



USER MANUAL FOR RESEARCHERS

Version 1.0

Developed by  IGCPHARMA

Table of Contents

A structured guide to completing a full aha™ harmonization session.

01. Part 1 - Introduction

- 1.1 What is aha™?
- 1.2 Core concepts

02. Part 2 - Getting Started

- 2.1 Uploading your datasets
- 2.2 The four-module pipeline
- 2.3 Navigation

03. Part 3 - Data Profiling Module

- 3.1 Overview
- 3.2 Switching between datasets
- 3.3 The variable table
- 3.4 Inspecting variable values
- 3.5 Searching and sorting

04. Part 4 - Strategy & Planning Module

- 4.1 Overview
- 4.2 Decision types
- 4.3 Plan confidence
- 4.4 Setting the human intervention threshold
- 4.5 The plan table
- 4.6 Reviewing variables
- 4.7 The review panel

05. Part 5 - Execution Module

- 5.1 Overview
- 5.2 The side-by-side preview
- 5.3 Proceeding to validation

06. Part 6 - Validation Module

- 6.1 Overview
- 6.2 The five validation checks
- 6.3 The validation table
- 6.4 Fixing failed variables
- 6.5 Pipeline processing time
- 6.6 Finalize the harmonization

07. Part 7 - Reference

- 7.1 Glossary
- 7.2 Frequently asked questions



Introduction

Understand the purpose, structure, and language of the platform.

What is aha™?

aha™ (Agentic Harmonization Assistant) is an AI-powered tool designed to help researchers harmonize clinical and observational datasets to a defined standard schema. It automates the most time-consuming parts of data harmonization - profiling variables, proposing mappings, suggesting transformations, and validating outputs - while keeping the researcher in control of every decision.

aha™ is structured as a four-step pipeline. Each step builds on the last, and the researcher can review, adjust, and re-run any step before proceeding. The goal is to produce a harmonized dataset that is semantically aligned, structurally consistent, and ready for analysis.

1.2

Core concepts

Before using aha™, it is useful to understand the following terms and concepts that appear throughout the tool.

- **Unharmonized dataset**

The source dataset uploaded by the researcher. This is the raw, study-specific data that will be transformed to align with the standard.

- **Standard dataset**

The reference schema that the unharmonized data will be mapped to. This defines the expected variable names, data types, coded values, and structure that the output must conform to.

- **Codebook**

A codebook (or data dictionary) is a reference document accompanying a dataset that describes its structure and contents, including variable names, labels, definitions, coding schemes, units of measurement, and handling of missing values. It may also include contextual information such as data collection methods, sampling procedures, and summary statistics.

- **Plan confidence**

A percentage score that reflects how certain aha™ is about its proposed harmonization plan for a given variable.

- **Human intervention threshold**

A confidence cutoff set by the researcher. Variables with a plan confidence score below this threshold are flagged for human review before the pipeline proceeds.

- **Harmonization**

The process of transforming the unharmonized dataset so that its variables, values, and structure match the standard. This may involve renaming variables, recoding values, converting data types, or deriving new variables from existing ones.



Getting Started

Upload the right files and move through the four-module pipeline.

2.1

Uploading your datasets

To begin a harmonization session, upload four files:

- 1 Your unharmonized source dataset (the study-specific data to be transformed).
- 2 The standard dataset that defines the target schema.
- 3 The data dictionary from the unharmonized source dataset.
- 4 The data dictionary from the standard dataset.

aha™ only accepts CSV formats for the data and PDF for the data dictionary. Once all files are uploaded, click Next, and the system will begin processing automatically and transition to the Data Profiling Module.



Before you begin

Make sure your datasets and data dictionaries are complete and correctly formatted before uploading. aha™ profiles variables as they appear in the file, missing column descriptions may affect harmonization results.

2.2

The four-module pipeline

aha™ guides the researcher through four sequential modules. Each must be completed before the next begins.

Module	Purpose
1. Data Profiling	aha™ extracts variables, data types, missing values, and coded value details for both datasets.
2. Strategy & Planning	Review and edit aha™'s harmonization plan, including decisions, target mappings, and transformations.
3. Execution	aha™ applies the approved plan and displays a side-by-side preview of the original and harmonized datasets.
4. Validation	aha™ runs automatic quality checks on every harmonized variable and flags any that do not meet the standard.

2.3

Navigation

Each module includes Back and Next buttons to move between steps. Changes made in earlier modules can affect later ones - if you return to Strategy & Planning from Validation, the pipeline will re-run from that point.



