Meaningful Integration of Data, Analytics and Services

Grant Agreement No. 727721

Contract Duration: 40 months (1st November 2016 – 29th February 2020)

This project is funded by
The European Union

H2020-SC1-2016-CNECT
SC1-PM-18-2016 - Big Data Supporting Public Health Policies

Deliverable 2.1

Good Practice Report 1

Circulation:	PU (Public)
Nature:	R (Report)
Version #:	7.0
Issue Date:	22/10/2017
Responsible Partner(s):	South Eastern Health and Social Care Trust
Author(s):	Paul Carlin, Dale Weston, Natasha Bloodworth, Debbie Rankin, Michaela Black, Jarmo Paakkonen, Juha Pajula
Status:	Final
Reviewed on:	11/12/2017
Reviewed by:	MIDAS Executive Board
Contractual Date of Delivery:	31/10/2017 (M12)

Grant Agreement No: 727721

Executive Board Document Sign Off

Role	Partner	Signature	Date
Project Coordinator	Ulster	Michaela Black	04/12/17
WP1 Lead	Ulster	Michaela Black	04/12/17
WP2 Lead	SET	Paul Carlin	11/12/17
WP3 Lead	VICOM	Gorka Epelde	04/12/17
WP4 Lead	KU Leuven	Gorana Nikolic	11/12/17
WP5 Lead	VTT	Juha Pajula	08/12/17
WP6 Lead	DCU	Regina Connolly	29/11/17
WP7 Lead	Ulster	Johnny Wallace	29/11/17
WP8 Lead	Ulster	Michaela Black	04/12/17
Scientific-Technical Manager	Austin Tanney (Analytics Eng)	Austin Tanney	04/12/17

Grant Agreement No: 727721

Abstract

T2.1 – Review and Analyse the Current Legislative and Good Practice (Lead: SET; All partners)
The description of the task is as follows: The work package group will work with the following target group(s): citizens, legislators, practitioners, health care managers, company managers, data analysts, academics, policy leads, health care professional groups and media to understand current legislative and best practice regarding consent, ethics, privacy, and access to Personal Identifiable Data (PID). The current literature will also be reviewed and analysed with the materials gathered through the engagements with the target groups. In addition, as part of this work, the public's perceived acceptability of different health and security-related uses of their Personal Identifiable Data (PID) will be assessed and collated.

This report will provide a review of the initial findings of a review of the literature with regard to the following areas:

Legislation
- Legislative frameworks

Good Practice
- Good practice models (that will lead to recommendations in the subsequent version of this deliverable, D2.2 Good Practice Report 2)

Current Models of Consent
- Current models of consent
- Data protection
- Data usage
- Data storage
- Data linkage

Current Models of Communication
- Current models of communication

Perceptions
- Public/private innovation in regard to data use
- Public perceptions concerning different uses of Personal Identifiable Data (PID)

Primarily this report will focus on the review of the literature. Whilst some work has been completed on engagement with a review of current partner arrangements using a questionnaire (Appendix 3), as well as a small focus group with data protection practitioners within government which remains anecdotal as it was unable to be recorded, work remains to be completed in respect of the review with the public, healthcare professionals, academics, policy leads and media. A large part of this will be informed by the session: MIDAS Consent, Ethics and GDPR workshop taking place in Belfast on 17 November 2017 (M13). This will be reviewed and will inform deliverable 2.2. A formal focus group is to be held within Northern Ireland with policy and Information Governance leads to discuss models at M16. This will replicated in England, Spain and Finland, as will a focus group for clinical staff at month 17. This work will be supplemented by PHE's work which is currently being finalised and will form D2.4, along with a piece of work using Twitter initially, as a tool for gaining meaningful insight into the creation, perception and acceptance of any model of use.

Grant Agreement No: 727721

Copyright

The MIDAS project is funded under the EC Horizon 2020 SC1- PMF-18 Big Data Supporting Public Health Policies

Grant Agreement No: 727721

Document History

Version	Issue Date	Stage	Content and Changes
v1.0	18/10/2017	Draft	Draft version for review by WP2 members
v2.0	22/10/2017	Draft	Draft version for review by consortium
v3.0	29/10/2017	Draft	Draft version for review by consortium
v4.0	02/11/2017	Draft	Updated draft version for review by consortium
v5.0	12/11/2017	Draft	Updated draft version for review by consortium
v6.0	29/11/2017	Draft	Updated draft version for review by consortium
v7.0	11/12/2017	Final	Final version approved by Executive Board

Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Grant Agreement No: 727721

Executive Summary

Work Package:	WP 2
Work Package leader:	Paul Carlin
Task:	T2.1 - Review and Analyse the Current Legislative and Good Practice
Task leader:	SET (Paul Carlin)

Table of Contents

1 Legislation

Legislation and adjudication must follow, and conform to, the progress of society. (Abraham Lincoln)

Legislation is the process or enactment of law, law being a principle, rule or guideline developed to guide social intercourse. These rules ensure that all members of society have a framework in which to exist and function, a framework that for the most part one would hope reflects our better selves. That being said, legislation can also reflect poorly on society, due to poor framing, interpretation and ideological perversity. One needs only examine the Nuremberg Laws (passed in September and November 1935) (Noakes & Pridham, 1974), which enshrined discrimination based on religion and race, leading to a society that permitted the worst excesses of human nature.

Legislation, its mechanism of construction, enactment and regulation are therefore essential in governing and managing systems of relationships, i.e. society (Younkins, 2000). Whilst one would aspire to perfection, in terms of the rules and regulation that control our abilities to act as individuals and members within a larger collective, there are obvious complexities and difficulties in achieving this aspiration.

Not least of these complexities and difficulties is the language in which law is framed, potentially creating difficulties with the interpretation of law, as well as enactment. This is of particular relevance, when considering the interrelationship and interdependence between different laws and different jurisdictions. This has specific resonance within a concept and organisational arrangement as large as the European Union (EU).

The context that the society exists within also has significance, in that, law and its enactment through a legislature can and often is relevant to time and place. This has become more and more obvious over the course of the last 50 years, when technology has moved at such a pace as to outstrip the legislative process (Wiener, 2004). This can lead to legislative frameworks, that at best fail to ensure that technology is exploited to maximise benefit for all, or at worst to, allows exploitation or lack of protection for the individual (OECD, 2000).

When one explores the idea of big data, one can be blinded by the anticipated benefits such as epidemiological tracking, precision medicine, economic modelling and forecasting, and banking (Chen, Chiang, & Storey, 2012), as well as the

potential in MIDAS, through policy integration. That being said, there are risks (Lenard & Rubin, 2015), such as:

- Improper use, leading to targeted discrimination for individuals and groups
- Poor control and regulation leading to data breaches, and identification / re-identification of individuals
- Market manipulation
- Privacy erosion

Therefore, the aim of any legislation program and regulatory framework for the use of data, at any scale, should be one, that exploits the benefit of the data and analytics, whilst ensuring the protection of the individuals needs and rights as defined within broader societal legislative constructs, such as the European Convention of Human Rights (ECHR) (Council of Europe: European Court of Human Rights, 1950), enshrined within the European Union in the EU's Charter of Fundamental Rights (European Union, 2012).

Within the ECHR, there are 18 articles and whilst all, to some extent may be informed and impacted by the use of data, two have a particular relevance, and potential to compete in relation to big data, **Article 8 – Privacy, and Article 10 – Expression.**

The notion of privacy and data (electronic and large volume) has been around since the latter half of the 20th century, with a number of reports and papers identifying changes in practice and a growing recognition of some of the issues arising from the use of personal data (Westin, 1979), as well as, reaffirming the notion of individual freedom for both persons and groups as a necessary component of any functioning democracy, in regard to the use of data (Cheng, 1969).

Building upon the Organisation for Economic Development's (OECD) recommendations that are articulated in the **Guidelines Governing the Protection of Privacy and Trans-Border Flows of Personal Data** (The Organization for Economic Co-Operation and Development, 1980) which describes 8 key principles:

- Collection Limitation Principle - There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject
- Data Quality Principle - Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date

- Purpose Specification Principle - The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose
- User Limitation Principle - Personal data should not be disclosed, made available or otherwise used for purposes other than those specified, except:
    a) with the consent of the data subject; or
    b) by the authority of law
- Security Safeguards Principle - Personal data should be protected by reasonable security safeguards against risks such as loss or unauthorised access, destruction, use, modification or disclosure of data
- Openness Principle - There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.
- Individual Participation Principle -
    a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him;
    b) to have communicated to him, data relating to him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him;
    c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and
    d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended
- Accountability Principle - A data controller should be accountable for complying with measures which give effect to the principles stated above,

Within the consortia, all partners jurisdictions have created systems for the use of data that is controlled according to statutory conditions for scientific research, statistics, government control and regulatory oversight, as well as the planning and investigation tasks of the authorities.

Within the context of registries, different registry administrators need to apply permissions for use of registry data for the same research and development projects individually and information is provided from different data registers on the basis of provisions in different laws. Population-based register data permit a comprehensive

Grant Agreement No: 727721

assessment of outcome on several levels, e.g., at the individual, institutional and population-based levels.

The Finnish government has proposed a legislative amendment aimed at ensuring that client and personal data registered in the social and health care sector in the future should be used as smoothly and safely as possible for the various purposes permitted by law. This legislative changes are under decision making and targeted to be in effect at the start of the year 2018. The changes propose that the data could be used also for teaching, information management, and for development and innovation. The permission for personal data use shall be granted by a single authority when data from several different administrators or from private social and health care needs to be compiled.

The European Union created the **Data Protection Directive (Directive 95/46/EC)**, which was translated into law in each of the member states, under seven principles as shown in Table 1.

Table 1: The seven principles (The European Parliament and the Council of the European Union, 1995)

Principle	
1	Subjects whose data is being collected should be given notice of such collection.
2	Subjects whose personal data is being collected should be informed as to the party or parties collecting such data.
3	Once collected, personal data should be kept safe and secure from potential abuse, theft, or loss.
4	Personal data should not be disclosed or shared with third parties without consent from its subject(s).
5	Subjects should granted access to their personal data and allowed to correct any inaccuracies.
6	Data collected should be used only for stated purpose(s) and for no other purposes.
7	Subjects should be able to hold personal data collectors accountable for adhering to all seven of these principles

Each of the partners within MIDAS operates within these codices described within the EU Data Protection Directive, enacted through each sovereign states legislature.

With the adoption of the **General Data Protection Regulations (GDPR)** by the Member states in 2016, and the regulations timeline set for implementation in May 2018, the securities offered by Data Protection Directive (The European Parliament and the Council of Europe, 2016) (Directive 95/46/EC) have become strengthened.

The protections offered by the GDPR can be summarised as follows:

- Harmonisation of regulation across the European Union
- Clear definition of “Personal Data”
- Separation of responsibilities between controller and processor
- Increased penalties for non-compliance
- Appointment of Data Protection Officers
- An understanding of privacy management
- Consent for processing
- Information must be provided at the point of data provision
- Consent for profiling
- Definition of legitimate interest
- Notification in respect of breach, which is not limited to theft
- The right of the individual to understand how their data Is being processed
- The right to the data being readily portable

Although the complexities and difficulties regarding e-commerce, privacy and the practicalities therein, will only become evident after use, our current understanding readily identifies key issues, described by Akter & Wamba (2016):

- Strategy, culture, leadership and organisation
- Marketing and sales
- Production and operations management
- Data quality, IT, infrastructure and security
- Human resources / talent management
- Overarching value

That being said, MIDAS proposes a system that ensures anonymity, limited intrusion and clear governance system to ensure adherence to legislation and regulation as described in Section 2 Good Practice, and the processes that have been enacted within the project to manage both the test data, and policy data that will be used as exemplars for the project.

Grant Agreement No: 727721

2 Good Practice

It is the framework which changes with each new technology and not just the picture within the frame.
(Marshall McLuhan)

Good / Best practice can be defined as “a working method or set of working methods that is officially accepted as being the best to use in a particular business or industry, usually described formally and in detail” (Cambridge English Dictionary, 2017). Within MIDAS, Work Package 2 (WP2) is tasked with delivering a framework that utilises technology, large volume data, policy imperatives and operational procedure to ensure that the data used in the project meets all legislative requirements and good practice requirements. In addition to this, there is an appreciation and recognition of the importance of ethical oversight, not only as an integral principle of the project management, but also as a core deliverable of the project outputs.

To better understand the notion of good practice, one is required to appreciate the environment, defined largely by the legislation previously described, that places Big Data and analytics within context, as this brings better understanding.

Nominally this notion of good practice relates to governance, and as such there should be a mechanism for control and oversight that is inclusive as possible, whilst being as pragmatic as possible, there is a balance to strike (Figure 2.1).

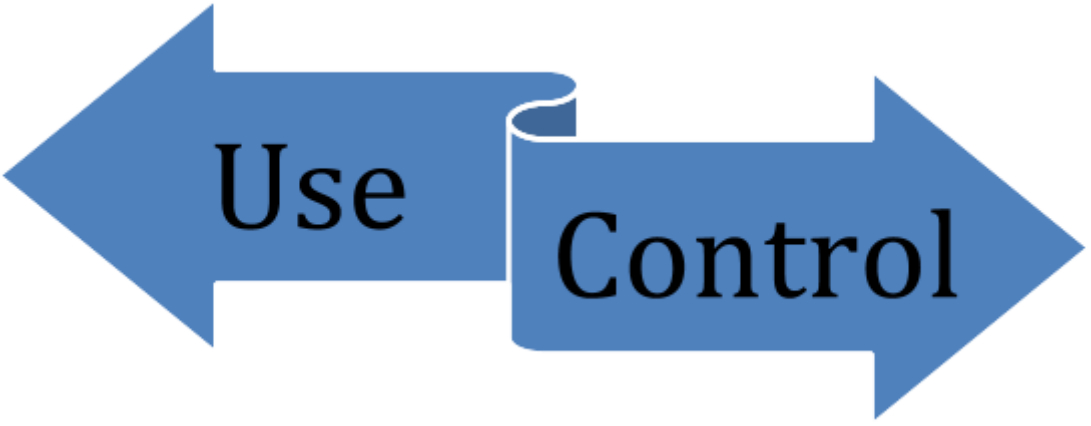


Figure 2.1 Balance

Control of data managed within a framework of good practice and informed by legislation, ethics and context, it would be hoped provides the scaffold to protect the

needs of all, both individuals, organisations and society as a whole. Pulling against this is the drive to **Use** data for a multitude of purposes.

