pruna [7], M. Massi Benedetti [8], on behalf of the BIRO Consortium*

*Serectrix snc (Pescara, Italy) [1], University of Malta (G'Mangia, Malta) [2], Joanneum Research (Graz, Austria) [3], University of Dundee (Scotland) [4],
NOKLUS (Bergen, Norway) [5], Ministry of Health (Nicosia, Cyprus) [6], Telemedica Consulting (Bucharest, Romania) [7],
Paulescu Institute (Bucharest, Romania), University of Bergen (Norway), University of Perugia (Italy) [8]

Aim. The right to privacy raises the question of how far a society can intrude into the personal lives of its citizens, while ensuring that a societal fundamental goal e.g. public health can be safeguarded. Despite of the many international instruments available as common terms of reference in the protection of privacy, issues surrounding data protection had been rarely addressed by design in the field of health information systems. The BIRO Information System involves the use of sensitive-medical data collected through diabetes registries within national boundaries and further processed for public health studies at the international level. The "Privacy Impact Assessment" (PIA) of the BIRO health information system envisages an in depth exploration of the topic and supports researchers and software engineers in the construction of a privacy protective system architecture.

Methods. There is no unique definition of PIA in the literature. It has been defined as a "process whereby a conscious and systematic effort is made to assess the privacy impacts of options that may be open in regard to a proposal. An alternative definition might be that a PIA is an assessment of any actual or potential effects that the activity or proposal may have on individual privacy and the ways in which any adverse effects may be mitigated. The adoption of PIA in B.I.R.O. seemed convenient and cost effective, as it allowed for privacy risks and concerns to be minimised by design. Ex-post adjustments were inherently reduced through the incorporation of mitigation strategies directly in the system design, whenever privacy risks could not be fully avoided. A multidisciplinary dedicated "PIA Team" (PT) was formed to carry out a structured procedure involving four consecutive steps: preliminary privacy impact assessment, data flow analysis, privacy analysis and PIA report. The **preliminary** part included a discussion on the data flow, focusing on the physical/logical separation of personal information/data. It involved a systematic review of the privacy literature, whose search strategy included use of Ovid Medline with criteria: {privacy AND [(registr* OR register) OR (health information system*) OR (health database*)]}, and limits [human AND English Language AND yr = 2001-2006]. A core set of fourteen papers was selected by comparing abstracts against main project objectives. Papers were reviewed by the PT to complete a comprehensive report of the first step and identify a short list of possible candidate architectures. The second step involved a **data flow analysis** for each of the alternatives identified. A Delphi Consensus Procedure was undertaken by the PT to define the best alternative by producing the following materials: data flow tables (DFT), including the possible scenarios for the collection, use and disclosure of personal information/data, with a number of possible options; an information flow questionnaire (IFQ), to assign marks to each scenario/option; an overall consensus table (OCT), ranking scenarios/options. Materials were assembled using the procedure presented in *figure 1*. DFTs were initially prepared by the PF and revised by the whole PT. They were finally approved and used to compile the IFQ. The IFQ provided a series of scenarios, broken down into separate sub-options, for any of which marks were assigned on the basis of a set of three essential criteria: privacy: the score on privacy was further split into three separate criteria: identifiability, linkability and observability; information content for diabetes: based on the information provided by the specific scenario/option in terms of relevance for diabetes; technical complexity (feasibility): based on the feasibility of the implementation of the specific scenario/optionScores ranged from 0 (not applicable) to 5 (high level).The overall mark for each option was based on the average of the three dimensions described above.The IFQ was distributed to the PT and each member was asked to assign independent marks to each variable. All results were included in the OCT, presenting options ranked by overall scores, with ties ranked by increasing threat to privacy. The best architecture was defined as the mix of best options for all dimensions examined. **Privacy analysis** covered any privacy issue arising in the transfer of data from the local centres to the central database. Potential privacy risks have been identified and thoroughly analysed through a summary table indicating mitigation strategies to be implemented. The level of risk has been classified according to an ordinal scale of intensity. The **final report** compiles results from all phases using a structured format.