Data Profiling Module

Review variable structure, data types,
missingness, and coded values.



The Data Profiling Module

is the first step of the pipeline. It gives the researcher a structured view of every variable in both the unharmonized and standard datasets before any harmonization begins.

3.1

Overview

For each variable in each dataset, aha™ automatically generates a description, identifies the data type, calculates the percentage of missing values, and gives the encoding details when applicable. The researcher can use this information to understand the content and completeness of their data before proceeding.

3.2

Switching between datasets

The module displays one dataset at a time. Use the tabs at the top of the table to switch between the unharmonized source dataset and the standard dataset. Each has its own set of variables and profiling details.

3.3

The variable table

Each row in the table represents one variable. The table includes the following columns:

Column	Description
Variable Name	The name of the variable as it appears in the dataset file.
Description	An extracted description from the data dictionaries of what the variable represents.
Data Type	The detected data type of the variable (e.g., categorical_nominal, binary, text, numeric).
% Missing	The proportion of rows in which this variable has no value.
Details	An icon that opens a panel showing the specific coded values for the variable, including any special missing value codes.

3.4

Inspecting variable values

Click the eye icon in the Details column of any row to open a side panel. This panel shows the key-value pairs for that variable (i.e., the numeric codes and their corresponding labels) as well as any codes used to represent missing data.



Note:

This detail view is particularly useful for categorical and binary variables where the numeric codes are not self-explanatory.

3.5

Searching and sorting

Use the search bar at the top of the table to filter variables by name. The sort icon next to the search bar lets you sort by missing rate.



Strategy & Planning Module

Evaluate the AI-generated harmonization plan
before execution.



The Strategy & Planning Module

is the main step of the pipeline. aha™ presents its proposed harmonization plan for every variable, and the researcher reviews, confirms, or edits each proposal before execution.

4.1

Overview

For each source variable, aha™ proposes a decision, a target standard variable, and a sequence of transformations. Each proposal is accompanied by a plan confidence score. The researcher sets a threshold to determine which variables require human review before the plan can be approved.

4.2

Decision types

Every variable in the harmonization plan is assigned one of five decision types. These reflect how the source variable relates to the standard.

Decision	Definition
Matched	The source variable is identical to the standard variable. No transformation is needed; the variable is used as-is.
Mapped	The source variable directly corresponds to a standard variable with consistent meaning. Full consistency is achieved after applying transformations.
Partially Mapped	The source and standard variables capture the same underlying concept but differ in operationalization. A partial relationship exists but full alignment is not achievable.
Calculated	The standard variable does not exist explicitly in the source data but can be derived from one or more existing source variables using a formula or logic.
Not Mapped	The source variable cannot be aligned with any standard variable, either directly, partially, or through calculation.

To see the definition of any decision type while working in the module, expand the Decision guide button in the toolbar or hover over a decision badge in the table; a tooltip with the definition will appear.

4.3

Plan confidence

Each row in the plan table includes a Plan Confidence score. This is aha™'s overall certainty that its proposed plan for a variable is correct.

4.4

Setting the human intervention threshold

At the top of the module, a color-coded slider allows the researcher to set the confidence threshold below which variables will be flagged for human review. The recommended default is 60%.

- 1 Variables with plan confidence at or above the threshold: Human intervention marked as Not needed.
- 2 Variables with plan confidence below the threshold: Human intervention marked as Needed.



Note:

The threshold can be reset to the recommended value at any time using the Reset to recommended button below the slider.

4.5

The plan table

The main table lists every source variable and its proposed harmonization plan. The columns are:

Column	Description
Source	The variable name in the unharmonized dataset.
Target	The corresponding variable name in the standard dataset. Empty for Not Mapped variables.
Decision	The proposed decision type (see Section 4.2).
Plan Confidence	The overall confidence score for the proposed plan (see Section 4.3).
Human Intervention	Whether the variable requires review before proceeding, based on the threshold setting.
Review	A button that opens the individual variable review panel.

Use the sort button next to the search bar to order variables by confidence, and the filter button to show only variables with a specic decision type.

4.6

Reviewing variables

Variables can be reviewed individually by clicking the Review button on any row, or in bulk using the buttons at the top of the module:

- 1 Review all: Opens the review panel starting with the first variable and steps through all of them in sequence.
- 2 Review only those that require human intervention: Opens the review panel filtered to only the variables flagged for human review.

4.7

The review panel

When reviewing a variable, the panel shows the source-to-target mapping at the top (e.g., RID → NACCID) and a progress indicator showing how many variables have been reviewed. The panel is divided into three sections.

