Mexico City Prospective Study (MCPS)
Showcase User Guide: Getting Started

1 Introduction

The Mexico City Prospective Study (MCPS) was established in 1998 and includes data on over 150,000 middle-aged adults. The study is being done in collaboration with the National Autonomous University of Mexico in Mexico City and has received funding from the Mexican Ministry of Health (Secretaria de Salud), the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnologia), the British Heart Foundation, the UK Medical Research Council, and the UK Wellcome Trust.

The Showcase aims to present the data available from the study in a comprehensive and concise way, and to provide relevant additional information for researchers considering applying to use the resource.

This user guide is designed to give you a brief overview of the MCPS data and provides some instructions on how to navigate your way through the system.

Suggestions and information for new users:

- **Read the background information on MCPS and details on data access procedures that can be found on the website (https://www.ctsu.ox.ac.uk/research/prospective-blood-based-study-of-150-000-individuals-in-mexico)**
- **Have a printout of this user guide handy, and take time to familiarise yourself with the Showcase structure, the accompanying documentation and the descriptions provided for each data-field before completing a preliminary application for access to the resource.**

If you encounter problems or faults, please email mcps@ndph.ox.ac.uk

2 Data included in MCPS

2.1 Data collected at the baseline visit

Between 1998 and 2004, the Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU) at the University of Oxford, in collaboration with the Mexican Ministry of Health, established a study in Mexico City, in which over 150,000 middle-aged adults (including 100,000 women and 50,000 men) provided information about their lifestyle and disease history, had physical measurements recorded (including weight, waist and hip circumference, blood pressure) and had a blood sample taken.

2.2 Data collected at the resurvey visit

A resurvey of 10,000 surviving participants (2015 - 2019) captured how lifestyles, physical and biological measurements and treatments for disease (e.g. diabetes) have changed over time. The resurvey also included various 'enhancements' (such as bioimpedance and the collection of a urine sample).

2.3 Linked follow-up data

All participants are now being tracked for mortality through linkage to Mexican national mortality databases; by January 2018, over 20,000 were confirmed to have died.

2.4 Requesting data and future data availability

The Data and Sample Access policy (available from the website, in both English and Spanish) gives full details on how to apply for access to MCPS study data, as well as information on updates and the data release schedule.

You can email us at mcps-access@ndph.ox.ac.uk with any specific queries about data access or about the study in general.

3 Finding data in the Showcase

BROWSE: Use this to navigate your way through hierarchical categories and subcategories of interest to data-fields (i.e. variables) of interest. **This will be the most appropriate tool for most researchers wishing to find and select data for their application to use the Resource.**

CATEGORIES: Related data-fields have been grouped together into categories, such as *Blood sampling data* or *Smoking*. You can find these in the **CATEGORIES** listing in the **CATALOGUES** section (https://datashare.ndph.ox.ac.uk/mexico/cats.cgi).

You can find the rest of the data that you need through the search facility:

SEARCH: The default search is a text search of the data-field name, its notes and data-codings. Entering a numeric value in the search box will return the data-field with that ID.
The **Data-Field Search** facility also allows you to conduct a search using specific criteria based on the type of data-field (see Section 5 for more details).

Text searches of Data-Codings, Categories, Resources, and Datasets can be conducted by selecting the relevant search type button.

