Commitment data cleaning and standardisation *(archived webpage)*

AUTHORS
Giacomo Zanello (lead author), Renata Micha

This archived webpage describes the process for commitment data cleaning and standardisation established prior to the Tokyo Nutrition for Growth (N4G) Summit in 2021 and has since being updated.

Introduction

This page describes the actions taken to retrieve, clean and standardise the nutrition commitments submitted through the Nutrition Accountability Framework (NAF).

In summary, the data is retrieved from the NAF Platform using an application programming interface (API) and then undergoes a semi-automated data cleaning and standardisation process to create a common format across all commitments. This includes assessing the eligibility of all commitments registered with the NAF and ensuring consistency of data types throughout the dataset. Commitments are also classified using the Nutrition Action Classification System and assessed for SMARTness using the SMARTness score.

No corrections are made to the self-reported commitments at this stage. Any standardisation that requires amending, regrouping or reclassifying results in the generation of a new variable and not the correction of an existing one. The commitment-verification process will seek clarity on commitments directly from stakeholders and commitments will be corrected and/or updated only with the approval of commitment makers. The data-standardisation process is independent of but complementary to the commitment-verification process.

Key terms used on this page

Commitment goal: A commitment goal is what stakeholders are committing to achieve and is used to track and assess progress made towards the commitment. Commitments goals must be measurable and should be nutrition-related, including nutrition-specific and nutrition-sensitive goals.

Nutrition Action Classification System: A taxonomy to name, define and classify nutrition actions based on common principles and shared characteristics, as described in the online resource The Nutrition Action Classification System.

Nutrition Commitment: The intent and pledge to address poor diets and malnutrition in all its forms through SMART nutrition actions.

SMARTness score: A novel scoring system for assessing the SMARTness of commitments registered through the Nutrition Accountability Framework (NAF). The five dimensions of SMARTness are defined in the online resource The SMARTness of nutrition commitments and the method of calculating the SMARTness score is described in Assessing the SMARTness score of nutrition commitments.

Commitment retrieval and eligibility

The data held within the NAF Platform is retrieved using an Application Programming Interface (or API) that enables all commitment data to be downloaded in a comma-separated values (CSV) file. The API automatically combines the information provided in the Sign up Form and the Commitment Registration Form, so each commitment is linked to the user information of the commitment maker.

Additional commitments were registered prior to the NAF Platform being launched, using either a Google Form or a PDF provided by the Global Nutrition Report (GNR). This data was manually mapped to the NAF Commitment Registration Form by the GNR team. Due to differences in the structure of the two registration forms, information was missing from the commitments registered via the Google Form/PDF. Before combining these commitments with the main dataset, each stakeholder was individually contacted by the GNR team and asked to supply the remaining information, either by registering their commitment on the NAF Platform or by sending it to the GNR team. The data from stakeholders who elected to send the additional information to the GNR team was then combined with the dataset retrieved from the API. The dataset was therefore restructured so the information provided in the Sign up Form was at the start of the spreadsheet followed by the commitment-specific responses in the same order as the Commitment Registration Form.

Commitments are grouped at commitment level (i.e., one row per commitment) in the dataset. Each one is automatically given a unique random 19-digit identification number when it is registered with the NAF Platform. To aid goal-level assessments (described in Section 4), a dataset with one row per goal is also created. This contains an additional unique identification number for each goal, consisting of the 19-digit identifier followed by sequential numbers separated by an underscore ('_'), where 0 represents the first goal and 9 represents the tenth goal, e.g., [19-digit identifier]_0.The application of SMART criteria in the formulation of nutrition commitments ensures that the type of the commitments (such as enabling, policy or impact), their goals and their expected outcomes are clear. Making commitments easier to classify and monitor also makes it possible to measure impact and demonstrate success. To facilitate the formulation and assessment of SMART commitment goals, we have identified a set of commitment ingredients (defined as the individual characteristics that describe each of the SMART dimensions); these have been mapped to

each of the five SMART dimensions, as detailed in the Appendix. These ingredients form the basis of the NAF Platform's Commitment Registration Form and are their assessments are at the base of the SMARTness score.

Commitment screening and inclusion

The combined dataset undergoes an eligibility assessment that removes all test submissions, all commitments that are not deemed eligible to be included in the NAF, and all duplicate commitments. Commitments are identified as potential duplicates manually by the GNR team when it appears that information in two commitments is either identical or too similar for the commitments to be independent of each other. The GNR team then contact the stakeholder to confirm the presence of a duplicate commitment and, if confirmed, the commitment specified by the stakeholder is removed.

Only commitments with nutrition-related goals are included in the NAF. Commitment goals that are not nutrition-related (e.g., increasing physical activity) or the information provided is not sufficient or clear enough to decide whether they are nutrition-related (e.g., increasing the number of beneficiaries receiving cash transfers to address poverty) are flagged and part of the verification process.

Data cleaning

Data is checked to ensure consistency in data type within variables. All numerical values are changed to float numbers (i.e., numbers with a decimal place). In cases where it is easily identified that a stakeholder used a comma as a decimal separator, such as for prevalence values, the comma is changed to a decimal point. If there is any ambiguity over the purpose of a comma, the value remains as originally submitted. The verification process will clarify with the stakeholder whether the comma has been used as a decimal separator and if confirmation is received, these commas will be changed to a decimal point.

