Aurum 2022 data issue

Page Owner: [redacted], OR. Last updated: 25/05/23 (latest updates highlighted).

AT SMT REQUEST, ALL EDITS SHOULD BE MADE BY OR QA'D BY [redacted] ALL SMT-APPROVED COMMUNICATIONS OR STANDARD LINES WILL BE CLEARLY FLAGGED.

Context

What were the data quality issues and do we know why these happened?

The first issue identified was random missing observations from the following CPRD Aurum builds: September, October and November 2022. All these builds were based on data received following the migration of [redacted] to [redacted] a cloud-based service. [redacted] updates data for the daily data updates ('daily deltas') one table at a time, but as there was no clear documentation on how this worked, we had been downloading data before all tables had been fully populated. [redacted] will now make documentation on their system processes available to HDS and measures will be put in place so that data are downloaded only when the tables are fully populated. We have come up with a recovery plan does not rely on the daily deltas.

The second issue presented itself as a 'leakage' issue from old builds over time. This should not happen as we need any given build to remain static to allow for reproducible research. This issue was discovered while the investigation into the first issue was ongoing. Essentially, if we re-ran the same extraction using a particular build (say, April 22) after September 22, the re-extraction would yield smaller numbers. We think we know the cause of this issue. When we moved to [redacted] sent a lot of data that they acknowledge was incorrect and advised CPRD to roll back. CPRD did this and the most likely reason for the current leakage could be internal 'cleaning' undertaken by us during this rollback which may have been too vigorous. Now that we understand what has happened, we can rectify this issue. The cause of this issue has not been communicated to clients. For any client that has extracted data against any CPRD Aurum database since September 2022, CPRD recommend that the data should be discarded and re-extracted using the previous CPRD Aurum May 2022 build that was made available on 23/12.

What are we doing to resolve these issues?

We will revert to using the unaffected CPRD Aurum May 22 back-up build which was unaffected by either issue - this is now available for use on the internal server. This build was made available to clients via the online data access portal on 23/12. We have removed the Sep, Oct and Nov 22 builds from the data access portal and already requested clients not to use these. In the revised comms we will clarify that the Sep, Oct and Nov 22 builds cannot be rescued (this has not yet been communicated) and researchers who have started using these builds are advised to either use the unaffected May 22 build or future builds.

We will use the unaffected May 22 CPRD Aurum build as the foundation for our recovery plan. We have cancelled the December 22 and January-March 23 builds. We were aiming to release a March 23 build with updated data to 2023 for 200 new practices plus all existing practices with data up to the point of the May 22 build, but data quality checks run by OR indicated issues with the referral data, which meant this was delayed to the end of April 23.

We will not be using the daily deltas for this process till we have confidence in these. This means that the data in future builds will not be as updated or indeed, updated to the same degree for all practices, during this recovery period. The June 23 build is planned to include data up to 2023 (but

different points in 2023) for all practices. July 23 is likely to be the first build including contemporaneous data for all practices.

March 2023 CPRD Aurum Build

The March 2023 release occurred on 28/04, including 228 new practices with data collected up to the end of January 2023, 1,491 practices with the same data as the May 2022 release, and 1 additional practice with data to March 2022. The output from our source data verification (SDV) checks was made available in the release notes to help researchers decided whether to use this database or not. The inclusion of 228 new practices was communicated to clients on the 06/03. The intention to publish reports from the SDV checks was communicated to clients 03/04.

Future CPRD Aurum Builds

Plans have been drawn up for a June 2023 release including a subset of priority practices. Data quality checks are to be agreed between HDS/OR by 24/04, before prioritised practices start to be downloaded. It has been noted that the timelines don't provide any time for OR to implement checks.

A steer is needed from IR/PCE on which practices to focus on for future builds so we prioritise practices in IR studies. IR have sent a lit of practices wanting to join studies, as well as those already involved, to HDS for review.

SMT are happy for a release cadence of every 2 months to give ourselves breathing space to review and ensure we have robust processes and checks in place. In July/August a paper will be presented by BM outlining the benefits / risks of changing the cadence of releases post-September.

Impact on CPRD Services

How does this impact the CPRD Aurum OMOP CDM plans?

The OMOP build has been generated based on the May 2022 build.

How does this impact IR services?

The mechanisms for running IR queries are now in place, with each study needing to be coded up with their specific patients or inclusion / exclusion criteria. Automation of queries is yet to be implemented by IR has pretty much returned to BAU.

Work on setting up the IR server and establishing the IR feed is almost complete. Due to new missing data issues highlighted by [redacted] in mid-March, the IRSP feed was not switched on as initially planned (IR data continue to be drawn from bulk feeds). There is currently no time-frame for daily delta updates given this new issue. Post-March release, IR and HDS will discuss future support and a working model for IR services, now that IR has decoupled from the data processing database.

