Fair Processing Information for the Education and Diabetes Project

About this document

This fair processing information document (also known as the privacy notice or patient information sheet) has been developed specifically for a Cardiff University project. This document is important because it concerns the processing of confidential patient information which are to be processed for use in research without patients' consent. This document explains how the organisations responsible for looking after the source datasets can share them under the direction of Cardiff University as the sole data controller for this project. This document also includes information on how patients can opt-out of having their data included in the study.

This document was last updated on 03/03/2025 (Version 6.1). Please email the project team at Steadfast@cardiff.ac.uk , or the project lead, Robert French (frenchr3@cardiff.ac.uk) with any queries relating to this document, if you require further information, or have suggestions on how we could present this information better.

Fair processing information for the STEADFAST project data linkage

Summary of the essential information on a single page

Our patient panel requested that we include the following information on the first page of our fair processing document.

Project information: Further information on the whole research project (going beyond the data processing covered in this document) is included on the Cardiff University project website, though a more user-friendly source is the page on the current funder's website (Welsh version here) or the blog at the start of the latest phase of this work.

How patients can 'opt-out' their data from the research: Patients can opt-out in two ways: (i) by opting out of the processing of their health data for planning and research using the national opt-out, (ii) by opting out of specific administrative datasets by contacting the controllers of those datasets. More details on these options are provided later in the document. Please note that once the processors delete the real-world person IDs (Name, date of birth, NHS number) from the project data, it is no longer possible for us to re-identify and remove people from the study.

Patients' views on this data processing: Cardiff University have ongoing consultation with people living with diabetes on how their data may be processed and used. Information about how we consulted young people is provided on the webpage for the first 2021 engagement workshop, the workshop report, and the blog from that event by the Diabetes UK lead for Type 1 Diabetes. We have also recorded the presentations given in our patient engagement sessions, Including Alex Bailey from the Medical Research Council Regulatory Support Centre speaking about using health data for research, Lucie Burgess speaking about the Diabetes UK perspectives on data, Rob French speaking about the specific data usage in the research study, and also feedback from three of the participants. In the second workshop, the young people produced a video explaining how researchers use data relating to young people with diabetes. Although the video and the report produced from that workshop have not been formally released, there is a version available on YouTube. One can also view the video for the paediatric diabetes audit from that session. We conducted a large set of engagement sessions with young people during 2022 in partnership with Diabetes UK; the official website for that work is here, during 2023 and 2024 we worked with other data users to refine the engagement toolkit created as part of that engagement. In February 2024 we secured funding for a one-off meeting with our young person representatives and the Children's Commissioner for Wales to discuss some of the themes in the data and research. There are further engagement plans over the coming year, though we are still finding ways to resource this work, please contact the project team at if you would like to get involved.

Introduction

This fair processing document aims to inform research participants about how their administrative health and education records will be processed and used for our research study in a fair, legal, and ethical way. This document will be made available to participants via the project website. Cardiff University will combine different datasets to create a linked research dataset which tracks the education and health of young people with diabetes; we show the data linkage mapped against age in Figure 1 below.

The key privacy concern for the data processing relates to the flow of identifying information (e.g., NHS number, name, postcode, date of birth, gender) from the health data providers NHS England (NHS England, previously known as NHS Digital) and Royal College of Paediatrics and Child Health (RCPCH). Because these identifiers help relate health data to an individual they are referred to as ‘confidential patient information’, such data has a special protection under common law known as the ‘duty of confidentiality’. While we describe all the data processing in this document, we include a paragraph on these key data flows in this introduction below.

The project has four data processors who use the confidential patient data, to create a ‘deidentified linkage ID’ (e.g., A12341234) on behalf of the data providers. For the Welsh linkage, the processor is Digital Health and Care Wales (DHCW, previously NHS Wales Informatics Service, NWIS). For the English linkage, there are approvals in place from HRA CAG or three processors: the Department for Education (DfE), NHS England Leeds Data Production Team (NHSD-LDP), and the Office for National Statistics Data Access Platform (ONS-DAP). Once the data processors have created the unique but deidentified linkage ID, the confidential information can be destroyed, and the linkage ID can be used to facilitate linking administrative health and education data. The purpose of replacing the confidential patient information with a deidentified linkage ID is to protect privacy by rendering the data not traceable back to an individual.