2.1 Ethics

This notional control, is framed using the legislation certainly, but should also be constructed according to clear ethical principles and standards. Ethics, the moral principles that govern a person's behaviour or the conducting of an activity (Oxford English Dictionary, 2017) must inform these good practice requirements, if this is not the case, then there is potential for harm and malfeasance.

Ethical constructs adapt over time, and as such act as a mirror that reflects to a large extent, society at a particular junction (Calman, 2004). Certain inalienable precepts remain constant in civilised societies, for example the right to life, but philosophical and political ideologies can corrupt these, diminishing the scope and protections offered by society as a whole (Strous, 2007).

One can perhaps, argue that big data, and its analysis falls with the paradigm of research. Whilst ethical and good practice frameworks within research, are derived principally from the Nuremberg Code and are designed for use within medical research (Fischer, 2006), subsequently enshrined within the Declaration of Helsinki (World Medical Association, 2013), there are a number of professional and regulatory codes drawn from these foundation documents, for example the ethical guidelines of the British Psychological Society (The British Psychological Society, 2009) which inform the thinking and outputs from the Behavioural Science Research Team within Public Health England ((DH) PHE), a consortia member.

Whilst the notion of ethics is clearly identified with medical research, it is perhaps less obvious within the arena of Big Data, although clearly there are issues about privacy (Martin, 2015).

2.2 Anonymity

Whether the notions of anonymity, pseudo-anonymity and (personal) identifiability are directly related to research is perhaps a moot point. The General Data Protection Regulation (The European Parliament and the Council of Europe, 2016) is relatively clear in codifying the distinctions:

"The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information, should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes".

With this, the importance of the particular processes of pseudonymisation and anonymisation become essential in the context of any good practice, this must be within a context that takes cognisance of the particular analytic need, whilst appreciating the risk of linked data sets leading to potential individual re-identification (Cohen, Amarasingham, Shah, Xie, & Lo, 2014). Clearly the key risk is around the identifiable nature of any data for a natural person, and the protections used to anonymise or pseudonymised data (although in the case of the deceased one should be mindful of concepts of confidentiality and common or case law that defines access in particular jurisdictions, Recital 27 (The European Parliament and the Council of Europe, 2016), is used.

One would also need to be mindful of the potential for re-identification, techniques such as k-anonymity (A dataset is said to be k-anonymous if every combination of values for demographic columns in the dataset appears at least for k different records (Ted is writing things, 2017)) have been described (Sweeney, 2002) and built upon within the health sector, for example Datafly (Sweeney, 2016), yet significant risks remain, that have the potential to hamper the utilisation and utility of big data for policy wide benefit (Asche, Seal, Kahler, Oehrlein, & Baumgartner, 2017). Re-identification scientist scientists have demonstrated that they can often "re-identify" or "de-anonymize" individuals hidden in anonymized data (Ohm, 2010).

Grant Agreement No: 727721

This again raises the notion of balance, balance in regard to the relative risk versus benefit of use, but also the balance in terms of responsibility as controller or processor. There is risk in the very act of exchange, and regardless of the level of anonymity or otherwise, all parties must recognise and appreciate the responsibilities inherent in their roles. Identifiable data can of course be exchanged between parties, but only within the overall construct of consent.

The Data Protection Ombudsman in Finland proposes in their additional statement of big data usage (Dnro 3424/41/2015) that even when anonymized data is used the register owners should evaluate the actual level of anonymization in time to time in their functions. The MIDAS EPAG is seen as a consortium level author against re-identification of the individuals, and MIDAS data could be evaluated from that perspective during the MIDAS time (e.g. yearly).

So far, in regard to good practice a number of core principles have been described. These have referred to as ethics and anonymity primarily. Within these overarching principles, processes must align with these key outputs, for example: data processing, data responsibility, adherence to the legislation, as well as some form of ethical oversight.

2.3 Quality

Perhaps more fundamental than this requirement for anonymity, is the quality of any data that may be used within the policy cycle and how this quality can be assured. Obviously, a controller or user needs the skills and resources to manage the quality process, but, issues do exist, with a continued shortage of data scientists expected over the next 3 years (Miller & Hughe, 2017) even with attractive remuneration evident throughout the industry (Burtch, 2017). That being said, there needs to be a defined standard for those undertaking this work as a reference.

Common themes, relating to the dimensions of data quality are easily discernible across the literature and described succinctly by (Askham, et al., 2013), as shown in Table 2. These ensure a core standard that must be met for all those involved in the accessing and analysis of data for a defined outcome, i.e. the core concept within MIDAS, is to utilise disparate data sources to define policy, and drive the policy cycle thereafter. Without an assurance as to the overall quality of these sources, one would risks losing the integrity and thus the value of the data itself.

Grant Agreement No: 727721

Table 2: Quality Dimensions

Dimension	
1	Completeness
2	Uniqueness
3	Timeliness
4	Validity
5	Accuracy
6	Consistency

This has real significance within the context of policy. Policy should be both informed and evaluated, to ensure that the established evidence base is used to benefit society as a whole, as well as being reviewed on a regular basis to assess impact.
This reaffirms the need for assuring quality, particularly for high level policy.
Obviously, one needs credible and accurate sources, and within the model development phase of the project classification (accreditation, if you will) of source will be examined.

MIDAS therefore ensures this by leveraging technology to enhance this policy cycle by ensuring the right data is presented in the right way at the right time to the right person/team to produce the highest quality output. This is how the project has been designed and is operated, with key deliverables linked together for delivery throughout the project life cycle:

- WP2
  - D2.2 - Good Practice Report 2
  - D2.3 - MIDAS Framework User Guide
- WP3
  - D3.3/D3.4 - Data Interoperability, Representation Report and Architecture Design versions 1 and 2
  - D3.8 - Enterprise Data Virtualization Layer Pilot 1
- WP4
  - D4.2 - Health Policy Decision Outcome Simulator 1
  - D4.5 - Framework for Combining Expert Knowledge and Data-analysis 1
- WP5
  - D5.1/D5.2 - Visual Analytics Tool(s) Concept versions 1 and 2

Grant Agreement No: 727721

2.4 Intuition

Whilst the idea of big data as a tool to drive policy seems self-evident, there are perhaps dangers that might have significant impact on our ability to formulate policy intuitively through over-reliance on data analytics. Adopting a data analytics approach wholesale would minimise the importance of theorising and intuitive extrapolation with regard to patterns of correlation established through automated analysis driven through algorithms (Bollier, 2010). It also raises risk, as automated systems without strong systems of interrelated processing between human and machine may not be adaptive enough to meet shifting ethical and privacy needs i.e. the system has no ethical framework to drive analysis and use to test ethical standards against, and even it it did, without adaptive design algorithms will quickly become redundant as society and acceptability/ perception changes.

One would suggest that there is a requirement that the right policy, within the right political context, must be a requirement for good practice, the palatability of a particular policy implementation and the data needed to drive it, must be a consideration, otherwise this work will wither on the vine before any meaningful implementation is carried out.

2.5 Good Practice Summary

Clearly, policy in and around the use of big data has importance for the MIDAS project, and will impact on how any solution may operate within a real world context. We have already made reference to how the data from the individual is made available, within three broad domains, anonymised, pseudonymised and identifiable data, and how this is described within the legislative and good practice domains. Each of these domains brings their own issues and challenges, yet one theme is relevant to all, **consent**.

Grant Agreement No: 727721

3 Current Models of Consent

Gregory (Scotland Yard detective): "Is there any other point to which you would wish to draw my attention?"
Holmes: "To the curious incident of the dog in the night-time."
Gregory: "The dog did nothing in the night-time."
Holmes: "That was the curious incident.
(Doyle, 1892)

Consent is defined as “Permission for something to happen or agreement to do something” (Oxford English Dictionary, 2017). This is certainly not a definition one would use in the context of healthcare. Within the healthcare setting, consent is only considered valid when classified as informed, defined as, “Permission granted in full knowledge of the possible consequences, typically that which is given by a patient to a doctor for treatment with knowledge of the possible risks and benefits” (Oxford English Dictionary, 2017).

Within the current legislative and good practice domains there are clear distinctions for data use outside of the provider's own control, that are broadly defined as that which is identifiable and non-identifiable data. Whilst, the GDPR does NOT require formal consent for data that is non-identifiable (albeit with caveats in regard to preparation and post processing), there remain significant requirements in respect of confidentiality, ethics and privacy (The European Parliament and the Council of Europe, 2016).

Once again there are competing demands, on the one hand, the need to respect and meet the rights of individual, whilst on the other the need to exploit the readily accessible data provided by the individual for business and development, all of which relate to the societal controls framed within legislation and good practice. These competing drives and demands are described in Figure 3.1.

Yet, the individual exists as part of that wider societal whole, and indeed, benefits in no small part from the innovation brought about by leveraging information from that collected by companies providing technology and/or services. Within this model, terms and conditions are agreed to by the end user described within the licence agreement, although the notion of informed consent is somewhat belittled by the scale, scope and construction of these End User Licence Agreements (EULA) (Solove, 2012), with only 20% of signatories actually reading the agreement

Grant Agreement No: 727721

(Nissenbaum, 2010), (Jensen, Potts, & Jensen, 2005). It raises the question of how important or otherwise the majority of users take the notion of privacy and control, particularly with regard to electronic systems and big data. The idea of public perception as a constraint for access will be discussed in Section 5.

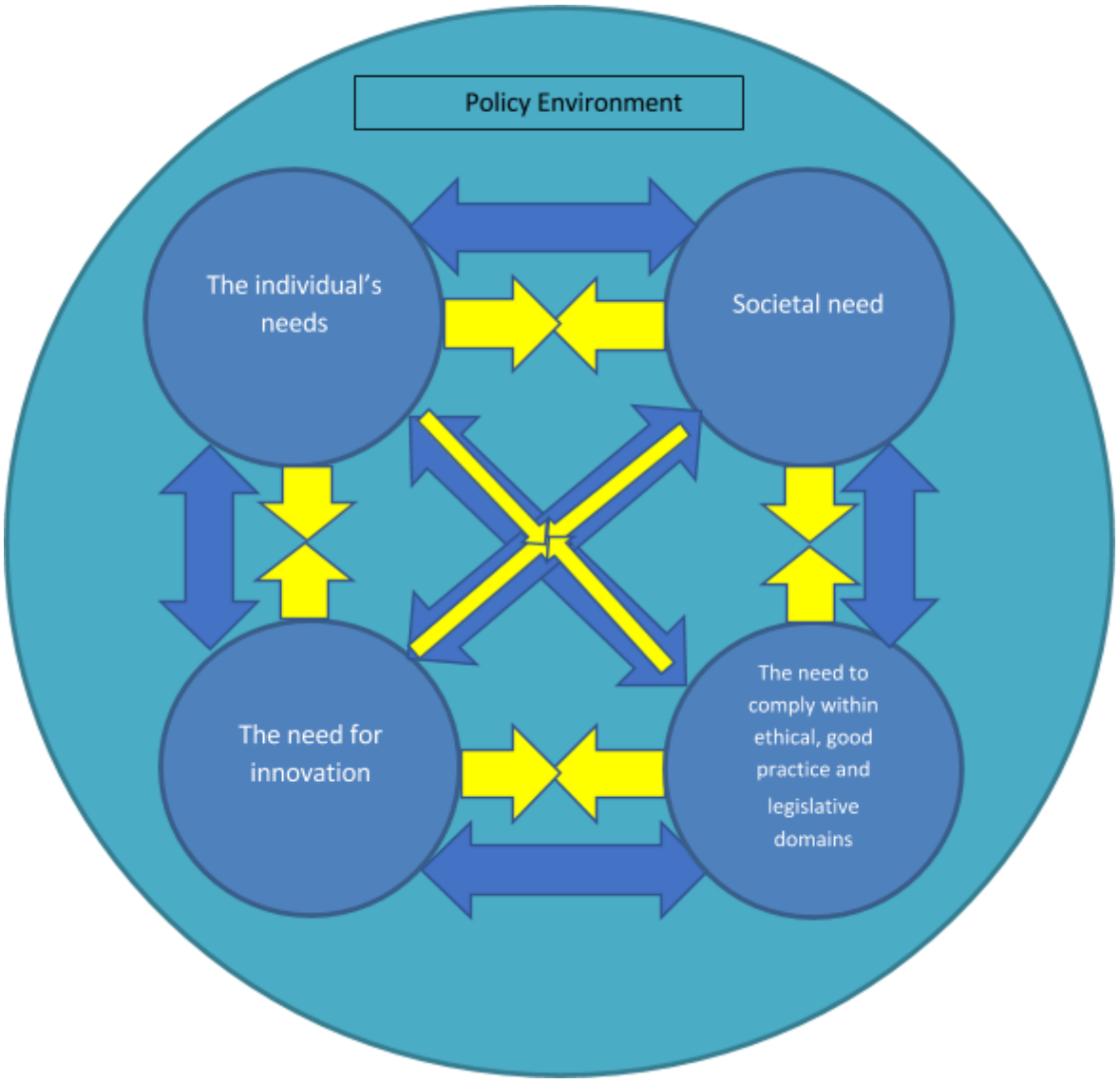


Figure 3.1 Push and Pull (Yellow arrow - Pull, Blue arrow - Push)

3.1 Process

There appears to be no real process or structure to how this informed consent is sought, processed and reaffirmed when examined in relation to Big Data, although this is a core concern and focus for the MIDAS project. Within the project structure,

the Policy Board and EPAG groups ensure relevant and appropriate oversight of all data use, creating a project specific governance structure.
Each of the MIDAS policy board members inform their use of data through adherence to legislation and the organisational policy and procedure relating this to ensure compliance. Whilst this might meet the standards as defined, one would contest that this leads to an over reliance on legal definitions rather on the spirit of the law.