Data quality checks have been documented.

How does this impact the Pregnancy Register and Mother Baby Link?

The CPRD Aurum Pregnancy Register was built on unimpacted May 2022 data, but updates are on hold. The Aurum Mother Baby Link was launched in December based on impacted May 2022 data. This means there are fewer matches in the Aurum MBL than there should be, but it is highly unlikely that this would result in incorrect matches. We plan to rerun the May 2022 version of the Aurum MBL on unimpacted May 2022 data once we have capacity within HDS. Complex questions from clients can be forwarded to the Product Owner [redacted] or back-up [redacted]

How does this impact the dataset delivery service?

Six datasets have been identified that have been impacted, five where data had already been delivered, and one that was in progress. All clients have been proactively contacted, and data is being redelivered using the newly restored May 2022 data where possible. Queries can be forwarded to the workstream lead [redacted] or [redacted]

How does this impact the linked data delivery service?

Linked data requests that are using one of the Aurum builds from September 2022, October 2022, or November 2022 should be rejected by the OR Team following the guidance provided by [redacted] (circulated to relevant OR Team members 03/01). If the linked data requests are from Aurum builds prior to September 2022 then it is fine to release the data as normal and we will ask clients to check when the study cohorts were defined and request a redelivery if needed. OR Team members must follow the guidance provided by [redacted] (circulated to relevant OR Team members 03/01) when releasing data or rejecting the request at any point during processing.

How does this impact commissioned research?

Two studies have been impacted: [redacted] and [redacted] deadlines have been extended, and the study is being rerun using the newly restored May 2022 data. [redacted] monthly feasibility reports have been delayed, and the interim study report will need to be run with whatever updated data are available in May 2023.

Plans for Rebuilding CPRD Aurum

How will we know these issues have been resolved in future builds?

HDS/OR/IR have drafted a plan of additional data quality checks to identify any issues. Results of the data quality checks will be added to a centralised document on SharePoint.

What work is ongoing to bring daily deltas back into processing?

No work has commenced as yet, as we are focussing on the bulk downloads to be able to release a new build. Once we are processing the bulk data, we can then start to look at other work we need to do to get the deltas back into processing. However, in mid-March 23, [redacted] notified CPRD of another issue with missing data from daily delta feeds (which is additional to that already known) - a separate work plan may be needed to handle the new issue. There are currently no time-frames for daily delta updates given this new issue and we will need to continue to use bulk data for longer to support builds. HDS hold fortnightly BAU calls with [redacted]

How will a surge in usage be handled?

Whilst there has not previously been a surge in usage when we have released a new Aurum build after a pause in releases, things are expected to be different this time. The real rush is anticipated to be in May/.June when we have more contemporaneous data for patients. HDS are discussing stress tests to establish the level of querying we can cope with.

FAQs

FAQs on the rebuild process for this wiki page are under preparation by HDS/OR/IR and are being coordinated by [redacted]

Plans for 2023/4

Internal Inquiry

CPRD would like to capture lessons learnt from the recent data quality issues to make us audit ready. There will be an MHRA internal enquiry that will be facilitated by [redacted] and [redacted] Members of HDS, OR, and IR may be contacted by a member of the Agency for an interview as part of the inquiry,

Business Planning

The 2023/4 CPRD Business Plan includes the following related activities:

HDS1: Maintain and update existing...datasets

HDS3: Improve data quality and assurance

The 2023/4 MHRA Business Plan will include the following CPRD deliverables:

Establish a data quality working group within CPRD, and commence implementation of external validation checks, to ensure that the data received and disseminated by CPRD meets data quality requirements for research and surveillance [redacted]

Set up a dedicated data quality page on the CPRD website to provide assurance to potential and existing CPRD clients regarding CPRD data quality [redacted]

Draft a data quality strategy including proposals for revised data quality checks for CPRD, and implement revised data quality checks focused on data processing, to ensure that the data received and disseminated by CPRD meets data quality requirements for research and surveillance [redacted]

FADE offer to Clients *(updated 25/05/2023)*

In May 2023 SMT approved that upon client request CPRD can offer the May 2022 FADE to MSL clients with no fee due to the issues with recovering CPRD Aurum continuing into May 2023

Fade will be provided upon request, without RDG or specific approval, for free, for the duration of the MSL licence period. (not for 12months from delivery but it will be linked to the licence period) without the need for any contract amendment.