There are three formal legal gateways for processing the data in the way we have described:

1. Because the data shared for linkage are ‘personal data’, the GDPR legal basis for Cardiff University to use such data is known as ‘Public Task’ ‘Article 6 1 (e)’ (processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller), Cardiff University (a public authority) will perform a task (research) in the public interest.
2. Because those parts of the data related to health are ‘special category’ (more sensitive) personal data, Cardiff University must satisfy one of the conditions available in GDPR to process such data. We use ‘Article 9 2 (j)’ “processing is necessary for archiving purposes in the public interest, scientific or historical

research purposes or statistical purposes", where our focus is on scientific research purposes.

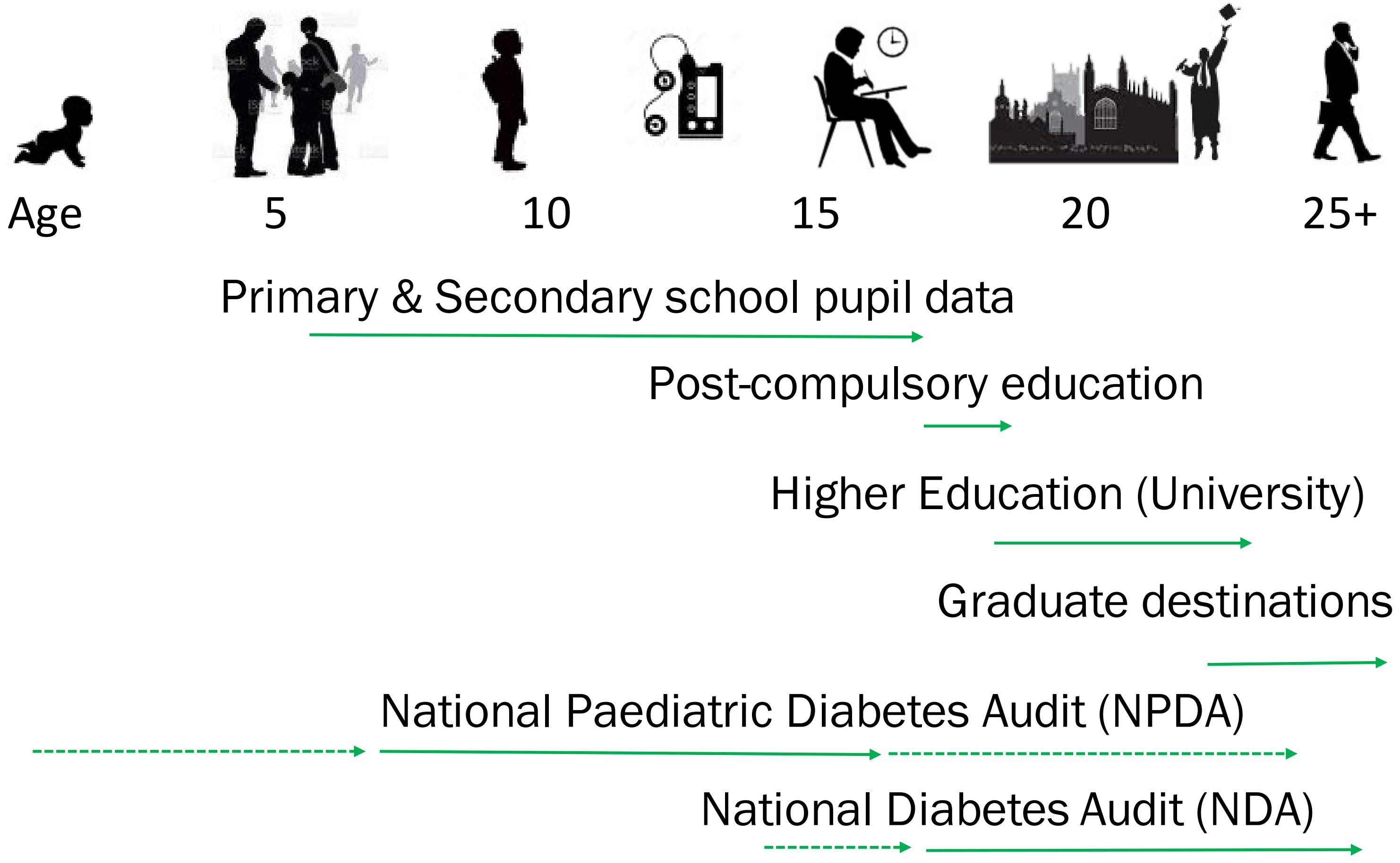
3. Because we are also processing confidential patient information (patient identifiable information) without consent, we also need to address the Common law duty of confidentiality. Cardiff University made a 'Section 251' application to the Health Research Authority (HRA) Confidentiality Advisory Group (CAG), who assess our application and advise the Secretary of State whether to set aside the common law duty of confidentiality for our specific purpose. The Section 251 approval letters for the STEADFAST data project are available on the project webpage. More information is given on the specific flows which relate to this in the 'How data will be flowed for linkage' section of this document.