Figure 1. Method for Selection of the Best BIRO Architecture



Data Flow Table

CANDIDATE ARCHITECTURE 2: AGGREGATION BY GROUP OF PATIENTS

Scenario 1: Grouping condition directly set by statistical object (e.g. ordered frequency distribution of LOS by CENTRE to compute variability of medians)

Description of personal information / Data clusters	Collected by	Type of format	Used by	Purpose of collection	Transmission to BIRO: de-identification	Security mechanisms for data transmission	Format of BIRO Database	Disclosed to	Storage or retention site
NO aggregation size limit OR min aggregation N=5 patients per cell OR min aggregation N=5, only applicable for high critical privacy variables e.g. service centre, geographical site etc	BIRO partner	One Record for each aggregation level	BIRO partner (local engine), BIRO Consortium (central engine)	Computation of single BIRO statistical object for local and SEDIS reporting	OPTION 1. All DATE fields transmitted as in original OPTION 2. DATE fields approximated to time interval (e.g. months)	OPTION 1. Password access for local administrator prompting client program to send encrypted bundles to BIRO OPTION 2. Client program automatically sending encrypted data (agent)	Separate sets of aggregated tables linkable by predefined statistical criteria	OPTION 1. BIRO database administrator OPTION 2. All local database administrators	OPTION 1. BIRO Coordinating Centre OPTION 2. EU (DG-SANCO)
Aggregation across service centres									

Data Flow Questionnaire

SCENARIO 1:
Question 1. PERSONAL INFORMATION/DATA CLUSTER: DECISION 1

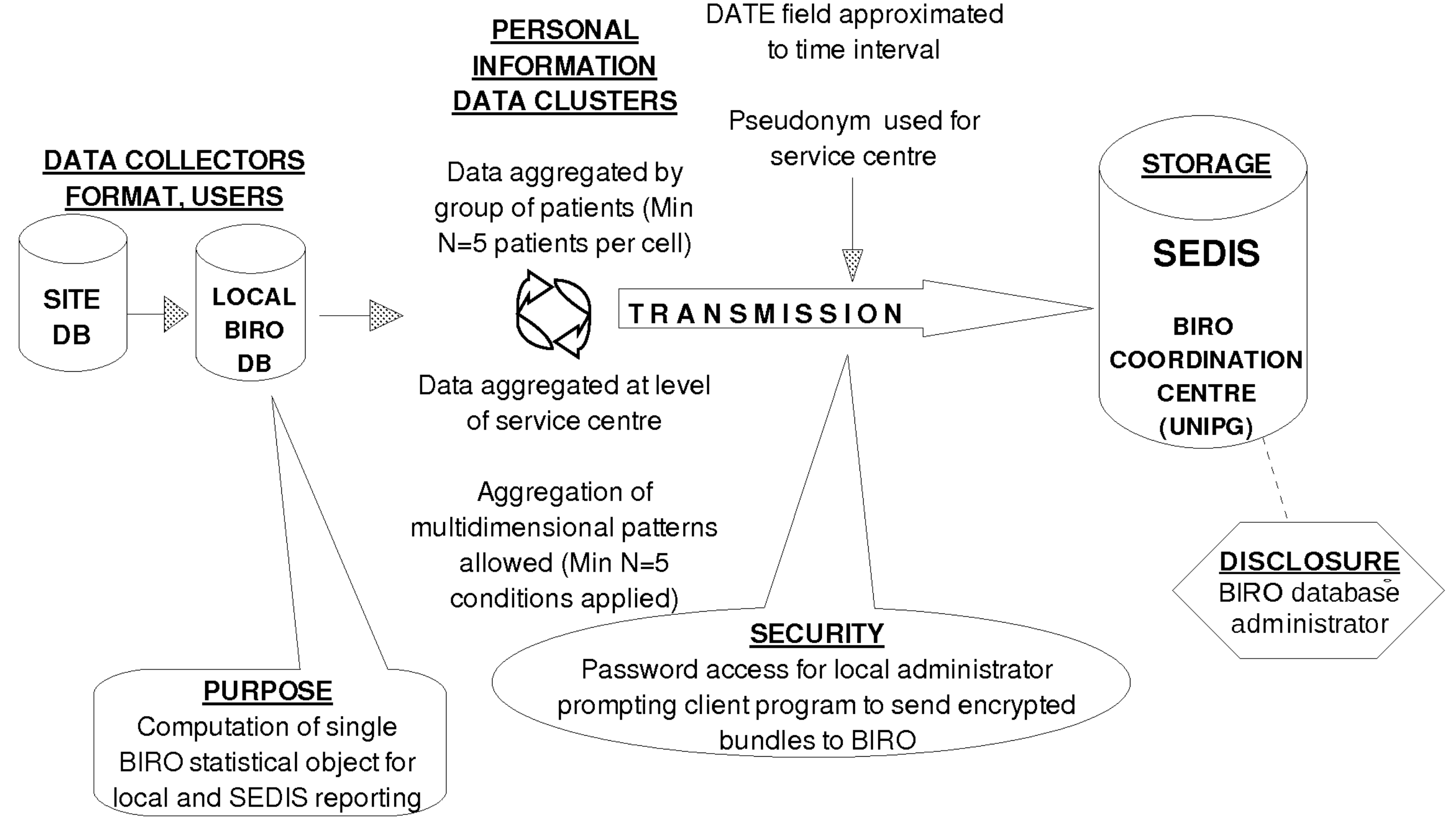
Option	Privacy				Information Content	Technical Complexity
	Identifiability	Linkability	Observability	Overall	Overall	Overall
No Aggregation size limit						
Min aggregation N=5 patients per cell						
Min aggregation N=5 patients per cell, only applicable for high critical privacy variables e.g. service centre, geographical site etc						