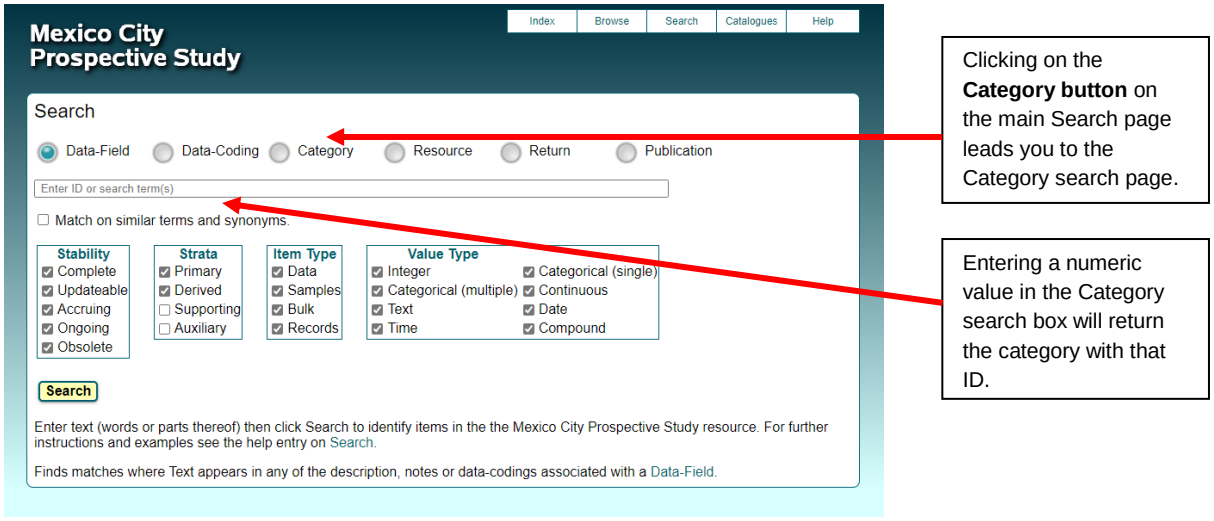

Figure 1. Illustration of the Category Search

Please see the **HELP** tab on the Search page for more details on conducting text searches.

A full list of data-fields, categories and documents can be found on the **CATALOGUES** tab**.**

4 Data categories and sub-categories

Data are organised in a tree structure, accessible via **BROWSE**, with the main categories based on the origin of data collection (Figure 2). These include:

- Baseline assessment
- Mortality follow-up
- Resurvey assessment

Please see the **HELP** page on ‘Browse’ for more details.

The data fields and/or sub-categories within a particular category are revealed by clicking on either the **Category ID** or **Description** in the displayed table (Figure 3).

