Plan Overview

A Data Management Plan created using DMPonline

Title: CARA

Creator: Akke Vellinga

Principal Investigator: Akke Vellinga

Project Administrator: Sana Parveen

Contributor: Nathaly Garzón-Orjuela

Affiliation: University College Dublin

Funder: Health Research Board (HRB) Ireland

Template: Health Research Board DMP Template

ORCID iD: 0000-0002-6583-4300

Project abstract:

CARA: Collaborate, Analyse, Research, Audit

CARA is a project (https://caranetwork.ie/) set out to develop a data-infrastructure to facilitate GPs to develop a deeper understanding of their patient population, disease management and prescribing through dashboards. The visualisations and comparisons can be used to generate audit reports.

The CARA infrastructure consists of a common data model (to combine data from different Patient Management Software (PMS) systems), CARAconnect to extract data and CARA dashboards for use in Irish general practice. The first exemplar dashboard focused on antibiotic prescribing to develop and showcase the proposed infrastructure, including automated audit reports, filters (within the practice) and between-practice comparisons.

ID: 141664

Start date: 01-01-2021

End date: 31-12-2025

Last modified: 10-01-2024

Grant number / URL: RL-2020-003

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not

imply that the creator(s) endorse, or have any relationship to, your project or proposal

CARA

Data description and collection or re-use of existing data

How will new data be collected or produced and/or how will existing data be re-used?

Data are recorded by GPs during the consultation with their patients with the use of a PMS. Each GP practice will have a server on which they save the database including all PMS tables.

To extract data, upload to the CARA data models and visualise through the dashboards, CARAconnect was created . The data extraction and loading process encompasses selecting relevant data from the practice database, data de-identification, and de-identified data upload to the CARA remote servers. In order to conduct this process in an automatic, structured, and secured way, a desktop application was developed to streamline this task and assist the GP. CARAconnect was envisaged as an easy-to-use application once the practice was registered. The link to download CARAconnect is sent in an email to the practice secure email account (see CARA registration process below) and can easily be downloaded and initiated by a double click. Upon activation, CARAconnect identifies the practice server(s). On the GP server, the database is identified and selected fields from different tables are automatically extracted and securely uploaded to the infrastructure and processed before saving into the new data models during the data transformations steps. At each stage, confirmation by the GP is requested to start extraction and to finalise and upload/send the data.

Through extensive exploration and elimination, a basic understanding of practice databases was developed, which was tested with a reference (anonymous) dataset to finalise specific tables needed to fulfil the data model requirements. CARAconnect facilitates the secure data extraction process based on the variables and tables identified.

What data (for example the kind, formats, and volumes), will be collected or produced?

A set of variables for the initial data extraction was identified, and a data transformation pipeline was created to transform the extracted data into bespoke data models. To this end, practice data was extracted from various tables stored on the practice server and processed before being moved into new data models.

Practices sign up to CARA and will upload data from the previous 5 years. The volume will be determined by the practice size (i.e. number of consultations/GPs).

Coding systems:

The data extracted includes the two main coding systems for classifying disease and therapeutic prescriptions, which are used by general practices in Ireland. The International Classification of Primary Care (ICPC) was designed to capture the interaction between the GP and the patient and is structured around the consultation. ICPC has fewer diagnosis codes than other systems, such as the International Classification of Diseases (ICD-10). PMS include both standard coding systems for coding consultations and diseases. However, GPs are not incentivised to code and few GPs code consultations, making the prevalence of common diseases difficult to measure accurately. Coding in Irish general practice has improved for a few selected conditions since the introduction of the CDM programme in 2020. CDM was integrated into the four accredited PMS systems resulting in the automatic recording of CDM disease codes.

Therapeutic prescriptions are coded within Irish PMS using the Anatomical Therapeutic Chemical (ATC) code, a unique code assigned to any medicine according to the organ or system it works on and its action. The classification system is maintained by the World Health Organisation (WHO). ATC codes have five levels, the highest level is a letter “X” for the main group (e.g. J), the second level is two numbers “##” for the therapeutic group (e.g.01), the third level is a letter “X” for the pharmacological subgroup (e.g. C), the fourth is a letter “X” for the chemical subgroup (e.g. A) and the final, fifth level is the active substance “##” (e.g. 01). This results in ATC codes written as X##XX## (i.e. J01CA01 for the antibiotic ampicillin). This allows easy classification of medicines to specific or larger therapeutic groups.

The data is not fully compatible with the desired output format for analysis and dashboard creation, and data quality, compatibility and usability must be ensured. In order to convert the data into the desired format, bespoke target data models and a number of data transformations were written to populate these models. The bespoke data models represent the main sources of data for the dashboard visualisations and constitute the final product of the data transformations process.