Aims of the project

The overall aim of the project is for Cardiff University to link data from multiple sources to generate a research dataset which tracks the education and health of young people with diabetes. This data will be used for research, in the first instance to quantify the associations between educational outcomes (e.g., school absence) and health outcomes (e.g., blood glucose levels), though there are many more nuanced questions that can be asked of this data. The benefit to health and social care is through providing evidence to practitioners to help them to take account of the links between education and diabetes management to help inform differential care, for example, for example whether some individuals struggle to cope with self-management amongst the other pressures of moving to university.

Children of school and university age with diabetes are most frequently living with type 1 diabetes, accounting for approximately 98% of cases, so the research primarily focuses on these individuals, however education and health outcomes for children with type 2 diabetes (and other rarer forms of diabetes) are equally important, and the analysis will be done for young people living with each type of diabetes.

Figure 1: Age Coverage for each of the linked datasets



Objectives of the project and data linkage

The Health Research Authority (HRA) ensures good governance for research studies in health and social care research and is responsible for the process under which we seek to share confidential patient information. This approval process facilitates the deidentified linkage across datasets. Specifically, we will link the diabetes-related health measures from diabetes audits to educational records of individuals collected by schools and universities. This data will then be used for health research, for example, to quantify (i) how trajectories of diabetes management (HbA1c measures) are associated with educational outcomes (ii) how educational outcomes affect trajectories of diabetes management.

To link datasets, the data providers will share identifiers (including NHS number, name, date of birth, gender, and postcode) with the data processors to generate deidentified linkage IDs to facilitate linkage in secure repositories. This processing is summarised in Figures 2 and 3 below. There are two data repositories in this study where the analysis of the deidentified linked data will take place. The Welsh data linkage repository is the 'Secure Anonymised Information Linkage Databank' (SAIL) at Swansea University. The English data linkage repository is the Office for National Statistics Trusted Research Environment. Currently this is the 'Secure Research Service' (ONS-SRS), though this is likely to change to the Office for National Statistics 'Secure Data Service' (ONS-SDS) in 2025/6.

The data is shared between the data provider and linkage repository in two separate flows, one flow for the identifiers and another flow for the substantive data. The patient identifiers are not sent directly to the data linkage repository, but rather to a third-party processor, who uses them to create a privacy-preserving linkage ID on behalf of the linkage repository. Once the processors have assigned the linkage ID to a dataset, they delete the real-world IDs as soon as possible, typically within three weeks of receiving the data. The processors then flow the linkage ID (without any real-world IDs, since these have now been deleted) to the repository. The data providers share the substantive health information directly to the data repository where they are joined to the linkage IDs provided by the processor. Different datasets can be linked using the linkage IDs. The data repositories check the linked data before it is shared with the researcher for analysis to ensure that the linkage has not inadvertently made cases re-identifiable.

How data will be flowed for linkage

Unfortunately, due to technical and information governance constraints, we are not able to flow all of the schools data, university data, paediatric diabetes audit data, and adult diabetes audit data for England and Wales into a single repository for linkage. Instead, we must use two separate repositories. The first has all the data except English schools data; the second has all the data except Welsh schools data. Researchers use statistical techniques to combine the analysis from the two repositories into a single model result.

Figure 2 shows how the datasets (except English schools data) will flow into the first repository. The dashed arrows coming from the top show the flows of substantive health and education data; these do not include real-world identifiers directly relating the data to the individual concerned. The solid arrows show the flows of identifiers containing information that directly identifies individuals (NHS number, name, date of birth and postcode). These identifier datasets do not contain any substantive data alongside the identifiers. To flow the identifiers relating to health data requires the research team to get a special approval from HRA CAG known as a Section 251 exemption, ensuring sufficient safeguards and patient information are in place to protect the patient’s privacy. One of the safeguards in the processing for this project is to flow and process the identifying information separately from the substantive health data, reducing the chance that in the unlikely event of a data breach, the extracts will not contain identifiable sensitive data. Another safeguard is using another organisation to replace the real-world IDs with a project-specific ID which cannot be traced back to the original real-world IDs. These project-specific IDs are reunited with the substantive data in the repository using an additional common ID created by the original data provider.