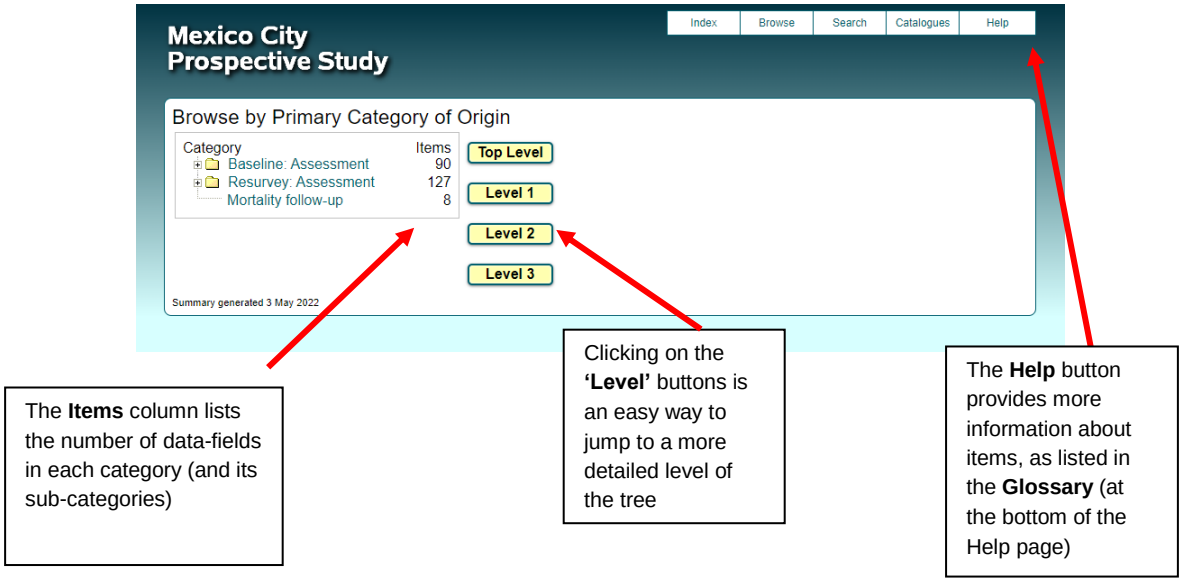

*Figure 2. Illustration of the tree structure via **BROWSE***

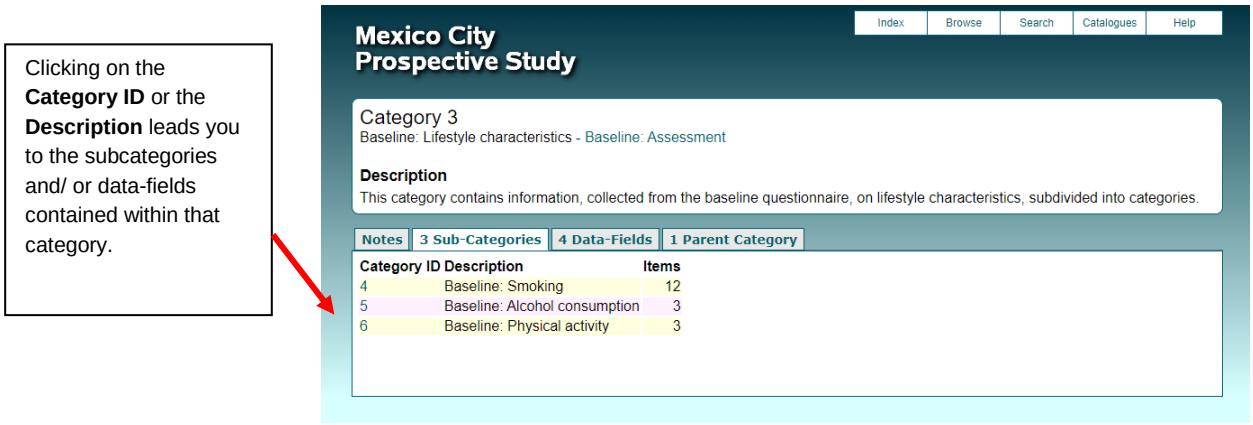

Figure 3. Illustration of sub-categories within the 'Lifestyle Characteristics' category

The tree structure assigns data-fields to one location only, and is not currently cross-referenced. It is therefore important to look in all parts of the tree that might contain data-fields relevant to your research question(s). In general, you should not rely on the **SEARCH** facility to find all fields of relevance for a particular topic.

5 Data-field information

The panel in the top-half of the data-field screen provides a brief description and category location of the data-field within the tree structure (Figure 4). It also includes more detailed technical information about each data-field. This includes information on:

- **Participants**: the number of participants that have the data item
- **Item Count**: the number of data items available

- **Stability**: whether the data-field is complete or changes over time
- **Value type**: the format and units of the data-field
- **Item type**: whether the data-field is a simple data point, relates to an inventory of biological samples, or is a large data object
- **Strata**: the likely relevance to researchers of the data-field
- **Sexed**: whether the data-field is available for both sexes
- **Instances**: how many occasions participants have this measurement performed
- **Array**: whether there are multiple data items for each instance. For example, Figure 4 shows that data on diastolic blood pressure is presented in an array with 3 values per measure (because the measurement was performed three times). Please see the **HELP** page for more details.