To reiterate, each organisation/region will have distinct processes in place, to ensure good practice but this is dependent on the use case and indeed the interpretation of the legislation, and one off agreements as defined in EULA's, highlight the less than perfect process and dynamic nature of consent within the private sector. It is clear that for research purposes anonymised data can and is used, yet, information obtained directly for clinical trial purposes, that is then anonymised for this specific purpose requires direct consent, within an established framework to ensure meaningful understanding. One is at a loss to understand the difference between the use of data for research purposes within the context of a formal testing of an hypothesis and a natural exploration of a large dataset that has the potential to affect individuals and society as a whole.

Models do exist for the use of large anonymised datasets that are accrued for specific purposes, for example the Honest Brokers Service[1] (HBS) in Northern Ireland, Administrative Data Research Centre[2] (ADRC) in the UK, Irish Social Science Data Archive[3] (ISSDA) and Eurostat[4]. These are assured, anonymised resources that potentially allow for data analysis and linkage at high policy level, but one needs to be aware that the utility is limited as one works with anonymised sources.

HBS[5] work on collated data sets, from Health Trusts that are anonymised for secondary use by researchers from a variety of institutions. The data remains securely stored within the HBS, with data never being allowed to leave the control of the service and each project review for potential risk for re-identification and statistical integrity, particularly when linking other data sets. Each application is then reviewed by a team drawn from the HBS governance board, which is then reported to the whole board.

[1] http://www.hscbusiness.hscni.net/services/2454.htm
[2] https://www.adrn.ac.uk/
[3] https://www.ucd.ie/issda/
[4] http://ec.europa.eu/eurostat
[5] www.hscbusiness.hscni.net/services/2454.htm

This model works, but requires both resource and expertise, but assures, confidentiality, integrity and quality for anonymised data that is accrued within the health sector in Northern Ireland.

The Finnish National Institute for Health and Welfare (THL) produces a range of statistics in the field of social welfare and health care to support decision-making, development and research. As a statistical authority THL is responsible for the maintenance and development of statistical and register resources. The authorisation is necessary from the THL for any use of confidential data for research purposes[6]. In addition, the Finnish Innovation Fund Sitra and THL are preparing a new, one-stop-shop operator that will collect and coordinate well-being data on the Finnish population for use in areas such as research, the operator is called: Isaacus[7] Several pre-production projects have been launched, which are building parts of the future digital health hub and the permit services. The experiences from these projects will be collected and integrated into a plan of action for the operator to enable the launch of its operation in 2018.

Within the big data life cycle of data generation, storage and processing, a number of mechanisms exist for ensuring privacy. Access restriction and falsifying data can be used during data generation, encryption and distribution in storage and safeguarding unsolicited disclosure and extraction without violating privacy (Jain et al, 2016).

Storage and access are thus important in the control of these datasets, and HBS reflects the suggestions of Jain (2016), storage is secure and confined, in that no data can leave the service, it is encrypted and all disclosure is strictly controlled. This approach is mirrored to some extent, by the technology in that MIDAS assures that data remains within the host organisations, who thus control access, encryption etc.

As we move more and more to cloud based solutions, the problems will change as will the challenges, these issues are described by Sun et al (2014) as:

"(i) how to enable users to have control over their data when the data are stored and processed in cloud and avoid theft, nefarious use, and unauthorized resale,
(ii) how to guarantee data replications in a jurisdiction and consistent state, where replicating user data to multiple suitable locations is an usual choice, and avoid data loss, leakage, and unauthorized modification or fabrication,

[6] https://www.thl.fi/en/web/thlfi-en/statistics/information-for-researchers
[7] https://www.sitra.fi/en/projects/isaacus-pre-production-projects

(iii) which party is responsible for ensuring legal requirements for personal information,
(iv) to what extent cloud subcontractors are involved in processing which can be properly identified, checked, and ascertained."

MIDAS side steps some of these issues by limiting the requirement for moving data from the host.

Consent (sic: Informed consent) is a dynamic, shared experience, which must be tailored to each individual's need, therefore the notional acceptance of terms and conditions, informed within an EULA, occasionally updated, but by its very nature opaque, creates issues for those wishing to access data, regardless of purpose. One must always keep in mind, that the value of any piece of data collected may only become apparent, over time, when analysis or indeed technologies and techniques enable the value to become evident (Ioannidis, 2013).

Is it enough to structure engagement as a one off, without further ***meaningful*** *contact?*

In essence, no consent is required for the processing of information, when it is being used for legitimate purposes, and there is a clear appreciation by parties submitting their data within the confines of that framework of legitimacy. But, the value and potential in the data may fall outside this initial constraint, and must therefore potentially be sought when this context changes, or indeed the potential use clearly articulated and consent sought up front with the individual.

For example, if a person's data is used to assess weight when entering a car automatically, which is then used this to assess vehicle range, and this data is then used to develop systems for creating insurance models, a legitimate output for that company's business, should this require consent? This is addressed in section 3.3, which explores consent as a tool to ensure good practice.

If data and its value are dynamic, then the process for consent must acknowledge this. Models of consent are clearly established within health related research which will be explored later in the report.

Of course, within the context of systems and processes that by definition allow secondary use of data (anonymised) for a variety of purposes, is this discussion no

more than self-reflection? To ensure this potential for secondary use, then the minimum requirement, by legal definition, is for the data to actually be anonymised.

3.2 Anonymisation

Obviously, there are a variety of techniques for anonymisation (see table 3), and these must assure that the data being used meets a minimum level of quality in specific regard to this anonymisation. This quality assurance is essential within the context of the GPDR, for both the original controller of the data and those that may have access to the data through the controller, with the added caveat that the controller must be assured and has some responsibility for assuring the ability of those accessing the data to meet this quality bar. This is a departure from the current legislative provision, and has potential impact for those sharing data.

Table 3: Anonymisation techniques

Technique		
Data Reduction	**Data Perturbation**	**Non-perturbation**
Removing variables	Micro-aggregation	Sampling
Removing records	Data swapping	Cross-tabulation of data
Global recording	Post-Randomisation Method	
Local Suppression	Adding Noise	
	Resampling	

The benefits of utilising this data is thus recognised within the GDPR, in that anonymised data being used for a legitimate purpose is permitted within the EU context, yet no standard is available to ensure anonymity and ensure minimal risk of re-identification (El Emam, Rodgers, & Malin, 2015).

Therefore, assuming that anonymisation and risk of re-identification have been reduced to a meaningfully low risk, how does this impact utility?

There is a general acceptance that with increasing levels of anonymity, the utility within datasets will decrease (Wu, 2013), although that is not to say that there is potentially huge value in discrete anonymised datasets. Rather the ability to add value, through integration of other anonymised datasets becomes more difficult as one drives to lower levels of granularity.

Grant Agreement No: 727721

Health data, with core identifiers, coherent through the record, will enable patient specific outcome measures to be readily identified and tracked over time, with potential benefit at a personal level, but only through the use of identifiable or pseudonymised data sets.

With anonymised data, the value lies at a population level, ideal for MIDAS, which is based at this level. Certain caveats thus become self-evident, for example, niche disease groups, which by their very nature have small numbers of patients associated with them, and thus pose a heightened risk of re-identification.

What is evident is that the requirement for standards that ensure anonymity, both in construction of the data and any analysis, are essential, but are not the only requirements for good practice if one excludes notions of consent. This creates effort, both in design and usage of any system, which will require resource and expertise as an integral part of any use, which obviously adds a burden of cost.

Synthetic data has the potential to assist in overcoming data privacy concerns. Synthetic data are “microdata records created to improve data utility while preventing disclosure of confidential respondent information. Synthetic data is created by statistically modeling original data and then using those models to generate new data values that reproduce the original data's statistical properties. Users are unable to identify the information of the entities that provided the original data” (US Census Bureau, 2017). Synthetic data must statistically resemble the original data from which it is modelled, as well as formally and structurally resembling the original data (Patki et al, 2016). The utility of this type of data may improve protections through all phases of the data cycle, particularly when considering the sensitivities and access to health related records (Choi, 2017).

3.3 Consent as a tool

How does a model that ensures ethical insight, consent and process, as previously referred to, fit with access, management and utility?

As previously stated, for anonymised datasets there is no requirement for consent, yet there may be value in leveraging the power of consent, in addressing some of the issues identified, particularly in regard to utility and adding significance to the potential of any number of datasets. This informed consent would ensure that the individual had access to information about a specific use case (or a more general

use case), ensuring that appropriate permissions were in place in before collection ensued, protecting the individual's rights, ensuring transparency and maximising potential.

On the face of it, this seems like a simple, and straightforward approach, and certainly within the sphere of health related research is a requirement for ethical review (and addressing issues of consent within this domain), but once again only for identifiable / pseudonymised data (World Medical Association, 2013) within the context of a scientific experiment. Routinely data is used within healthcare settings for service evaluation, service development and audit, without any process of formal consent, rather consent is inferred and ethical considerations, which include issues of consent, are assumed to apply only to formal research (Dixon, 2017).

Good practice, within the paradigm of non-research, would contain the four principles for ethical review: Autonomy, Beneficence, Non-Maleficence and Justice (Macklin, 2003). At a minimum ethical oversight is required to examine the impact the project may or may not have on the individual (Dixon, 2017). This is as true for a local ward based project as for a project requiring data from a multitude of sources at a substantive population scale.

Therefore when we consider Big Data, the paradigm is not significantly different, with the impact at an individual level requiring attention when moving a project forward, as well as at larger scale. Other considerations impact on this assessment, for example, the systems for aggregation, quality control, analytical tools and use/ outcome, requiring review. For example, consent and mechanisms for granting and continued validation of consent should be reflected on “in the contexts of direct marketing behavioural advertising, third-party data brokering, or location-based services” (Tene & Polonetsky, 2012). Yet, there needs to be a certain pragmatism, when leveraging the potential in this data, as the significance of data can sometimes only become evident over time, for example, Kaiser Permanente’s analysis of Vioxx (United States Senate: Committee on Finance, 2004).

3.4 How not to do it

Vioxx was manufactured by Merck and after carrying out 8 randomised studies, with a group of 5400 subjects, gained Federal Drug Administration (FDA) approval in 1999. Also in 1999 the company launched the Vioxx Gastrointestinal Outcomes Research study (VIGOR), with a target recruitment of 8000 patients. The aim of the

study was to compare the toxicity of rofecoxib (Vioxx) with Naproxen (both oral painkillers) for gastrointestinal toxicity.

In October 1999, there was a clear indication of Vioxx's superiority in respect of gastrointestinal toxicity, after review of the Data and Safety monitoring boards review (DSMB). The next meeting of the DMSB, in November 1999 examined episodes of cardiac toxicity, which showed that 79 patients out of 4000 on the Vioxx arm, as opposed to 41 out of 4000 on the naproxen arm had serious cardiac events including death. The panel decided that whilst "disconcerting" the event numbers were small. At the meeting of the DMSB in December 1999, the relative risk of Vioxx v Naproxen was shown at this stage to be 2:1 but the committee decide to continue the study, with a recommendation to analyse the cardiovascular results before study end, they postulated that Naproxen may have had a protective effect, thus skewing the outcome in this arm. This was communicated to Merck. In January 2000, Merck wanted to avoid a specific analysis at this time, rather they wanted to wait to study end and combine an overall report on cardiovascular events from all studies to date. Dr Michael Weinblatt, the DMSB chair wished for an immediate analysis. Agreement was reached in February 2000, that this would be completed by 10 February 2000, and would be limited to events to date. There followed a short period of contracting and financial disclosure that showed that Weinblatt had developed a commercial interest with Merck.

In May 2000, the study was submitted to the New England Journal of Medicine (NEJM), but with data missing. 3 out of 20 heart attacks within the Vioxx arm were excluded. This was highlighted in July 2000 by the statistician, Deborah Shapiro. Two sets of corrections were then submitted to the NEJM in July and November 2000, with this data still missing. The VIGOR results were published in the NEJM, 23 November 2000, with trial data submitted to the FDA on 13 October 2000. In February 2001, the FDA had an advisory meeting on Vioxx and released the VIGOR data on its Web site. On 22 August, 2001, cardiologists, Debrabrata Mukherjee, Steven Nissen and Eric Topol published their meta-analysis of the VIGOR dataset in the Journal of the American Medical Association (JAMA), they cast serious doubt on the notion of a Naproxen effect.

Numerous studies between January 2002 and August 2004 point to increased cardiovascular risk with Vioxx. Merck withdrew the drug in September 2004, after the APPROVe study (Bresalier, et al., 2005) showed an increased risk of heart attack after 18 months.. Between July 2005 and November 2007, there are court proceedings and formal requests by the NEJM for corrective action. Merck settle the

Grant Agreement No: 727721

case for $4.85 billion November 2007. In all, the Lancet estimates that 88,000 Americans have had heart attacks as a result of taking Vioxx, with 38,000 dying (Prakash & Valentine, 2007). This highlights the effect time has on effect, but also exemplifies the importance of timely and accurate analysis, and how individuals can be affected by the methods used for aggregation, review and examination of data. It is not the only example, and the complexity inherent in drawing conclusions from big data should must also be recognised as with the criticism of the Google Flu trend data, once considered the poster child of Big Data analytics (Lohr, 2014).