- **4.7.1 Source & Target Info**

To understand the source and target variables in more detail before deciding, expand the Source & Target Info panel. This shows each variable's data type, number of unique values, missing rate, and a plain-language explanation of how the two variables relate semantically.

- **4.7.2 Decision**

This section shows the five possible decision types, each with its individual confidence score. aha™'s recommended decision is highlighted and labeled as aha™'s pick. The researcher can select a different decision type if appropriate.

- **4.7.3 Target**

This section rank candidate standard variables. The aha™'s recommended target is shown at the top and labeled aha™'s pick. The researcher can select a different target if needed. Click the information icon next to any candidate to open a detail panel showing its description, coded values, data type, and other relevant metadata.

**Note:**

Additional candidates beyond the top four are available by scrolling down in the target list.

- 4.7.5 Transform

Expand the three-dot menu below any transformation option to read the full reasoning behind it. To confirm a transformation, select it using the Selected button.

If none of the suggested transformations are appropriate, use the Ask aha™ chat for another transformation option. Describe the required transformation in plain language, and aha™ will generate a new option.

Once all three sections are complete, use the Save & Next Variable button to save the review and move to the next variable.

- 4.7.4.1 Transformation types

Each transformation sequence is built from one or more of the following transformation types. Understanding what each one does helps the researcher evaluate whether the suggested plan is appropriate.

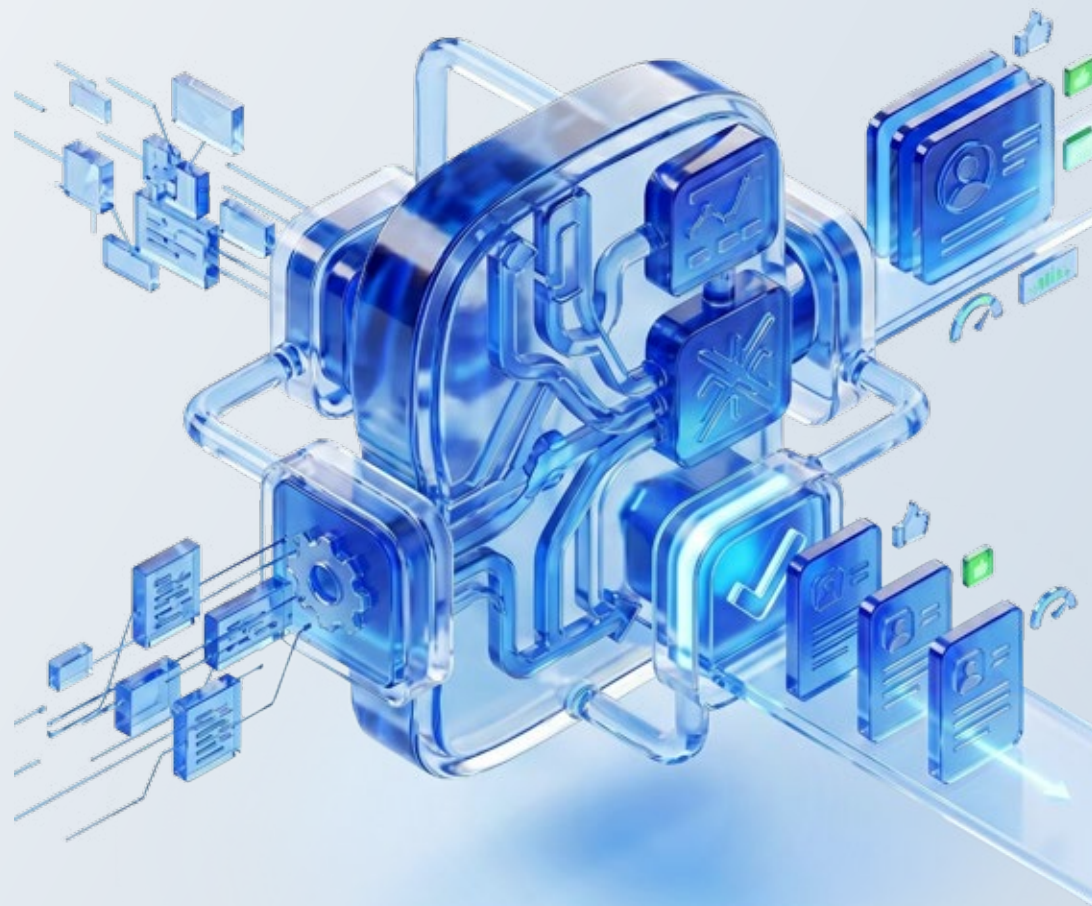
Transformation	What it does	Example
Rename	Changes a column's name without modifying its values.	RID → NACCID
Map	Replaces specific values in a column with new ones according to a defined mapping.	1 / 2 → Male / Female
Add Suffix	Appends a tag to a column's name. Used when a variable only partially matches the target and the name must signal the approximation.	age → age_approx
Add New Category	Introduces a new permitted value to a standard column that does not yet include it.	Adding Unknown to a column that previously only had Yes / No
Trim	Caps numeric values to a defined range. Values below the minimum become the minimum; values above the maximum become the maximum.	Ages above 90 are set to 90