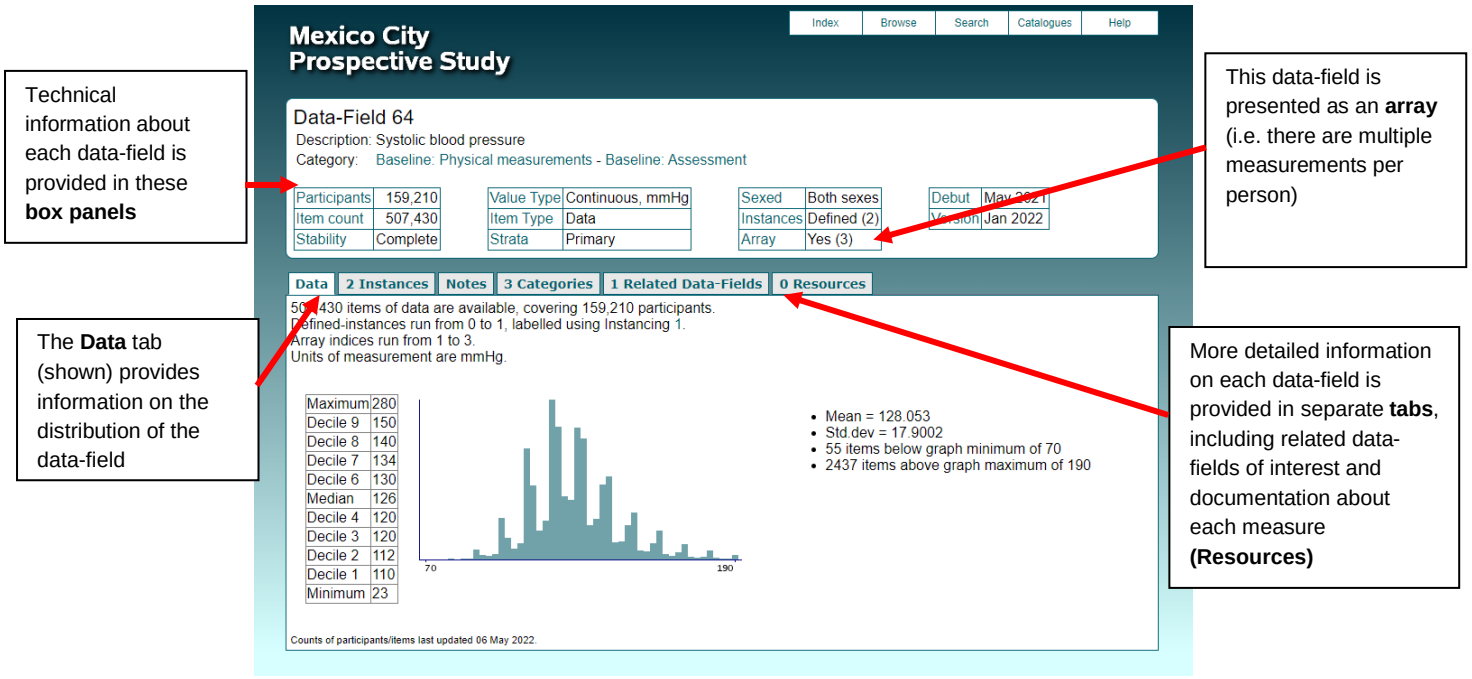

Figure 4. Illustration of a data-field page

The univariate distribution of each data-field is presented in graphical or tabular format (or both) in the **Data** tab (Figure 4). Data are not presented if they are free-text or other unsuitable data items.

The **Instances** tab provides the univariate distribution of each data-field at each instance (e.g. for Data-Field 64 the data are presented separately for the initial baseline assessment (Instance 0), and re-survey assessment (Instance 1).

The **Notes** tab includes the full description of the data-field, together with other details.

The **Categories** tab lists the categories and sub-categories of which the data-field is a member. This is also shown horizontally in the category tree, at the top of the page.

The **Related Fields** tab lists other data-fields to which the current data-field is related. For example, the data-field for ‘diastolic blood pressure ’ (ID: 65) is related to the data-field `Systolic blood pressure’ (ID: 64).

The **Resources** tab contains explanatory documentation related to each data-field, if appropriate.

6 Data cleaning

Data at baseline (1998-2004) was entered directly into handheld devices and at resurvey (2015-2019) into electronic tablets (Samsung Galaxy 10.1). All devices were programmed to take the trained field workers through the questionnaires in the same prescribed manner, automatically skipping questions that were not relevant (for example, questions about obstetric history for male participants). The data recorders queried moderately extreme physical measurements and prevented the input of highly implausible values.

At the end of each working day, the data collected was automatically uploaded to the study's electronic database. The electronic procedures minimised the costs of data collection, processing and checking, while improving data accuracy and completeness. A database-checking program was used to identify possible data errors in the study database, which were regularly reviewed by a data monitoring team and, when necessary, checked by field workers (e.g. by revisiting participants).

Data for the Showcase is provided ‘as is’ from the study’s electronic database; that is, there has been no additional validation, data cleaning or cross-checking between variables, over and above that mentioned above.

Data obtained via linkage (i.e. death records) are subject to validation checks and cleaning. This involves identifying ambiguities in the data, such as invalid clinical classification codes, or mismatches of participants’ records.