This success and failure impacts on awareness which, when wedded to public perception and understanding of both the capability and potential around Big Data analytics creates prospective issues for those wishing to maximise the benefit from data.

3.5 A brief note on opt-in/opt-out

In a broader sense, there can be two approaches: *opt in* and *opt out*. Opt in, is one in which the individual makes an informed choice and selects to become involved. Opt out, is one in which the company (controller), defines use, and use the data as the default unless contacted by the individual. Opt out is seen as preferential for innovation and productivity (McQuinn, 2017). There are obvious parallels between consent for organ donation and data “donation”. This will be explored more fully in D2.2, as a model is defined and tested, in light of end user engagement.

3.6 MyData

Within MIDAS, the Finnish model of MyData describes a system for matching the requirements of legislation and good practice that potentially offers a pragmatic solution to the issues around consent.

There are three core principles embedded with the MyData paradigm.

1. Human Centric control, with the technology, systems and processes for individuals to manage their own data, thus dealing directly with issues of consent.
2. Usable data, taken from a variety of sources and structured in such a way as to aid linkage and analysis
3. An open platform that allows access by industry to ensure compliance with quality standards and legislation whilst increasing utility.

Grant Agreement No: 727721

This model will be assessed within the MIDAS project for usability and against the benefits referred to above in policy creation. The deliverable D2.8 (M08) MyData Potential Report gives a MyData overview in healthcare and in the second phase the document D2.9 will give more details and information on the MyData potential in policy making as a part of MIDAS.

MyData architecture offers one of the possible legal bases for processing personal data. All data processing with MyData is based on consents (Alén-Savikko et al, 2016)[8]. It is characteristic to consents that they are always issued by the account owner (data subject), who can also change or withdraw the consent at will. The data authorization and consent model has been described in details within MyData Authorisation model[9].

3.7 Basque Side model

To handle ethical issues related to the Basque pilot site focused on child obesity, BIOEF, as coordinator of the Basque side, has agreed a methodology with the Basque Agency for data protection, the Basque Health system legal department and the Basque Clinical Ethics committee. This methodology is summarised in figure 3.2.

[8] https://hiit.github.io/mydata-stack
[9] https://github.com/HIIT/mydata-stack/raw/gh-pages/mydata-data-authz.pdf

Grant Agreement No: 727721

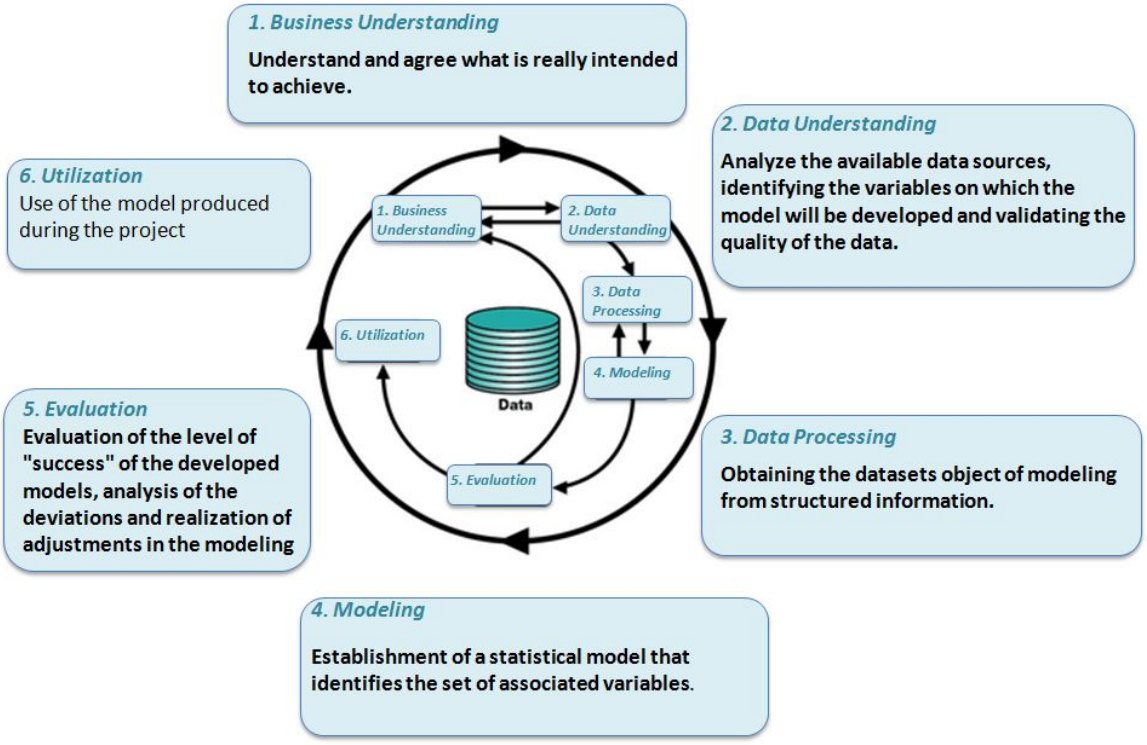


Figure 3.2 Basque Policy Site Ethics Methodology

At this stage of the project, it has focused on the first 3 steps:

1. **Business understanding**: Several meetings have been held at Basque Health system's central services in order to analysis the selected issue, Child Obesity, as well as the associated variables. A working group has been created comprised of paediatricians, IT technicians, policy makers, healthcare director, Vicomtech and coordinators from BIOEF. On the other hand, a collaboration framework has been agreed among BIOEF and EROSKI, and BIOEF and Euskaltel, as a data providers.

2. **Data understanding:** The Basque side working work has analysed the available data sources, following the short description of the available data sources:
    - Clinical data of more than 850.000 patients of the Basque Health system. More than 120 variables have been selected by involved paediatricians.
    - Open Data Euskadi[10]

[10] http://opendata.euskadi.eus/inicio/

Grant Agreement No: 727721

- Basque Grocery (Eroski) data focused in a large questionnaire related to the consuming habits among the young people (400 people) as well as access to the data obtained in Ekilibria[11] program (Clients' caloric consuming descriptions)
- Geolocation of the Basque people (Euskaltel).

3. **Data processing**: The first export of 500 patients have been done taking into account the requirements agreed within the working group. This will aid to understand the technical characteristics of the Basque Health system.

Related to the legal aspects, the continuous communication with the Basque Agency for data protection as well as the elaboration of the Data protection Impact assessment by the external legal agency will ensure the correct use and treatment of the sensible data.

The BIOEF legal department is working on the needed DAAs with data providers (Basque Health System, Eroski and Euskaltel) and MIDAS developers.

4 Current Models of Communication

The single biggest problem in communication is the illusion that it has taken place (George Bernard Shaw)

Thus far we have discussed end user involvement in an abstract sense, in that, models of engagement are driven with the legislative bounds, ethical frameworks and good practice controls, with little appreciable, meaningful input, except as a the provider of consent to a particular project, if required. This is important, in that it mirrors to some extent the historical attitudes and practical application of data use within the medical research domain. Traditionally, medical research has been seen as the domain of academics and/or business, with Big Pharma maintaining a preeminent position of power. Certainly, this world view, supported through paternalistic systems of delivery and control in respect of healthcare, was well established (Coulter, 1999). The unassailability of this approach, one in which the imbalance by those accessing a service (the patient) to those delivering the service

[11] https://www.eroski.es/ekilibria-club-salud

(healthcare providers), was seen to be, all things being equal, appropriate. This perception was from the end of the 1980's onwards under challenge.

This challenge was driven, primarily by two separate, yet in some ways related events, the emergence of the Human Immunodeficiency Virus (HIV) and the internet. As HIV ravaged the communities of San Francisco and New York, the medical hagiography was upended by militant groups of patients, who fought to have their voices and expertise considered, not only in driving treatment choices, but also in defining the research agenda for HIV (France, 2016). Wedded to this is the readily available source online, that offered/offer anyone the opportunity to research and understand information relevant to their personal needs, thus demystifying the sometimes arcane and protected knowledge so beloved of professionals (Laing, Newholm, Keeling, & Speier, 2010).

4.1 An example of effect

There are similarities with the emergence of Big Data as a resource in the modern world in regard to health with the emergence of some of the great discoveries in healthcare over the last 100 years. Big Data has as much potential to change the delivery of medicine as the discovery of penicillin.

It is remarkable that Fleming, whilst discovering penicillin and describing its antibiotic properties, had no faith in its potential as a pragmatic solution for the treatment of infection. Rather, the work of Florey and Chain, building on that of Fleming (Jacobs, 2004) ensured a firm grounding for the development of deep-tank fermentation by Pfizer, which guaranteed the production and widespread availability of high quality medical grade penicillin (American Chemical Society, 2008). It is curious that the development of Data Science in specific regard to health care is now on the cusp of enabling true personalised medicine by, in very simple terms, linking the individual genome of a person, with pathology and treatment (McCarty, et al., 2011). Here once again we have potential that can only ever be exploited through the development and use of technology specific to data analytics.

Yet, the difference in how we drive and develop utility may lie in how we communicate the benefit as well as the risk. The use of penicillin, along with sulphonamides, heralded a new era in medicine, which has dramatically changed outcomes for patients (Aminov, 2010) with public understanding and appreciation of these benefits ensuring widespread endorsement and use. This has continued

relatively uninterrupted for the last 60 years, with a step progress in the range and efficacy of antibiotics, with a commensurate expectation by the public, of the sustained benefit and efficacy of these drugs (McNulty, Boyle, Nichols, Clappison, & Davey, 2007). Unfortunately, this perception is leading to what some have defined as a medical Armageddon (The Independent, 2017), with inappropriate prescribing, non-completion of antibiotic regimens and incorrect use by patients, among the reasons identified as contributors to the emerging crisis (The Lancet Infectious Diseases Commission, 2013).

As a roadmap for the development of Data Science as a change agent for healthcare, there is little doubt that the history of antibiotics has lessons for those involved in driving the Data Analytics agenda. The availability of data, particularly in light of mobile technologies, that continue to grow and improve, as well as the ability to access supercomputing infrastructure through cloud based solutions, is driving the integration of technological solutions with the clinical need, particularly in respect of personalised medicine (Broner, 2017). Whilst some professional, technologists and healthcare providers, are starting to understand the import of this, it is far from universally accepted or indeed understood, not only by these professionals (Mullich, 2013), but more importantly the public (Andrejevic, 2014).

The story of antibiotic use, and its success (although now being questioned), is one where the success and rapid change in individual and population outcomes is easy to communicate. It has the benefit of being relatively immediate, easily accessible and has a model that is familiar and manageable: one sees a doctor, is examined, is prescribed and administers, with an effect that is noticeable (if successful) in the short term. For the public at large, there is a clear and easy familiarity with the technology (the pill) and the model of use. The use of data for health care benefit has some way to go before this understanding, however inappropriately followed, with antibiotics is reached. Therefore there is a need to raise awareness urgently to show value and educate the public that will allow meaningful collaboration, between those providing the data, and the teams maximising any potential from this data, at individual and / or population based levels (Habl, Renner, Bobek, & Laschkolnig, 2016).

It is essential that this communication speaks to people at a personal level, and conveys a message that articulates the potential benefits in the use of Big Data, addressing the issues we have already discussed such as privacy, ethics and consent. If this is not a priority, then we will have both misunderstanding as well as misrepresentation, which we will now address within perceptions.

5 Perceptions

"If the doors of perception were cleared everything will appear to us as it is, infinite" (Oscar Wilde)

There is a general acceptance in the literature that there is ongoing generation of a large and complex datasets, derived from population monitoring (e.g. clinical records, demographics and survey data) as well as other indirect determinants of health, such as environmental and behavioural data (AbouZahr, Adjei, & Kanchanachitra, 2007). This, as has been stated previously, raises the potential for re-identification of an individual, when combining datasets for analysis, jigsaw identification (Wellcome Trust, 2016). This is but one example of the dilemmas, which exist, when the use of data is considered, reflecting the scope, scale and complexity of the data (Ekbia, et al., 2015).

For those familiar with the technology, its potential and the complexities that exist, the risk and rewards are easily recognised. Yet, for those in positions less well informed, there is a danger that the risk / reward relationship will be less than appropriately defined.

The perception of the general public, is informed through a variety of sources, but in the 21st century it is primarily through media, be that online, satellite, terrestrial or digital (Gitlin, 2003). This perception and the tools that inform are not impassive or indeed unbiased, and yet, to achieve an informed picture one must become aware of the facts, facts grounded in reality, rather through misreporting, or "Fake News".

We live in an age where the validity of the data we regularly review, can be manipulated, not only statistically, but through a specific perspective lens. Thus, the debacle over Care.data was framed within an environment that was toxic to the innovation, due in no small part to the poor management, miscommunication, professional disagreements and inadequate provision for data protection within the project design (Presser, Hruskova, Rowbottom, & Kancir, 2015). That being said, the idea of a coherent primary and secondary healthcare data repository (NHS England, 2013) was far from senseless, but the lack of thought foresight, honesty and communication ensured that it was viewed in the most negative of lights, with the Review of Data Security, Consent and Opt-Outs (Caldicott, 2016) nailing the coffin firmly shut on the project.

A core theme running through this report is the need for dynamic, appropriate and recurrent communication. It also asserts that there is a requirement to actually use the data we have within health (we should note that within MIDAS, one of the core concepts is that this should be used in conjunction with other sources beyond health to maximise policy effect) for benefit, meeting the high levels of trust expected by the public, and having the systems and technologies in place to manage this most sensitive of data (Caldicott, 2016).

Unfortunately, lessons are not always learned, and one could argue that the setbacks, as well as tone, set by Care.Data was further compounded by the recent DeepMind project, created between Google and a NHS Trust.