Data transformations include date formatting, field calculation (such as age), mapping to different standards (such as ATC, ICPC, or ICD-10), mapping to created taxonomies (such as Green and Red antibiotics) and data aggregations (such as antibiotics aggregations per consultation).

For the purpose of visualisations, the ICD-10 codes were mapped to the ICPC codes, taking a pragmatic approach and considering their occurrence in general practice. For the exemplar dashboard for antibiotics, the antibiotic ATC code J01 was used and subdivided into classes J01C. Additionally, a second categorisation implemented as part of a national AMS initiative divides antibiotics into green (preferred) and red (non-preferred) antibiotics guidelines, was used.

Data includes:

anonymised patient number	**consultation type**
year of birth	**coded diagnosis**
gender	**medication prescribed**
medical card status	**date of consultations**

Documentation and data quality

What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?

TBC

What data quality control measures will be used?

Question not answered.

Storage and backup during the research process

How will data and metadata be stored and backed up during the research process?

Procedures, data models, code, information etc (no real data), is shared on the CARA GitHub channel.
GP Data will be stored on ICGP servers.

How will data security and protection of sensitive data be taken care of during the research?

The CARA network registration process for GP practices follows a well-defined workflow to ensure secure and efficient onboarding. Initiating the registration, new practices provide necessary details, and the system validates that the email is associated with the closed and secure email service adopted in Irish general practice and hospitals as the primary mechanism for secure communication between health systems (hospitals, laboratories, general practices, pharmacies). Upon entering the registration details, aone-time password (OTP) is generated and sent via email to the provided secure email address. Users are required to enter the OTP within a specified time limit. In case of delay, an option to regenerate a new OTP is available. Following successful OTP validation, users proceed to fill out the registration form, where confirmation of terms and conditions (GP agreement) is a prerequisite before finalising the registration. The GP agreement includes a detailed list of the data that will be extracted, an explanation of the aggregated use of this data for practice comparisons and for research purposes. Afterwards, an email is sent to the secure email address, facilitating the download of the CARAconnect application. The users can subsequently login using their registered credentials, ensuring a seamless and secure experience throughout the CARA network registration and login process. At every subsequent data upload, the GP agreement has to be re-confirmed.
GP can only register with their health mail account. Healthmail is a secure clinical email service that allows health care providers to send and receive clinical patient information in a secure manner. The service is provided by the Primary Care Directorate of the HSE and is managed by eHealth Ireland and supported by the ICGP and the Irish Pharmacy Union.
Healthmail is a closed mailing system. CARA has a healthmail account and can send GPs mails from this account.

Legal and ethical requirements, codes of conduct

If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

CARAconnect extracts de-identified data from the practice database and ensures data security through a practice specific login-based access with 2-factor-authentication. GPs can view their practice data but are only allowed to view aggregated practice data from all other participating practices to avoid possible identification of another practice. Uploads are for a static period to irrevocably de-identify the practice data. Any new data upload overwrites previously uploaded data. Data extraction does not include any patient identifiers or free text, nonetheless, specific technical identifiers needed to link together data in different tables are hashed and extracted. As combinations of specific variables with other external identifiable data sources may potentially lead to identification, two additional processes were applied to facilitate the use of specific technical identifiers to link data in different tables:

- Salted hashing is unidirectional encryption and decryption is almost impossible. Salting introduces an additional random part (a

set of strings of fixed length) to a hash function to create a one-way function. The random part remains the same during extraction but is unique for every extraction.

- k-anonymisation is applied to the data from all other practices, with k set at 5. This guarantees that a minimum of 5 similar patients are included in any comparison (between practice) visualisation. A k-anonymised dataset implies that each record is de-identified from at least k - 1 other.

To accommodate the patients' right to object to the processing of personal data, which provide the option to exclude their data from data processing for research purposes in accordance with General Data Protection Regulation (GDPR) requirements, a data entry field was identified in the PMS to indicate the exclusion of a record.

How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

TBC

What ethical issues and codes of conduct are there, and how will they be taken into account?

Legal basis for processing: Article 89 of GDPR. Section 42s (1)(b); 42 (2) and 42 (3) of the DPA 2018. The processing of Article 9 type data (including health data) requires that a condition in Article 9 must be found; In this case the Art 9 ground will be 'is necessary for reasons of substantial public interest'.

As explicit individual level consent would be impossible to obtain this research will require a consent declaration from the Health Research Consent Declaration Committee (HRCDC).

Data sharing and long-term preservation

How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Data requests from researchers will be considered by the CARA board (CARA team members, ICGP members and GP/patient representative) and a platform, application form and procedure will be designed. To be finalised (and updated) by the end of 2024.

How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

See procedures outlined in application process (CARAnetwork.ie)

What methods or software tools are needed to access and use data?

None

How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

TBC

Data management responsibilities and resources

Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

CARA team led by Prof Akke Vellinga, UCD

What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Ongoing investment