Figure 3 shows how the datasets (except Welsh schools data) will flow into the second repository. Again, the dashed arrows show substantive health and education data flows, but no real-world identifiers directly relating the data to the individual concerned. The flows of real-world identifiers shown by the solid arrows present a greater challenge than the equivalent flows for the first repository shown in Figure 2. This is because the National Paediatric Diabetes Audit records include the patients NHS number but not the patient’s name, however the processor used to create the project-specific linkage ID from the real-world IDs can process names but not process NHS numbers. To overcome this, we use NHS England to add each person’s name to the list of real-world IDs. As for the first repository, Cardiff University must obtain support (Section 251) from the Health Research Authority (HRA) to ensure the proper safeguards and information sharing are in place.

Figure 4 shows the flows for linage described in Figure 3, but using alternative processors for the real-world IDs for linkage in ONS. The reason we need to use an alternative processer is because we think they may be able to create better quality linkage IDs, improving the linkage rate and hence the data and analysis creating more robust evidence to better inform the improvement in clinical and educational practice.

Figure 2: Data flows (excluding English schools data) into the first data repository

Figure 3: Data flows (excluding Welsh schools data) into the second data repository using DfE to create the linkage IDs

Figure 4: Data flows (excluding Welsh schools data) into the second data repository using ONS to create the linkage IDs

The identity of the researcher and their relationship with the University/department

The Data Controller for the project is Cardiff University. Cardiff University is a controller in the sense that they are choosing and directing the processing used to create the linked research dataset, Cardiff University do not themselves handle or store the data, this is all carried out by other parties. The principal investigator (lead researcher) is Dr Rob French, Principal Research Fellow in the School of Medicine at Cardiff University. The primary statistician working on the data will be the lead researcher. Subject to other approvals, other researchers within the project team may also work on the data. For example, this may include work by Dr Lowri Allen, clinician and PhD student at Cardiff University School of Medicine, to quantify the links between parents' and children's diagnosis of diabetes and how these affect child outcomes, and by Alisha Bhanot, a medical student training with Cardiff University, modelling the associations between school exclusions and diabetes health outcomes. All staff accessing data must meet the stringent requirements set out by the Office for National Statistics.

Though real-world person identifiers are shared at the outset of the process, these identifiers are deleted before the data reaches the repository, so it is technically impossible for these identifiers to be made available to any researchers, including the lead researcher.

Since 2023 we have been working with the Health Research Authority and the data providers to try to widen access to this research dataset to researchers beyond Cardiff University. We hope this would allow a broader range of research, to more generate evidence that can generate by a single research team, which can be used to improve policy and practice. The requirements for such access are extremely stringent and will take significant time and resources to resolve, however if we are able to agree the requirements with the data providers and the regulators this document will be updated to include more details of that processing.

Opting out of the study

To opt out of their confidential patient information being used research and planning a participant one can use the National Data Opt-Out. Details of the process is provided on the National Opt-Out information page. A persons' data is still used for clinical direct care purposes and for the diabetes audits purposes but will be deleted from any extracts used for research purposes, including our study.

In theory participants can also directly opt-out of their data being used for research for any of the dataset which we propose to use. These include (i) the National Paediatric Diabetes Audit (ii) The National Diabetes Audit (iii) Higher Education Statistics Agency Dataset (iv) Welsh Government schools data (v) DfE English schools' data. The option to opt-out directly through the relevant data provider is not frequently used, and so the process can be unclear, and may not always be feasible in practice. We advise participants to follow the processes on each data provider’s website, and to contact our project team to help escalate this with data providers if they are struggling.

Opting out of this study must happen prior to the data flows from the data providers to the data processors. This has already occurred for Wales in 2019 and is scheduled for February 2025 for England. Once the initial flows are established, there will hopefully be annual refreshes each November, which may provide an option to retrospectively opt out, but such processing is not currently used, and the approvals are not yet in place to investigate these options further. After the processing takes place, the identifiers dataset will have flowed to the processors to create a linkage ID. Beyond this point, the processors do not retain any identifiable information, so there is no scope for re-identifying and removing people from the study.

The Cardiff University Data Protection Officer (Andrew Lane) is responsible for ensuring that the university complies with data protection laws, including liaising with the Information Commissioner’s Office (ICO) where necessary, and acts as a point of contact for data subjects, including those from this study. He can be contacted by email at inforequest@cardiff.ac.uk or by phone at 029208 74164, or post Compliance and Risk, University Secretary's Office; Cardiff University; McKenzie House; 30-36 Newport Road ; Cardiff; CF24 0DE.

If you wish to report any issues or make a complaint regarding this study that cannot be resolved with the project lead or the Cardiff University Data Protection Officer, you may wish contact the Information Commissioners Office (using live chat on their website or 0303 123 1113), or the Health Research Authority Confidentiality Advisory Group (cag@hra.nhs.uk or 0207 104 8100).