Once again the core message, one around the development of a tool to help manage Acute Kidney Injury (AKI) was lost in the scramble to sensationalise the issues. Issues, there certainly were, as elaborated upon by Elizabeth Denham in her role as Information Commissioner (Dehanm, 2017) and further supported by the DeepMind Health Independent Review Panel Annual Report (Bracken et al., 2017). What is notable is once again the identification of communication, and thus perception and lack of understanding is highlighted throughout. It should also be stated that DeepMind health had initiated the formation of the Independent Review Body as a core concept within the overall project architecture, a fact that was somewhat under reported to the public.

When it comes to sensitive information such as health and banking, the public demand the highest standards, they are aware of the significance of the data and the importance of it. Yet, data rich sources, such as mobile phone data, supermarket loyalty schemes and app data appear to rate much lower in respect of appropriate protection and use. Why is this?

It may relate to the level of perceived risk and the utility of the data that the user provides for a return? With app data the use case drives utility; google maps user data usage concerns anyone?

With a coherent healthcare record, the benefit is both complicated and much less immediate and tangible, thus leading to reticence and mistrust when innovating within these domains. This sensitive data is also readily identifiable, as opposed to data being collected on a store card or credit card, it’s remote and unseen.

There is obviously work to do in relation to driving engagement between the innovator and the data provider. This is a requirement for good practice, ethical viability and ultimately utility.

5.1 Ongoing Public Engagement Work

A portfolio of work is currently underway within WP2 to examine public perceptions related to consent, data sharing, and anonymisation. One aspect of this work is a research study, led by PHE designed to assess how acceptable a sample of the British public finds the sharing of different types of health and security related data. Using an online survey template, participants will be presented with different scenarios during which there is the potential for data to be shared. These scenarios cover a range of different security and health related contexts (e.g., terrorist attacks, chemical fires, mental health, and cancer) in which anonymised or non-anonymised data (e.g., concerning the incident, disease, or treatment) might be shared. For each scenario, participants will be asked to provide ratings regarding their perceptions of the acceptability of sharing anonymised or identifiable data with a range of different organisations. This study will provide us with a first look at the contexts and situations in which members of the public may find it more or less acceptable to share different kinds of data. Ethical approval has been granted for this public perception study within WP2 with data collection expected to commence in November 2017 for projected completion in February 2018.

This study is only one aspect of WP2's public engagement work. For instance, a further collaboration between IBM, SET, and PHE is also underway to explore the potential for using a currently in-development Twitter bot to conduct primary research, based on the study described above, among Twitter users. This presents an opportunity to both enable further public engagement data collection and also to trial an exciting new data collection method within the MIDAS consortium.

6 The Beginnings of a Model

Within MIDAS we have the operational good practice arrangements for managing and delivering on the project itself, as well as an overall project deliverable. A deliverable that will create a practical and pragmatic model of good practice that is transferable into the everyday, in tandem with the technology.

Grant Agreement No: 727721

6.1 Policy Board Baseline Questionnaire

There was a requirement first and foremost to understand the policy leads' current thinking and structures in respect of ethics and governance, and to this end a baseline assessment was carried out using a simple questionnaire (appendix 3). There was general concurrence in regards to use of data for research and core business, use of sensitive and publically available data, as well as the use of appropriate codes of conduct. What was less clear was the use of opt out/opt in models and the use of public consultation to understand perception. This latter consideration may support some of the thinking described within the preceding sections of this report, in that the technology and its potential has moved on, yet public perception does not appear to be assessed in any real dynamic way, thus those driving use have little real understanding of this perception, or indeed the information necessary to discover and address any perceived or real public issues.

Partners exist in three primary domains, academia, industry and public sector, and through discussion, some differences in respect of understanding and constraints have been identified. Again these mirror some of the preceding discussion in the report, as to the push and pull of utility versus good practice versus innovation and outcomes.

6.2 The Ethics and Privacy Advisory Group (EPAG)

To ensure control in respect of datasets and their use within MIDAS to help manage this differential, a structure and model was created to assure control of data access and activity within the consortia, through the Ethics and Privacy Advisory Group (EPAG). EPAG thus became the focus for governance in respect of ethics and privacy, governance being defined as "the action, manner, or power of governing" (Houghton Mifflin Harcourt, 2016), manifesting itself as structures that ensure ethical, legal and good practice domains are addressed within a system, and that there is clear accountability and responsibility aligned within the systems that in the end answer to society, or in this case primarily to the consortia (see Appendix 1 and 2).

The consortia has access to datasets for testing that are anonymised, and to ensure good practice the terms of reference for EPAG were constructed using the using the 7 Caldicott principles[12]:

[12] https://www.igt.hscic.gov.uk/Caldicott2Principles.aspx

Grant Agreement No: 727721

Principle 1

Justify the purpose(s)

Every proposed use or transfer of patient-identifiable information within or from an organisation should be clearly defined and scrutinised, with continuing uses regularly reviewed by an appropriate guardian.

Principle 2

Don't use patient-identifiable information unless it is absolutely necessary

Patient-identifiable data items should not be used unless there is no alternative.

Principle 3

Use the minimum necessary patient-identifiable information

Where use of patient-identifiable information is considered to be essential, each individual item of information should be justified with the aim of reducing identifiably.

Principle 4

Access to patient-identifiable information should be on a strict need to know basis

Only those individuals who need access to patient-identifiable information should have access to it, and they should only have access to the information items that they need to see.

Principle 5

Everyone should be aware of their responsibilities

Action should be taken to ensure that those handling patient-identifiable information, (both clinical and non-clinical staff) are made fully aware of their responsibilities and obligations to respect patient confidentiality.

Principle 6

Understand and comply with the law

Every use of patient-identifiable information must be lawful. Someone in each organisation should be responsible for ensuring that the organisation complies with legal requirements.

Principle 7

The duty to share information can be as important as the duty to protect patient confidentiality

Health and social care professionals should have the confidence to share information in the best interests of their patients within the framework set out by

these principles. They should be supported by the policies of their employers, regulators and professional bodies.
(Caldicott, 2013)

EPAG acts in the broadest sense, as the final ethical and good practice arbiter for the MIDAS project (Appendix 1 and Appendix 2). This process and reference group, with external membership from outside the project team, ensure an oversight function, and offers opinion, to robustly challenge potential assumptions within the team in specific relation to ethics and privacy. This is a requirement within the grant agreement, although the preceding section clearly articulates the need for ethics and governance within the legislative environment that this project is clearly established within. MIDAS is at its core a project about the appropriate valued use of data that can be structured and analysed to inform the policy cycle. For this to work it must operate within the current legislative frameworks. To facilitate this work, the consortia has created a system of governance and review that ensures external validation and quality assurance of both the data and use under scrutiny, but also the intent for which it is being used.

6.3 Risk Assessment Tool

To help assess risk, a risk tool was created to give broad assessment of potential issues with data for partners to simplify and expedite process (Appendix 4). This process uses a portal for submission and review with communication and elaboration through established consortia channels. The model therefore resembles:

Stage 1
- Identification of dataset for use by partner
- Definition of project constraints (technical at the moment, not policy)

Stage 2
- Minimal dataset sent for review: By Policy Board
- EPAG to assess potential ethical issues
- Operational team actions review outcomes
- EPAG approval

Stage 3
- Project carried out
- Review by Policy Board

Thus far, MIDAS is at the stage of testing technology and constructs. In the next phase this will develop into an integrated model of policy definition and evaluation aligned to the EPAG process, that will act as a template for all prospective groups/teams using MIDAS.

6.4 Data Access Agreements

Data access and control is strictly controlled at this stage, with the use of Data Access Agreements (DAA) (appendix 5) mandatory for all parties accessing test data-sets to deliver the overall project.

6.5 Summary

Primarily the model of self-regulation created within MIDAS aligned the overall project requirements and operational needs of the project, within a framework that assesses both the validity of any proposed outcome and the practicalities of the assessment process itself. This will ensure meeting these project's needs, within a clear ethical and privacy framework, ensuring alignment with core principles inherent within the legislation.

7 Conclusion

Data use is complex, difficult and necessary. Legislation and good practice frameworks create environments that help assure appropriate use, but the interpretation, rapid advance of technology and the sometime lack of clarity in regard to utility lead to difficulties for both those who provide the data and those wishing to use it. Within all frameworks, consent is recognised as a potentially useful tool, yet, as has been seen, consent creates issues in respect of utility, in that for meaningful consent, processes must reflect the needs of the end user in respect of ongoing meaningful communication. This communication forms a core component for engagement, whether formal consent is to take place or not. When it comes to the idea of big data use, there is a clear need to challenge misunderstanding, and misrepresentation that appear to be creating a negative public perception, particularly when the data being used is derived from public services.

Grant Agreement No: 727721

There is now an understanding that data when used appropriately, with the necessary protections has the potential to bring real and meaningful insight into complex problems and issues, at both the individual and population levels. This review endorses this potential, but appreciates the challenges that might influence and make difficult the actualisation. Whilst legislation exists, and will be robustly controlled within the GDPR, the huge potential, variety and breadth of users creates difficulties when considering any single model of good practice, rather the good practice model would be better served by creating key recommendations, to ensure the creation of bespoke, pragmatic and effective models of performance.

7.1 WP2 Plan of Work

To ensure that the final project deliverables are achieved for Deliverable 2.2, WP2 are preparing a program of work exploring a variety of issues, for example, perception and consent. This includes, but is not limited to, ongoing work by Public Health England designed to examine the British public's response to sharing their data in a variety of health and security related contexts (this will form D2.4). Furthermore, work is ongoing to explore the potential to use Twitter as a tool for gaining meaningful insight into the creation, perception and acceptance of any model of use. The findings from these engagements will build upon this review, and will be supplemented by the use of a technological project within MIDAS using social media to drive feedback from the public.

What is obvious is that there are requirements for the project going forward:

- A process to understand the actual perceptions of a target population
- A communication plan to articulate the importance of data science/data analytics for the general public, policy and clinical staff
- A method to assess the need for consent for data controllers
- A process to integrate communication as part of ongoing information giving to persons
- A broad ethical and good practice model of use that is both dynamic and flexible

The next phase of Work Package 2 will address these requirements as follows:

- Creation of a process of engagement to define needs to articulate benefit and challenge misunderstanding - this will build on the work of PHE and the Twitter-bot project led by IBM, and will include focus groups for public, policy and clinical staff.

Grant Agreement No: 727721

- A deeper understanding of the use of consent - Is it necessary, not specifically in regards to GDPR, but also in regards to acceptance to society? Again the twitter bot project and work of PHE will help drive this work
- The creation of a model that builds on the internal mechanisms created within MIDAS, using the following template, derived from current understanding:

A WP2-specific proposed model for acceptability and review by policy, industry and public sector organisations is as follows (and is aligned with similar activities being carried out throughout the entire MIDAS project and all work packages):

Stage 1- Policy data identification

- Identification of policy need by policy team/ lead: applicant provides a minimal data description that describes policy need - These policy needs have been identified within MIDAS.
- Minimal data description sent for review by technology and operational partners
- Technical feasibility agreed - Data set and systems identification
- EPAG type group to assess potential ethical issues
- Operational team actions review outcomes
- Regional Policy pilot case approval

Stage 2 - Project delivery stage

- Project team defined
- Project timelines agreed
- Evaluation criteria set

Stage 3 - Project access

- Data Access – permissions as per nation/organisation
- Technical access
- Analytics
- Policy report

Stage 4 - Impact

- Policy Board review
- Policy formulation
- Policy plan: Implementation and Evaluation
- This will align with the work of WP6

Grant Agreement No: 727721

A Gantt chart describing the WP2 plan is shown in Figure 7.1.

Task Name	Start	End	Duration (days)
Process of Engagement - Focus groups/ engagement days	17/11/2017	28/03/2018	131
Understanding perception of model of consent	17/11/2017	28/03/2018	131
Creation of model for testing	28/03/2018	30/05/2018	63
Model testing	30/05/2018	30/08/2018	92
D2.2 Report	30/08/2018	30/10/2018	61

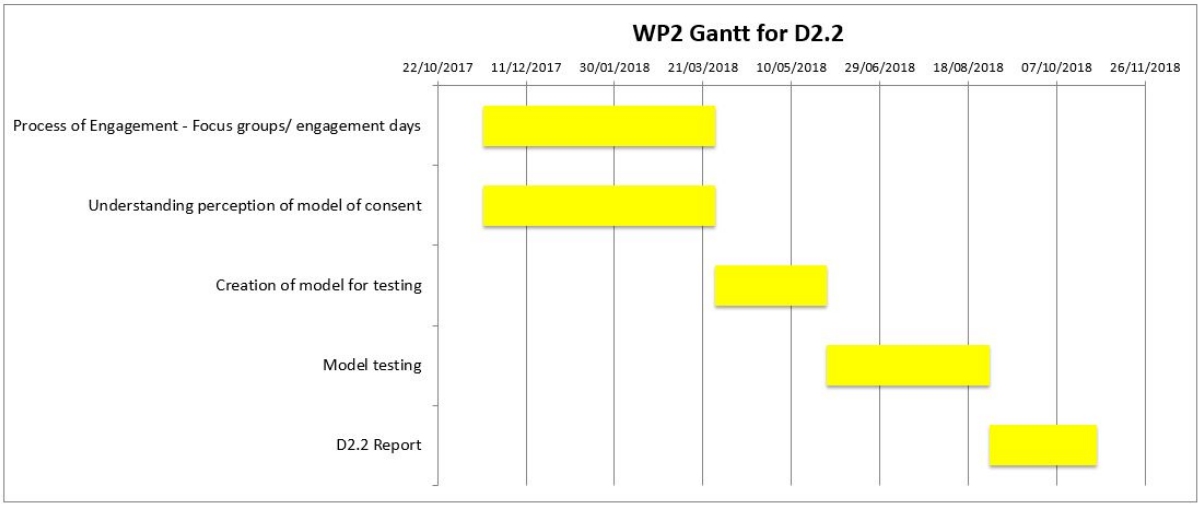


Figure 7.1 WP2 Plan for D2.2: Good Practice Report 2 (due M24)

8 References

AbouZahr, C., Adjei, S., & Kanchanachitra, C. (2007). From data to policy: good practices and cautionary tales. The Lancet, 369, 1039-1046.