The current plan is to apply the same policy to the next FADE when CPRD Aurum is fully refreshed later in FY 23/34

Communications and Meetings

What are the next steps for client comms?

We have not provided details to clients of when we would send out the next round of comms.

A data quality page has been added on the CPRD website where the data quality checks can be published, and other key data quality publications can be included. Content will report broadly on the range and type of checks (e.g. structural, data quality, etc.) and at which part of the process these are undertaken.

BM will organise 1:1 drop-in sessions with users to address any concerns.

When will SMT next be meeting to discuss the issues?

05/05.

NB: SMT will continue to attend data quality meetings up to 12/05. Thereafter, SMT will withdraw from weekly meetings but should be copied into notes. [redacted] and [redacted] will continue to support the meetings.

External Communication Sent to Clients (SMT Approved Communications)

Email	Date sent	Mailing list	Content
Customer message_December 2022 CPRD Aurum release_021222.msg	02/12/2022	Nominated Users on Active MSL Contracts | Salesforce	Data anomaly found and CPRD Aurum Dec 2022 build will be delayed
	05/12/2022	Nominated Users on Active MSL Contracts | Salesforce Release Notes Mailing List | Salesforce	CPRD GOLD Dec 2022 build is available and CPRD Aurum Dec 2022 build release is cancelled
Customer message_December 2022 CPRD Aurum release_051222.msg	05/12/2022	Nominated Users on Active MSL Contracts | Salesforce Release Notes Mailing List | Salesforce	CPRD Aurum Dec 2022 build cancelled. Reduction in the number of observations in the CPRD Aurum Sept, Oct, and Nov 2022 builds found. Clients should use the CPRD Aurum May 2022 build
Customer message_CPRD Aurum Recommendations_151222.msg	15/12/2022	Nominated Users on Active MSL Contracts | Salesforce	CPRD Aurum May 2022 build will be available on the data portal, and affected studies with linked data can be re-requested at no charge.
Customer Message_Update CPRD Aurum Recommendations_201222.msg	20/12/2022	Nominated Users on Active MSL Contracts | Salesforce	Data quality issues impacting data extracts after 01/09/2022. Clients to use the newly restored CPRD Aurum May 2022 build, if data was extracted after 01/09/2022. There will be no CPRD Aurum Dec 2022 or Jan 2023 build. (NB: the wording of this message incorrectly refers to a "full data extract" in the potential scenarios section - this should simply read "release")
Customer message_CPRD Aurum Data Quality Issues_170123.msg	17/01/2023	Nominated Users on Active MSL Contracts | Salesforce	Detail about the level of inaccuracy in CPRD Aurum Sept, Oct, and Nov 2022 builds, deletion of data from CPRD Aurum builds, there will be no CPRD Aurum Jan or Feb 2023 builds.
Customer message_Update_CPRD Aurum Data Quality Issues_010223.msg	01/02/2023	Nominated Users on Active MSL Contracts | Salesforce Contact Relationship Roles for Accounts | Salesforce Main Contacts from MSL organisations only	Detail what the March 2023 CPRD Aurum build will contain, and the quality checks that will now be performed on the data.
Customer message_April 2023 CPRD GOLD release and CPRD Aurum update_030423.msg	03/04/2023	Nominated Users on Active MSL Contracts | Salesforce Release Notes Mailing List | Salesforce	CPRD GOLD Apr 2023 build release, and CPRD Aurum Mar 2023 build release would be delayed to complete further data verification checks from [redacted]
Customer message_Update CPRD Aurum data quality_190423.msg	19/04/2023	Nominated Users on Active MSL Contracts | Salesforce Contact Relationship Roles for Accounts | Salesforce Main Contacts and Research roles from MSL organisations only	Table showing [redacted] data verification checks, CPRD Aurum Mar 2023 build would be released soon.
Customer message_March 2023 CPRD Aurum release_280423 .msg	28/04/2023	Nominated Users on Active MSL Contracts | Salesforce Release Notes Mailing List | Salesforce	March 2023 CPRD Aurum build is released with limited temporal coverage
Customer message_ CPRD Aurum update_260723 .msg	26/07/2023	Nominated Users on Active MSL Contracts | Salesforce Contact Relationship Roles for Accounts | Salesforce Main Contacts roles from MSL organisations only	Next CPRD Aurum release would not be 31 July 2023 as previously advised, will be delayed.

Handling & Tagging Client Enquiries

Queries received and recorded in Salesforce (SF) have been tagged with the keyword "Aurum 2022 data issues" and are shown in this report Aurum 2022 data issues Queries all time | Salesforce. CPRD colleagues who receive direct communication from clients about this are advised to create a query in SF so we can record the conversations and enable visibility across CPRD teams.