Transformation	What it does	Example
Count	Creates a value by counting how many columns satisfy a given condition.	Number of cognitive tests a participant completed
Case When	Sets a column's value based on conditional logic applied to the same or other columns.	If score < 24 then Impaired, else Normal
Convert to Null	Replaces values that meet a condition with a missing value (null).	Placeholder -99 becomes blank
Formula	Computes a new column value from a mathematical expression using one or more existing columns.	BMI = weight / height ²
Convert Data Type	Changes how values are stored without altering their content.	Text "5" → integer 5
Convert Numerical to Categorical	Groups a continuous numeric range into discrete labeled bins.	Age → <65 / 65-80 / >80
Assign	Fills an entire column with a single fixed value for all rows.	Set source_study = "AIBL" for every row
Normalize	Cleans and standardizes text so that values are consistent and comparable regardless of original formatting.	" Male", "MALE", "male" → male
Split	Divides text on a defined separator and retains the selected part.	"1990-05" split on -, keep year → 1990
Group By Cumulative Count Unique	Orders rows within a group and assigns each a sequential number, resetting for each new group.	Patient visits numbered 1, 2, 3... in date order



Execution Module

Apply approved mappings and preview the harmonized output.



The Execution Module

applies the approved harmonization plan to the source dataset and produces the harmonized output. No input is required from the researcher at this stage.

5.1

Overview

Once the researcher approves the plan in the Strategy & Planning Module, aha™ executes every decision, target mapping, and transformation automatically. The process typically takes a couple of minutes.

5.2

The side-by-side preview

When execution is complete, the module displays two tables side by side:

- 1 Preview Unharmonized Dataset: The original source data with its original variable names and values.
- 2 Preview Harmonized Dataset: The transformed output with renamed variables, recoded values, and converted data types applied.

5.3

Proceeding to validation

Click Next to proceed to the Validation Module, where aha™ will run automatic quality checks on the harmonized output. If something in the preview looks incorrect, click Back to return to the Strategy & Planning Module and adjust the plan before re-running.



Validation Module

Check outputs against the standard and resolve failed variables.



The Validation Module

is the final step of the pipeline. aha™ runs five automatic quality checks on every harmonized variable and presents the results for the researcher to review.

6.1

Overview

At the top of the module, a summary panel shows key statistics for the current harmonization session:

Statistic	Description
Total Variables	The total number of variables in the unharmonized dataset.
Harmonized	The number of variables that passed all five validation checks.
Not Harmonized	The number of variables that failed one or more checks and require attention.
Not Mapped	The number of variables for which no standard mapping was found.
Harmonization Rate	The percentage of variables that were successfully harmonized.

6.2

The five validation checks

Each variable in the harmonized output is evaluated against the following five criteria. A variable must pass all five to be marked as harmonized.

#	Check	What aha™ evaluates
1	Concept Alignment	Do source and standard measure the same thing at the same semantic level and units? This check flags mismatches such as age vs. age group, subject ID vs. visit ID, or categorical vs. continuous concepts.
2	Data Type Compatibility	Are source and standard data types compatible? If not, does the harmonization plan define a valid conversion (e.g., numeric to categorical, datetime to year, string to coded category)?
3	Value / Category Coverage	Do mapped values cover all required standard categories, including edge cases such as multiracial, unknown, or consolidated groups? Any standard categories that cannot be produced are explicitly flagged.
4	Missing & Placeholders	Are all placeholder codes and missing value handling strategies explicitly accounted for? This includes patterns such as sentinel values, nulls, and recoding strategies.
5	Stats Sanity Check	Do output distributions and value counts look statistically reasonable? For numeric variables this covers ranges and means; for categorical variables it covers unique values and counts.

6.3

The validation table

The main table lists every harmonized variable and its validation result. The Harmonized column shows TRUE for variables that passed all checks and FALSE for those that did not. The remaining columns show a brief explanation for each of the five checks, with a View more link to read the full detail.