Akter, S., & Wamba, S. F. (2016). Big data analytics in Ecommerce: a systematic review and agenda for future research. Electron Markets, 26, 173-194.

Alén-Savikko, A, Byström, N, Hirvonsalo, H, Honko, H, Kallonen, A, Kortesniemi, Y, Kuikkaniemi, K, Paaso, T, Pitkänen, OP, Poikola, A, Tuoriniemi, S, Vainikainen, S & Yli-Kantola, J 2016, MyData Architecture: Consent Based Approach for Personal Data Management. Helsinki Institute for Information Technology, Finland.

American Chemical Society. (2008, June 12). Development of Deep-tank Fermentation. Washington, District of Columbia, United States.

Aminov, R. I. (2010). A brief of the antibiotic era: lessons learned and challenges for the future. frontiers in MICROBIOLOGY, 1, 1-7.

Andrejevic. (2014). The Big Data Divide. International Journal of Communication, 1673-1689.

Asche, C. J., Seal, B., Kahler, K., Oehrlein, E. M., & Baumgartner, M. G. (2017). Evaluation of Healthcare Interventions and Big Data: Review of Associated Data Issues. PharmacoEconomics, 35(8), 765-759.

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., et al. (2013). The Six Primary Dimensions For Data Quality Assessment.

Bollier, D. (2010). The Promise and Peril of Big Data. Queenstown: The Aspen Institute.

Bracken, M., Bromiley, M., Buggins, E., Burbidge, E., Horton, R., Huppert, J., et al. (2017). DeepMind Health Independent Review Panel Annual Report. Soapbox.

Bresalier, R. S., Sandler, R. S., Quan, H., Bolognese, J. A., Oxenius, B., Horgan, K., et al. (2005). Cardiovascular Events Associated with Rofecoxib in a Colorectal Adenoma Chemoprevention Trial. The New England Journal of Medicine, 352(11), 1092-1102.

Broner, G. (2017). Supercomputing Is the Future of Genomics Research. Genetic Engineering & Biotechnology News, 37(3), 18-19.

Burtch, L. (2017). The Burtch Works Study: Salaries of Predictive Analytics Professionals. Evanston: Burtch Works Executive Recruiting.

Caldicott, D. F. (2013). *Caldicott review: information governance in the health and care system.* Department of Health.

Caldicott, F:. (2016). Review of Data Security, Consent and Opt-Outs. William Lea.

Calman, K. C. (2004). Evolutionary ethics: can values change. Journal of Medical Ethics, 30, 366-370.

Cambridge English Dictionary. (2017, September 7). Best Practice. Retrieved from Cambridge English Dictionary: http://dictionary.cambridge.org/dictionary/english/best-practice.

Chen, H., Chiang, R. H., & Storey, V. C. (2012, December). Business Intelligence and Analytics From Big Data to Big Impact. MIS Quarterly, 36(4), pp. 1165-1188.

Cheng, K. (1969). Privacy and Data Bank. Chitty's Law Journal, 90-94.

Choi, E., Biswal, S.,Bradley, M.,Duke, J,.Stewart, W, F.,Jimeng, S. (2017). Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. Machine Learning in Healthcare. pp 1-20

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics in Healthcare. Health Affairs, 33(7), 1139-1147.

Coulter, A. (1999). Paternalism or partnership? Patient have grown up - and there's no going back. British Medical Journal, 319, 719-720.

Council of Europe: European Court of Human Rights. (1950). European Convention on Human Rights. Strasbourg: Council of Europe.

Dehanm, E. (2017, July 3rd). Letter to Sir David Sloman. London.

Dixon, N. (2017). Guide to managing ethical issues in quality improvement or clinical audit projects. Healthcare Quality Improvement Partnership.

Doyle, A. C. (1892). The Memoirs of Sherlock Holmes.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., et al. (2015). Advances in Information Science, Big Data, Bigger Dilemmas: A Critical Review. Journal of teh Association for Information Science and Technology, 66(8), 1523-1545.

El Emam, K., Rodgers, S., & Malin, B. (2015). Anonymising and sharing individual patient data. British Medical Journal, 1-6.

European Union. (2012). Charter of Fundamental Rights of the European Union. 2012/C 326/02.

Fischer, B. A. (2006). A Summary of Important Documents in the Field of Research Ethics. Schizophrenia Bulletin, 32(1), 69-80.

France, D. (2016). How to Survive a Plague: The Story of How Activists and Scientists Tamed AIDS. Picador.

Gitlin, T. (2003). The Whole World Is Watching: Mass Media in the Making and Unmaking of the New Left. University of California Press.

Habl, C., Renner, A.-T., Bobek, J., & Laschkolnig, A. (2016). Study on Big Data in Public Health, Telemedicine and Healthcare. Luxembourg: European Commision.

Houghton Mifflin Harcourt. (2016). *American Heritage® Dictionary of the English Language* (5th ed.). American Heritage® Dictionary of the English Language, Fifth Edition. Copyright © 2016 by Houghton Mifflin Houghton Mifflin Harcourt Publishing Company.

Ioannidis, J. P. (2013, March). Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research. The American Journal of Bioethics, 13(4), 40-42.

Jacobs, F. (2004). Breakthrough: The True Story of Penicillin. iUniverse.

Jain, P., Gyanchandani, M., Khare, N. (2016). Big Data Privacy: a technological perspective and review. Journal of Big Data. 3 (25).

Jensen, C., Potts, C., & Jensen, C. (2005). Privacy practices of Internet users: Self-reports versus observed behavior. International Journal of Human-Computer Studies, 203-227..

Laing, A., Newholm, T., Keeling, D., & Speier, D. (2010). Patients, Professionals and the Internet: Renegotiating the Healthcare Encounter. National Institute for Health Research.

Leonard, T., & Rubin, P. (2015). Big Data, Privacy and the Familiar Solutions,. Journal of Law, Economics and Policy, 11, 1-32.

Lohr, S. (2014, March 28th). Google Flu Trends: The limits of Big Data. Retrieved from The New York Times: https://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/

Macklin, R. (2003). Applying the four principles. Journal of Medical Ethics, 29, 275-280.

Martin, K. E. (2015, June). Ethical Issues in the Big Data Industry. MIS Quarterly Executive, pp. 67-85.

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., et al. (2011). The eMERGE Network: A consortium of biorepositories linked to

electronic medical records for conducting genomic studies. BMC Medical Genomics, 4(13).

McNulty, C. A., Boyle, P., Nichols, T., Clappison, P., & Davey, P. (2007). The public's attitudes to and compliance with antibiotics. Journal of Antimicrobial Chemotherapy, 60 (Supplement), i63-i68.

McQuinn, A. (2017, October 6th). The Economics of "Opt-Out" Versus "Opt-In" Privacy Rules. Retrieved October 18th, 2017, from itif: https://itif.org/publications/2017/10/06/economics-opt-out-versus-opt-in-privacy-rules

MIDAS Consortia. (2016). MIDAS-DOA-20160731.

Miller, S., & Hugh, D. (2017). The Quant Crunch: How the demand for data science skills is disrupting the market. Boston: Burning Glass Technologies.

Mullich, J. (2013). Closing the Big Data Gap in Public Sector. Bloomberg Businessweek.

NHS England. (2013, October 16th). NHS England. Retrieved from News: NHS England sets out the next steps of public awareness about care.data: https://www.england.nhs.uk/2013/10/care-data/

Nissenbaum, H. (2010). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford: Stanford University Press.

Noakes, J., & Pridham, G. (1974). Documents on Nazism 1919-1945. New York: Viking Press.

OECD. (2000). Reducing the Risk of Policy Failure: Challenges for Regulatory Compliance. OECD.

Oxford English Dictionary. (2017, October 2). Consent. Retrieved from Oxford English Dictionary: https://en.oxforddictionaries.com/definition/consent

Oxford English Dictionary. (2017). Ethics. Retrieved from Oxford English Dictionaries: https://en.oxforddictionaries.com/definition/ethics

Oxford English Dictionary. (2017, September 30). Ethics Definition. Retrieved September 30, 2017, from Oxford English Dictionary: https://en.oxforddictionaries.com/definition/ethics

Oxford English Dictionary. (2017, October 2). Informed Consent. Retrieved from Oxford English Dictionary: https://en.oxforddictionaries.com/definition/informed_consent

Patki, N., Wedge, R., Veeramachaneni, K. (2016). The Synthetic Data Vault. *In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Montreal, QC, Canada, 17-19 October 2016, IEEE.

Paul Ohm (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymisation. 57 UclaLaw Review 1701-1777

Prakash, S., & Valentine, V. (2007, November 10th). Timeline: The Rise and Fall of Vioxx. Retrieved October 21st, 2017, from NPR: http://www.npr.org/templates/story/story.php?storyId=5470430

Presser, L., Hruskova, M., Rowbottom, H., & Kancir, J. (2015). Care.data and access to UK health records: patient privacy and public trust. Technology Science.

Solove, D. J. (2012). Privacy Self-Management and the Consent Dilemma. Harvard Law Review.

Strous, R. D. (2007). Psychiatry during the Nazi era: ethical lessons for the modern professional. Annals of General Psychiatry, 6(8).

Sweeney, L. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570.

Sweeney, L. (2016). Datafly: a system for providing anonymity in medical data. In T. Lin, & S. Qian, Database Security XI: Status and Prospects (pp. 356-382). Springer.

Ted is writing things. (2017). On privacy, research, and privacy research: k-anonymity, the parent of all privacy definitions. https://desfontain.es/privacy/k-anonymity.html

Tene, O., & Polonetsky, J. (2012). Privacy in the Age of Big Data: A Time for Big Decisions. Stanford Law Review.

The British Psychological Society. (2009). Code of Ethics and Conduct: Guidance published by the Ethics Committee of the British Psychological Society. Leicester: The British Psychological Society.

The European Parliament and the Council of Europe. (2016). On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union.

The European Parliament and the Council of the European Union. (2016). REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union.

The European Parliament and the Council of the European Union. (1995, October 23). DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Brussels.

The Independent. (2017, October 12th). World Leaders Urged to Act on "Post-Antibiotic Apocalypse" by Chief Medical Officer. Retrieved from Independent: http://www.independent.co.uk/life-style/health-and-families/health-news/antibiotics-resistance-apocalypse-warning-chief-medical-officer-professor-dame-sally-davies-drugs-a7996806.html

The Lancet Infectious Diseases Commission. (2013). Antibiotic resistance - the need for global solutions. London: The Lancet.

The Organization for Economic Co-Operation and Development. (1980). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. OECD.

United States Senate: Committee on Finance. (2004, November 18th). Testimony of David J.Graham, MPH, November 18. 2004. Washington, Washington DC, United States.

US Census Bureau.(2017). Information Quality Glossary. https://definedterm.com/synthetic_data.

Wellcome Trust. (2016). The One-Way Mirror: Public attitudes to commercial access to health data: Report prepared for the Wellcome Trust. London: Ipsos MORI.

Westin, F. A. (1979, September 12). Computers, Personnel Administration, and Citizen Rights (1979). Retrieved from Heinonline: file:///C:/Users/paul.carlin/Downloads/AlanFWestinComputersPerso.pdf

Wiener, J. B. (2004). The regulation of technology, and the technology of regulation. Technology in Society, 483-500.

World Medical Association. (2013). Declaration of Helsinki (7th revision). Geneva: World Medical Association.

Wu, F. T. (2013). Defining Privacy and Utility in Data Sets. University of Colorado Law Review, 117-1178.

Younkins, E. W. (2000, April 5th). The Evolution of Law. Montreal, Quebec, Canada.

Grant Agreement No: 727721

9 Glossary

DMSB - Data and Safety Monitoring Board
ECHR - European Convention on Human Rights
EPAG - Ethics and Privacy Advisory Group
EU - European Union
EULA - End User Licence Agreement
FDA - Food and Drug Administration
GDPR - General Data Protection Regulations
HIV - Human Immunodeficiency Virus
MIDAS - Meaningful Integration of Data, Analytics and Services
OECD - The Organisation for Economic Co-operation and Development
PHE - Public Health England
PID - Personalised Identifiable Data
VIGOR - Vioxx Gastrointestinal. Outcomes Research

Grant Agreement No: 727721

10 Appendix 1: MIDAS Ethics and Privacy Advisory Group (EPAG) - Terms of Reference

1. Constitution

The MIDAS Project Policy Board (POLICY BOARD) hereby resolves to establish a project specific group to be known as the Ethical and Privacy Advisory Group (EPAG).

2. Membership of the Project Group

- Independent Lead: Dr Siobhan McGrath
- Paul Carlin, South Eastern H&SC Trust (WP2 Lead)
- Professor Jonathan Wallace and Dr Michaela Black, University of Ulster (WP1, WP7 and WP8 Leads)

3. Quorum

A quorum shall be a total of 3 of the members of the committee or their nominated deputies, if the independent lead is not present all decisions must be ratified before actioned.

4. Frequency of Meetings

The committee shall meet every four months until the project has been designed. Thereafter every six months until overall project completion. The committee members may delegate in their absence, if on any occasion the member/ representative cannot attend then the agenda will be assumed to be agreed and all voting will be carried forward in the affirmative, excluding as stated in 3.

5. Authority

The Committee is authorised by the POLICY BOARD to undertake any activity within its terms of reference. In particular, it may seek advice from whatever source it deems to be appropriate in order to fulfil its function.