Overall Consensus Table

ARCHITECTURE 2	Personal Data Decision 1	No Aggregation size limit	3.5	4	3
		Min Aggregation N=5 patients per cell	2	3	3
		Min aggregation N=5 patients per cell, only applicable for high critical privacy variables e.g. service centre, geographical site etc	2	4	3
	Personal Data Decision 2	Aggregation across service centres	2	2	2.5
		Data aggregated at the level of service centre	2.5	3	3
	Personal Data Decision 3	Aggregation of multidimensional patterns (e.g. risk adjustment) NOT allowed	2	2	2
		Aggregation of multidimensional patterns (e.g. risk adjustment) allowed	3	3.5	2.5
		Aggregation of multidimensional patterns (e.g. risk adjustment) allowed, Min N=5 condition applied	2	4	3
	Transmission Decision 1	All DATE fields transmitted as in original	3	3	2
		DATE fields approxim[illegible]rval (e.g. months)	2		
	Transmission				

Results. The accomplishment of PIA tasks provided essential input for the development of all major components of the B.I.R.O. system. Three main candidate architectures were identified, with differing levels of data sharing. The first alternative required the transmission of "individual patient data, de-identified through a pseudonym", secured by an encryption algorithm and privacy protective communication technologies. The second alternative envisaged data shared as "aggregation by group of patients, with Centre's IDs available in de-identified form, securely encrypted", transferred using privacy protective solutions. he third alternative was based on "Aggregation by Region", optimised to impede reverse engineering, with the usual secure data transfer. Details of the three alternatives were used to compile the DFTs and DFQ. The Delphi panel selected the best alternative by ranking the three alternative scenarios, including options for their implementation. The resulting B.I.R.O. system architecture is shown in *figure 2*, whose criteria were duly taken into account for implementation.