Note:

Use the sort button at the top of the table to bring failed variables (FALSE) to the top for easier prioritization.

6.4

Fixing failed variables

If one or more variables failed validation, use the “Review not harmonized variables” button at the bottom of the page. This returns the researcher to the Strategy & Planning Module with the variable list pre-filtered to show only those that need attention. Once the plan is updated, the pipeline re-runs automatically from that point.

6.5

Pipeline processing time

The Pipeline Processing Time panel, accessible by expanding the button of the same name, shows how many minutes were spent in each of the four modules during the current session. This is provided for transparency and can be useful for estimating session time in future harmonization projects.

6.6

Finalize the harmonization

Below the validation table a “Finalize & Save” button allows to finish the harmonization once you are satisfied with the process. The final harmonized CSV will be saved to your workspace, and you will be redirected to the main menu. Changes in the harmonization can be made later by resuming the run in the “Run History” panel.



Reference

Use the glossary and FAQ to answer common workflow questions.

Glossary

Term	Definition
Codebook - Data Dictionary	A structured document or file that describes the variables in a dataset. Including their names, definitions, data types, coded values, and any special codes used for missing or sentinel values. aha™ uses codebook information to understand the datasets and inform the harmonization plan.
Confidence score	A percentage reflecting how certain aha™ is about a specific aspect of its proposed harmonization plan. Sub-scores exist for decision type, target selection, and transformation.
Data type	The format in which a variable's values are stored. Common types include categorical_nominal, binary, numeric, text, and datetime.
Decision type	One of five categories assigned to each variable: Matched, Mapped, Partially Mapped, Calculated, or Not Mapped. Describes how the source variable relates to the standard.
Harmonization	The process of transforming a source dataset so its variables, values, and structure conform to a defined standard schema.
Harmonization rate	The percentage of source variables that passed all five automatic validation checks after execution.

Term	Definition
Human intervention	A flag indicating that a variable’s plan confidence is below the researcher-defined threshold and requires manual review before proceeding.
Missing value code	A numeric or symbolic code used in the source dataset to represent a missing or unknown value (e.g., -4, 88, 99).
Plan confidence	The overall AI confidence score for a variable’s harmonization plan, calculated as the average of the decision, target, and transform confidence sub-scores.
Sentinel value	A specific code in a dataset used to indicate a special condition, such as not applicable or not assessed, rather than a true missing value.
Standard dataset	The reference schema that defines the expected variable names, data types, and coded values that the harmonized output must conform to.
Threshold	The plan confidence cutoff below which variables are flagged for mandatory human review. Set by the researcher in the Strategy & Planning Module.
Transformation	A sequence of operations applied to a source variable to convert it into the format required by the standard target variable.
Unharmonized dataset	The raw, study-specific source dataset uploaded by the researcher before any harmonization has been applied.
Validation check	One of five criteria evaluated by aha™ in the Validation Module: concept alignment, data type compatibility, value/category coverage, missing & placeholders, and stats sanity check.

7.2

Frequently asked questions

- **Can I upload more than two datasets?**

Each aha™ session processes one unharmonized dataset and one standard dataset. To harmonize multiple source datasets, run a separate session for each.

- **What happens if I change a decision in the review panel?**

Changing the decision type for a variable may affect the available target options and transformation suggestions. aha™ will update the recommendations accordingly. The plan is not finalized until you save the review and proceed to the next module.

- **Can I skip variables I do not want to harmonize?**

Variables assigned the Not Mapped decision type are excluded from the harmonized output. If you do not want to include a variable in the harmonized dataset, set its decision to Not Mapped in the review panel.

Frequently asked questions

- **What should I do if none of the suggested transformations are correct?**

Use the Ask aha™ chat to generate other transformation, in the Transform section in the review panel. Describe the required transformation in plain language - for example, 'recode value 1 to 0 and value 2 to 1' - and aha™ will generate a new transformation option based on your description.

- **Can I return to a previous module after completing it?**

Yes. Use the Back button in any module to return to the previous step. If you make changes in the Strategy & Planning Module, the pipeline will re-execute and re-validate automatically when you proceed forward again.

- **What does it mean if a variable shows FALSE in the validation table?**

A FALSE result means the variable failed one or more of the five validation checks. The columns in the validation table show which checks failed and why. Use the Review not harmonized variables button to return to the Strategy & Planning Module and address the issues.



Developed by  **IGCPHARMA**

www.igcpharma.com