Grant Agreement No: 727721

6. Role and Responsibilities of EPAG

The role of the EPAG is to act as the Ethics and Governance group for the MIDAS project. It is be responsible for the design, application to all regulatory bodies, co-ordination, monitoring and facilitation of the overall project, with particularly regard to WP2.

The main responsibilities of the Group are:

- To guide, support and monitor all project activity in regards to Ethics, Privacy and Governance.
- To produce clear guidance in developing the project/s, to ensure that it meets the needs of the MIDAS project and the European Commission moving forward.
- To develop a robust project research protocol/s that meets the required scientific standard.
- That all materials are reviewed in a timely manner when informing the design of the overall project and the specific projects in WP2.
- That standard operating procedures are developed within the WP groups that ensure that all partners are aware of the ethical and good practice requirements and work together to achieve the overarching aims of WP 2 within the overall project aims and objectives.
- To coordinate with stakeholders in the development of the protocol.
- To manage the risk involved in the project.
- To provide regular (project driven) updates to the POLICY BOARD.
- Provide final sign off on all project materials before submission to all relevant regulatory bodies.

7. Operational Arrangements for Meetings

7.1. Administrative support to EPAG

EPAG shall be supported administratively by the University of Ulster :

- Preparation of the agenda in conjunction with the Chairman and issue of agenda on behalf of the Chairman;
- Distribution of papers sufficiently in advance of each meeting to facilitate their full consideration and discussion at the meeting (nominally 1 week in advance);

Grant Agreement No: 727721

- Ensuring appropriate arrangements are in place for the servicing of the committee including the taking of minutes and keeping a record of matters arising and issues to be carried forward.

7.2. Conduct of meeting

All questions arising will be decided by a simple majority of those present. In the case of equal votes, the Chair will have a casting vote. It is intended that meetings will not last more than 2 hours. Those stakeholders not in attendance will be taken to vote in the positive unless other information is received by the group before a meeting.

7.3. Agenda items and papers for meetings

Agenda items should be submitted to Ulster 14 days in advance of the meeting. He/she will agree the content of the agenda prior to issue with the chairman of the committee.

Ulster will issue the agenda/papers for the meeting approximately 7 days in advance of the meeting.

Should an item need to be raised on the day, this can be covered under Any Other Business, subject to there being available time for discussion. If separate papers require circulation, these should, wherever possible, be issued with the agenda. This is intended to enable the members to have the opportunity to read information in advance.

7.4. Minutes of meetings

Ulster will provide the secretariat for the meeting. Minutes of meetings will be produced and agreed with the chair prior to issue. These will be circulated as soon as possible after the meeting listing topics discussed, actions agreed and individuals responsible for undertaking those actions.

7.5. Review of Terms of Reference

The Policy Board will review its terms of reference on an annual basis and should endorse these formally.

Grant Agreement No: 727721

8. Reporting

The minutes of EPAG shall be formally recorded and distributed to the members of EPAG and presented to the next Policy Board meeting for information and noting.

9. EPAG within Proposed Governance Arrangements

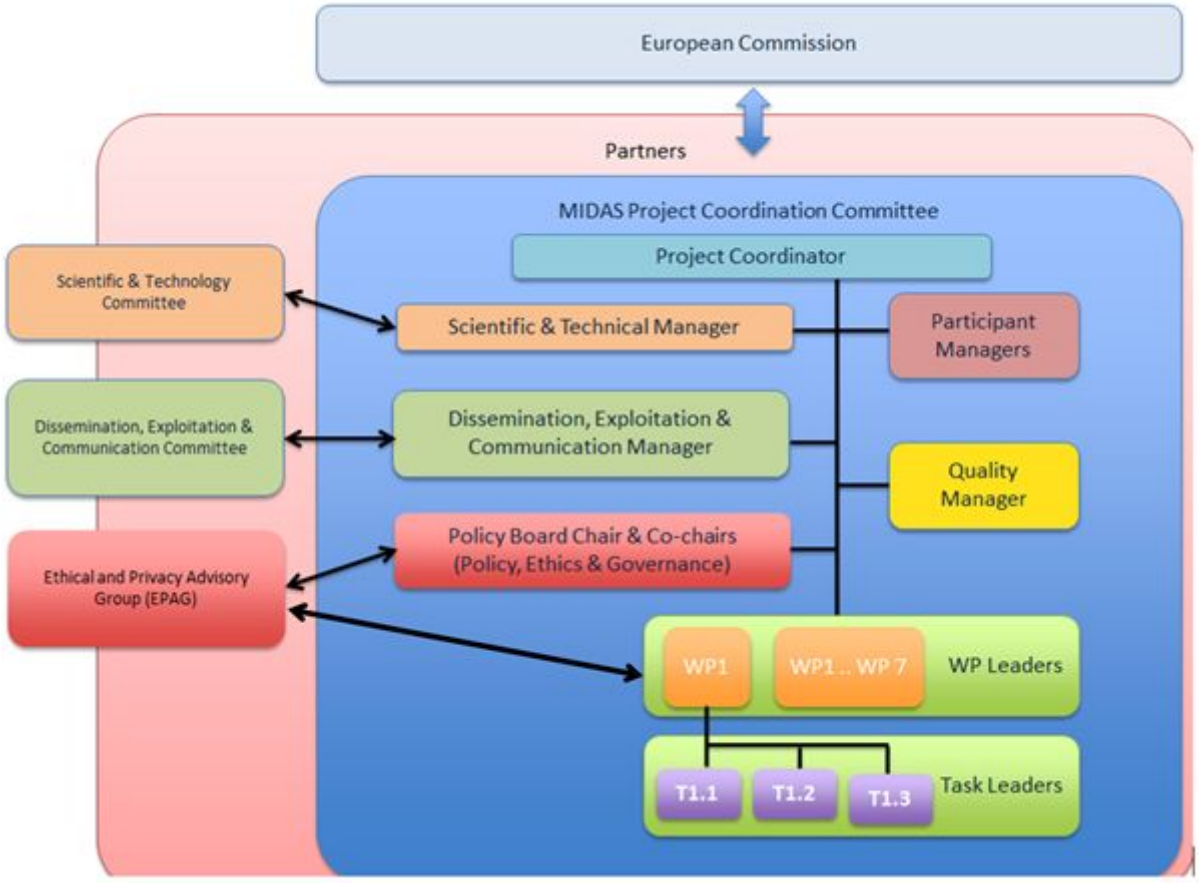
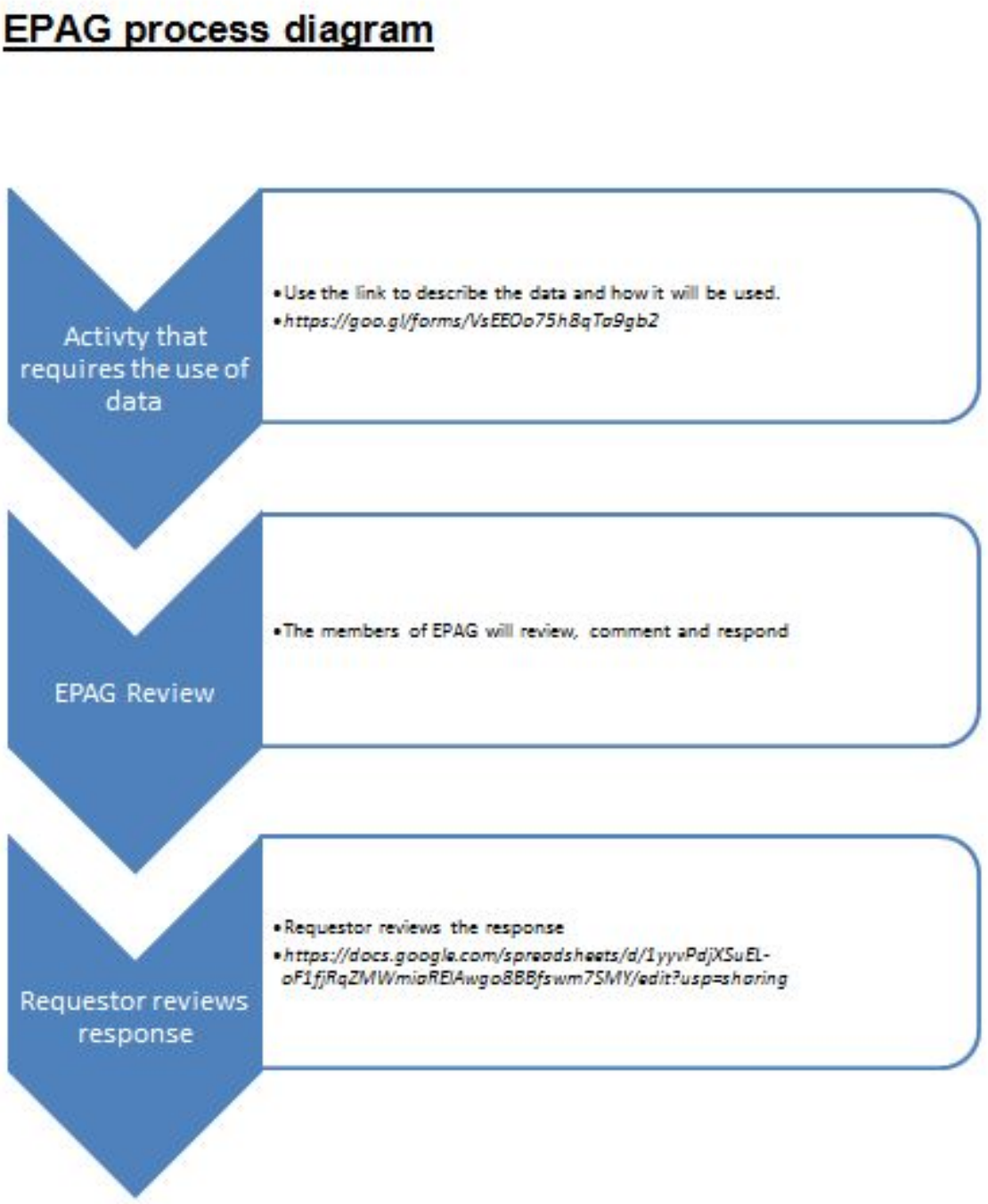


Figure 10.a EPAG within Proposed Governance Arrangements

Grant Agreement No: 727721

11 Appendix 2: EPAG Process Diagram



Grant Agreement No: 727721

12 Appendix 3: Policy Board Baseline Questionnaire

Version 1
March 16, 2017
ETHICS AND GOVERNANCE BASELINE

Work Package 2: Good practice and legislative baseline review

Partner Organisation:
Name:
Country:

Type of Organisation:

- Academic
- Industry
- Public sector
- Other

Questions:

1	Do you use sensitive data (e.g. identifiable information): Yes No
2	Do you use publically available data: Yes No
3	Do you use this data for core business purposes, or for research purposes or both? Yes No Both

4	Is this data currently used to inform government policy? Yes No (if the answer is yes, please proceed to 5, if no, please proceed to question 6)
5	Please provide the process/ procedure documents that allows the use of data for policy makers (link to e-document)
6	Do you have a code of a practice you must adhere to for sensitive data? (this will include ethics and Governance processes) Yes No (if the answer is yes, please proceed to 7, if no, please proceed to question 8)
7	Please provide the code of practice documents for sensitive data (link to e-document)
8	Do you have a code of a practice you must adhere too for public data? (this will include ethics and Governance processes) Yes No (if the answer is yes, please proceed to 9, if no, please proceed to question 10)

9	Please provide the code of practice documents for sensitive data (link to e-document)
10	What legislation is relevant for the use of sensitive data in your Country? (please give title and link to document)
11	Has there been a public consultation in the last 10 years for personal data use and sharing in your country? Yes No (if the answer is yes, please proceed to 12, if no, please proceed to question 13)
12	Please provide the consultation report for sensitive data (link to e-document)
13	Do you have an opt in/opt out model (if yes please give title of report and link to e-document) Yes No

Grant Agreement No: 727721

13 Appendix 4: Risk Assessment Tool for Datasets

Risk Assessment MIDAS Datasets

Version: 1.00	
Action Required:	This risk assessment tool is intended to be done quickly beforeusing any data for MIDAS.
	It will help the Organisation assess any concerns about the data which may delay project delivery. In certain circumstances it may indicate that the project should not be considered further.
	Your assessment is proportionate to the risks associated with undertaking the project.
	All questions should be answered based on your understanding of data management, legislation and good practice within both your local context and the context of MIDAS.
	As a planning aid, please add a short note on concerns and management actions (proportionate to the risks associated with undertaking theproject) needed to address these concerns during assessment and project delivery.
	Where the response to any question is unclear (e.g. because it is not possible to discuss quickly with the relevant person) then this would indicate a risk that may need to be managed later in the process, and so this may indicate option 3 or 4 is appropriate.
Background:	**This risk assessment will quickly highlight any concerns that the Participating Organisation may have with a project early in, or before commencing the project within the context of MIDAS.**
	The assessment is based on theMIDAS teams capabilities to support a specific project at a specific time (i.e. Organisation capabilities may vary over time for operational reasons).
	Generally it is completed by the Participating Organisation's MIDAS lead.
	Some areas may require early a short discussion between partners as well as other contacts in the Organisation (e.g. service managers) to share opinions on the project.
	Understanding any concerns quickly means that they can be addressed early within the MIDAS consertium.
Notes:	The 'Participating Organisation' (or similar expressions) is used to mean any partner organisation within the MIDAS consortia.
	A 'Participating Organisation' may be providing data, hosting data, analysing data or preparing data. This may include display of data and interaction.
	The Risk Tool can be used when there is a significant change to any of the areas described to support changes to management actions.
	This assessment should be undertaken for each project by each Participating Organisation.
Assumptions:	The Organisation has been provided with sufficient and relevant information so as to assess its readiness (and risks) appropriately.
Accountability:	The Participating Organisation's board

Grant Agreement No: 727721

Risk Assessment MIDAS Datasets Version: 1.00

Organisation:
EPAG reference:
Completed by: on:
Project/

The assessment of risk in respect of the data to be used and the ability to contribute to the project for delivery.