Figure 2. Selected BIRO Architecture



DE-IDENTIFICATION
DATE field approximated to time interval

PERSONAL INFORMATION DATA CLUSTERS
Data aggregated by group of patients (Min N=5 patients per cell)

Pseudonym used for service centre

DATA COLLECTORS FORMAT, USERS

SITE DB

LOCAL BIRO DB

TRANSMISSION

STORAGE
SEDIS
BIRO COORDINATION CENTRE (UNIPG)

Data aggregated at level of service centre

Aggregation of multidimensional patterns allowed (Min N=5 conditions applied)

DISCLOSURE
BIRO database administrator

SECURITY
Password access for local administrator prompting client program to send encrypted bundles to BIRO

PURPOSE
Computation of single BIRO statistical object for local and SEDIS reporting

Conclusion. The BIRO Information System involves the use of sensitive-medical data collected through diabetes registries within national boundaries and further processed for public health studies at the international level. According to the BIRO architecture, participating centres apply procedures for data anonymisation before any transfer to the BIRO central database is made. The central server processes aggregate records solely for statistical and scientific purposes. According to Recital 26 of the EU Data Protection Directive, anonymisation allows personal data processing without consent, placing anonymous data outside the scope of the data protection principles therein contained. The BIRO system processes only statistical objects that are stored as aggregate table into flat text comma delimited files. Hence, there is no possibility, according to the state of the art, to identify a patient, either directly or indirectly, with a reasonable effort. Clinical centres also receive privacy protection through the use of pseudonyms. Aggregate data are processed by the local database engine and sent to the central statistical engine through an "ad hoc" communication software ensuring secure information exchange and compliance with security requirements enshrined in EU and international data protection norms. Considering that data are rendered anonymous by local BIRO centres and transmission occurs in a secure environment, the further processing at the level of the global statistical engine cannot pose any privacy risk, either directly or indirectly. Trans-border data flow envisaged in BIRO is legally viable according to the EU legislation. Publication of project results is performed to avoid any direct/indirect identification of data subjects and/or local centres. Potential privacy risks have been analysed through a summary table, which allows to estimating the better privacy protective alternative in the processing of data.At a general level, the BIRO Information System processes only de-identified data. Hence, the level of risk can be considered, in most of the cases described, low. As highlighted in the privacy summary table (*Table 1*), efficient mitigation strategies have been implemented in the context of BIRO. Consequently, the aforementioned potential privacy risk could be considered fully avoided and/or removed.

Privacy impact assessment shows that the selected BIRO architecture fulfils privacy protection requirements by addressing and resolving broad privacy concerns from different angles. The architecture of the system flexibly affords the best privacy protection in the construction of an efficient model for the continuous production of European diabetes reports. The methodology identified can be usefully applied in other fields of health information, particularly where disease registers are involved as primary units for data collection and statistical analysis.

Table 1 Privacy contingency risks

		Level of risk*				
Element	Nature of risks	Low	Medium	High	Comments	Mitigating mechanisms
Individual data: pseudonym used for patients' IDs + data aggregated (n = 5 patients per cell)	Individual privacy	X			Pose an indirect risk to individual's privacy	Non-reversible de-identification
Pseudonym used for centres' IDs	Non-individual privacy		X		Pose an indirect risk to centres' privacy	Reversible de-identification + reporting system: percentage
Data transmission	Security measures	X			Pose an indirect risk to individual's privacy	Encryption
Access to the BIRO network	Security measures		X		Pose an indirect risk to individual's privacy	Secure applications; hacking tests
Global statistical analysis	Individual privacy + non-individual privacy + security measures	6			Pose an indirect risk to individual's privacy and centres privacy	Non-reversible de-identification + encryption

*Low, risk can materialise but mitigating factors exist; moderate, risk is likely to materialise if no corrective measures are taken; high, there is a high chance that negative effects will materialise if no corrective measures are taken.
BIRO, Best Information Through Regional Outcomes; ID, identification.

All documentation available at: http://www.biro-project.eu/results.htm