1. using existing/proven processes and arrangements in Organisation.
2. With only minor changes processes and arrangements in Organisation.
3. Using new/changed processes and arrangements in Organisation.
4. May not be practical to achieve within a reasonable timetable.
0. Unknown. Sufficient information not yet provided.

1	2	3	4	0

Select the most appropriate option *Add notes on any management actions required*

		1	2	3	4	0
A	**Study Legal and Regulatory Needs** Please assess the the reuqirements for local and regulatory needs	◉	○	○	○	○
B	**Local Alignment** Is the Organisation already aware of any constraining factors currently that may impact on their ability to deliver the project?	◉	○	○	○	○
C	**Organisational Team** Is it likely that the Organisation will be able to provide support to MIDAS to meet the demands of the project .	◉	○	○	○	○
D	**Further Agreements** Are there any requirements for further agreements/ contracts required in respect of the project	◉	○	○	○	○
E	**Consent** Is it likely that the project will require the use of data that requires consent? If so can the organisation assure that this will be faciliatated and operated to all good practice requirements?	◉	○	○	○	○

1. Data is open source
2. Data is anonymised.
3. Data is psuedoanonymised.
4. Data is identifable.
0. Unknown. Sufficient information not yet provided.

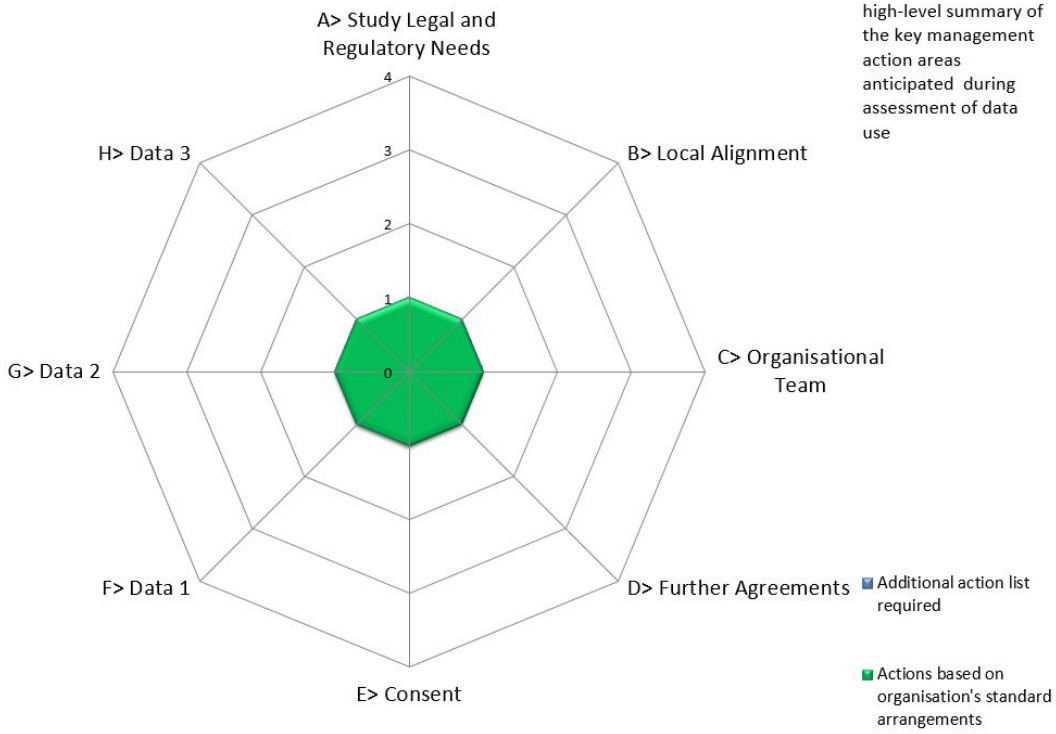
1	2	3	4	0

Select the most appropriate option *Add notes on any management actions required*

		1	2	3	4	0
F	**Data 1** Has the organisation identified an appropriate data source for MIDAS?	◉	○	○	○	○
G	**Data 2** Has the organisation assessed the status of the data?	◉	○	○	○	○
H	**Data 3** Has the organisation assessed the technology required for analysis and/or visualisation within the consortia?	◉	○	○	○	○
I	**Data 4** Has the organisation assessed the potential for reidentification of the data?	◉	○	○	○	○

Grant Agreement No: 727721

Participating Organisation: Summary of Key Management Action Areas



Grant Agreement No: 727721

14 Appendix 5: Data Access Agreement Template

midas

Meaningful Integration of Data, Analytics and Services

Grant Agreement No. 727721

Contract Duration: 40 months (1st November 2016 – 29th February 2020)

This project is funded by
the European Union

H2020-SC1-2016-CNECT
SC1-PM-18-2016 - Big Data Supporting Public Health Policies

MIDAS – Data Access Agreement

Non Identifiable Data for Use between partners in the MIDAS project

Grant Agreement No: 727721

Grant Agreement No:
727721

Copyright

The MIDAS project is funded under the EC Horizon 2020 SC1- PMF-18 Big Data Supporting Public Health Policies

Grant Agreement No: 727721

Grant Agreement No:
727721

Document History

Version	Issue Date	Stage	Content and Changes
1.0	dd/mm/yyyy	Draft/Final	

Grant Agreement No: 727721

Grant Agreement No:
727721

Table of Contents

Grant Agreement No: 727721

Grant Agreement No:
727721

Introduction

There is a requirement for demonstrator projects within MIDAS for datasets to be accessed by a number of partners. To ensure that all partners understand the limits of use and what is permitted within the context of MIDAS, the following defines the process and information required for approval before any output can be released outside of the project.

Partners requiring access to the datasets hosted by Ulster University should follow this process.

For accessing datasets outside of the project, each partner must follow their local processes and work within the relevant legislative frameworks.

Any data attributed to the project or knowledge extracted from it cannot and must not be released, discussed, presented, published or reviewed outside of the consortium without following this process.

In the event of a breach of this agreement, which results in a financial penalty, claim or proceedings, the parties agree to co-operate to identify and apportion responsibility for the breach and the defaulting party will accept responsibility for any such claim.

Please refer to Appendix 1, 'Principles Governing Information Sharing' for guidance.

The form is divided into Sections (A-G) as follows:

Section A: Details of Requesting Organisation
Section B: Commissioning Organisation
Section C: Details of data items requested
Section D: Consent issues
Section E: Data Protection
Section F: Declaration: Requesting Organisation
Section G: Declaration: MIDAS Executive Board

Appendix 1: Principles Governing Information Sharing

Grant Agreement No: 727721

Grant Agreement No:
727721

MIDAS DATA ACCESS REQUEST FORM

MIDAS Data Access Request Reference: _____

MIDAS Data Access Request Contact Name (*contact for whom the data is being requested from*): _____

MIDAS Data Access Request Organisation: _____

Title of Agreement: _____

Date of Request: _____

Date Access Begins: _____

Date Access Ends: _____

Review date if on-going agreement: _____

Please state if this is an update of a previous agreement or a new request for information.

An update of an earlier extract ☐ New application ☐

6

Grant Agreement No: 727721

Grant Agreement No:
727721

Section A: Details of Requesting Organisation

(A) Details of Requesting Organisation	
Name of Requesting Organisation *Please note that the Data Access Agreement will be immediately returned unless the requesting organisation has signed section F.*	
Name of Authorised Officer Requesting Access to MIDAS Data *Please print*	
Position/Status	
Organisation	
Email Address	

7

Grant Agreement No: 727721

Grant Agreement No:
727721

Section B: Commissioning Organisation

If you require the data to carry out work on behalf of another partner, please complete section (B) below. If not, please go straight to section (C).

(B) Commissioning Organisation	
Name of Commissioning Organisation	
Contact Name	
Title	
Email Address	

Grant Agreement No: 727721

Grant Agreement No:
727721

Section C: Details of Data Items Requested

(C) Details of Data Items Required:	Rationale for Data Items
Please provide a list and description of the data to which the request applies, e.g. include all identifier attributes, (e.g. Postcode, Date of Birth, Gender, Diagnosis Code, Religion etc.) **(broader descriptors can be used for large numbers of variables)** ***This section should be completed in conjunction with the host and/or owner of the data.***	Please indicate the reasons for requiring each of these data items. This can be generic.

Processing of Data	
Please indicate how you propose to process the data once received	
Please state in as much detail as possible the purpose for which the data are required by the partner named in section (A) including any record linking or matching to other data sources. Please continue on a separate sheet if necessary or attach any relevant documentation.	

Grant Agreement No: 727721

Grant Agreement No:
727721

Other Data that may be linked (please list other data sets that may be linked)	
Will any other Data contain Identifiable Details?	*Please tick* Yes ☐ No ☐
Frequency of transfers	Once Only ☐ Other ☐ *Please specify if other:* _____

Section D: Consent Issues

(D) Consent Issues	
Do you have the individuals' consent if using any other data set?	*Please tick* Yes ☐ No ☐
If yes, please provide a copy of the consent form i.e. explicit consent should be obtained for the processing of sensitive personal data.	

Grant Agreement No: 727721

Grant Agreement No:
727721

Section E: Data Protection

(E) Data Protection (of Requesting Organisation)	
Do you have a confidentiality / privacy policy that complies with the Data Protection Act 1998? (As data is hosted in UK)	*Please tick* Yes ☐ No ☐
Are confidentiality clauses included within contracts of all staff with access to the data?	*Please tick* Yes ☐ No ☐
Are all staff trained and aware of their responsibilities under the Data Protection Act 1998 and do they adhere to the eight Data Protection Act Principles?	*Please tick* Yes ☐ No ☐
If you answered no to the question above, how can you assure the reviewers that you meet a minimum standard in respect of data protection?	Please insert text as appropriate:

Section F: Declaration: Requesting Organisation

(F) Declaration: Requesting Organisation

Data Protection Undertaking on Behalf of the Organisation Wishing to Access the Data

My organisation requires access to the data specified and will conform to the all good practice requirements defined below:

I confirm that the information requested, and any information extracted from it,

- Is relevant to and not excessive for the stated purpose
- Will be used only for the stated purpose
- Will be not be removed, copied, transferred, screenshot, or any way duplicated form the protected hosted environment
- Will be used no longer than is necessary for the stated purpose

I *(name:printed)* _____, as the Authorised Officer/ signatory of *(Organisation)* _____, declare that I have read and understand my obligations and adhere to the conditions contained in this Data Access Agreement.

Signed: _____
(Authorised Signatory/Officer)

Date: _____

Section G: Declaration: MIDAS Executive Board

(G) Declaration: MIDAS Executive Board

DATA ACCESS AGREEMENT

I CONFIRM THAT:

Ulster University on behalf of the MIDAS consortium consents to the access of the data specified, to the organisation identified in Section A of this form.

Executive Board Chair *(Michaela Black)*

Signed: _____

EPAG Chair *(Paul Carlin)*

Signed: _____

Scientific-Technical Manager *(Austin Tanney)*

Signed: _____

Date: _____

Any loss, theft or corruption of the shared data by the requesting organisation must be reported immediately to the Personal Data Guardian of the owning organisation. Please also note that any serious breaches, data loss, theft or corruption should also be reported to the MIDAS Executive Board by the Data Controller.

1. *Ulster University/EPAG/the data owner will assist in completion of: the Introduction and Section C, and will complete Section G upon receipt of completed DAA*
2. *Requesting partner to complete: Section A, B (if applicable), Section C (Processing of Data and Linked Data sections), Section D, Section E, Section F*
3. *For specific data access requirements (including modifications to the terms of the DAA) with partners outside the consortia please contact EPAG in the first instance*

Grant Agreement No: 727721

Appendix 1: Principles Governing Information Sharing[1]

Code of Practice 8 Good Practice Principles[2]	DPA Principles	Caldicott Principles[3]
1. All organisations seeking to use confidential service user information should provide information to service users describing the information they want to use, why they need it and the choices the users may have. 2. Where an organisation has a direct relationship with a service user then it should be aiming to implement procedures for obtaining the express consent of the service user. 3. Where consent is being sought this should be by health and social care staff who have a direct relationship with the individual service user. 4. 'Third Party' organisations seeking information other than for direct care should be seeking anonymised or pseudonymised data. 5. Any proposed use must be of clear general good or of benefit to service users. 6. Organisations should not collect secondary data on service users who opt out by specifically refusing consent. 7. Service users and/or service user organisations should be involved in the development of any project involving the use of confidential information and the associated policies. 8. To assist the process of pseudonymisation, the Health and Care Number should be used wherever possible.	1. Data should be processed fairly and lawfully. 2. Data should be processed for limited, specified and lawful purposes and not further processed in any manner incompatible with those purposes. 3. Processing should be adequate, relevant and not excessive. 4. Data must be accurate and kept up to date. 5. Data must not be kept longer than necessary. 6. Data must be processed in line with the data subject's rights (including confidentiality rights and rights under article 8 of the Human Rights Act). 7. Data must be kept secure and protected against unauthorised access. 8. Data should not be transferred to other countries without adequate protection.	1. Justify the purpose(s) for using confidential information. 2. Only use it when absolutely necessary. 3. Use the minimum that is required. 4. Access should be on a strict need-to-know basis. 5. Everyone must understand his or her responsibilities. 6. Understand and comply with the law.

[1] These principles must be followed by health and social care organisations when considering use and disclosure of service user information.
[2] Code of Practice, paragraph 3.17.
[3] PDG Principles are adopted from the Caldicott Principles established in England and Wales.