The start month and start year and end month and end year (GX.7 of the Commitment Registration Form where X refers to the goal number) are converted into date format (DD/MM/YYYY) to create one start date variable and one end date variable. The first day of the chosen month is used for the start date and the final day of the chosen month is used for the end date.

Multiple choice answers are converted into binary variable columns for convenience and to simplify analysis. For example, an answer to the N4G thematic area question (Q9 of the Commitment Registration Form) selecting both 'Health' and 'Food' is converted into two binary columns in the dataset where Health=1 and Food=1. Country names across the dataset are standardised to be consistent with the spelling used within the GNR (e.g., 'USA' to 'United States of America').

Missing values are checked. Compulsory questions in the commitment registration form have no true missingness, however values are considered missing for responses such as 'NA'

('not applicable'), 'TBD' ('to be determined'), 'TBC' ('to be confirmed') or any text indicating that 'this has not been decided yet'.

Finally, the language of each commitment is identified. While commitments must be submitted in English, occasional non-English commitments have been registered with the NAF. Non-English commitments are translated into English before being analysed.

Nutrition Action Classification and SMARTness score

Using the goal-level version of the dataset, all measurable goals for eligible commitments are classified in the Nutrition Action Classification System and assessed for SMARTness score (see Assessing the SMARTness score of nutrition commitments). Commitment goals that are not explicitly nutrition-related (e.g., increasing physical activity) or that do not contain sufficient or clear enough information to decide if they are nutrition-related (e.g., increasing the number of beneficiaries receiving cash transfers to address poverty) are classified and assessed for SMARTness. The connection of these goals to nutrition can then be determined with stakeholders in the verification process.

The SMARTness assessments of each ingredient are combined to calculate a goal-level commitment SMARTness score together with a Nutrition Action SMARTness Index.

Data standardisation

Due to the extent of the variation in self-reported responses, the commitment dataset requires a degree of standardisation before any data can be analysed. Standardisation enables comparison between commitments while allowing the substance [LC1] [GZ2] of the commitment to remain the same. This section describes the standardisation of organisation name and stakeholder, funding information, indicator name and geographical action area.

Organisation name and stakeholder type

Organisation names for both the organisation making the commitment (from the Sign up Form) and any additional organisations involved (Q5 on the Commitment Registration Form) are checked for consistency in names and spelling, and abbreviations are expanded, where known. For example, 'WHO' becomes 'World Health Organization' for all instances.

Stakeholder type is also checked to ensure that the correct type was selected when the stakeholder signed up to the NAF Platform. Where stakeholders selected 'Other' as stakeholder type (in the Sign up Form), the free text answers were assessed to find the best fit out of the other stakeholder types.

All expansions of acronyms and reclassifications of stakeholder type will be checked with stakeholders through the verification process and corrections to the dataset will be made if required.

Funding information

Funder, funding mechanism and amount of secured funding (Q11 on the Commitment Registration Form) are standardised by regrouping. These groupings, along with descriptions of the types of answers in that group, can be found in Table 1, Table 2, and Table 3 respectively. The groupings for funder and funding mechanism are standardised as binary values, as they are not mutually exclusive. For example, one commitment could have multiple funders and be classified under 'Government', 'Global organisation' and 'Mixed'. The amount of secured funding is mutually exclusive, and stakeholders cannot be classified into more than one of the groups. The total cost of the commitment (Q10a on the Commitment Registration Form), where provided, is standardised by converting the cost from the specified currency into US$ using the World Bank exchange rate at the time.[1]

Table 1: Standardisation groups for funder

Funder group	Description
Governments	National governments or ministries
Global organisations	UN, WHO, UNICEF, Gates Foundation, WFP, World Bank, Food Foundation, etc
Donor organisations	Donor organisations, partners, development banks and institutions, EU, and relevant offices
Businesses	Business such as Google, Griffith Foods, Quorn, Cargill, etc
Aid organisations	National aid organisations such as USAID, Irish Aid, Nutrition International, Eat well programme, etc
Private sector	Family institutions, individuals, grands, business stakeholders and funds
SUN	Scaling Up Nutrition networks

Funder group	Description
Mixed	Any combination of two or more of the other groups
Unknown	Funder described as 'unknown', 'TBD', 'TBC', or text indicating that this has not been decided yet; funder described as 'not applicable'; funder described as 'no'

Table 2: Standardisation groups for funding mechanism

Funding mechanism group	Description
Public	The answers name public, public budget, public agency, organisation or institution
Private	Private sector, private fund, individual
Governments	When a national government or government ministry is mentioned
Donors	Donor organisations, global organisations, WHO, UN, UNICEF, etc
Self	Self-funded
Mixed	Any combination of two or more of the other groups
Unknown	Funding mechanism described as 'unknown', 'TBD', 'TBC', or text indicating that this has not been decided yet; funding mechanism described as 'not applicable'; funding mechanism described as 'none'

Table 3: Standardisation groups for amount of secured funding

Amount secured group	Description
Full	'Fully funded', '100%', 'secured', all and any other text indicating the commitment is fully funded.
High	Any amount, number or word indicating between 51–99%
Half	Any amount, number or word indicating 50%
Low	Any amount, number, word indicating between 1–49%; any amount, number, or word indicating no funding has been secured
Partial	No numerical estimate provided other than words indicating funding is partially secured
Unknown	Amount secured described as 'unknown', 'TBD', 'TBC', or text indicating this has not been decided yet; amount secured described as 'not applicable'; amount secured described as 'no'

Indicator name

Indicator names (GX.8a on Commitment Registration Form where X refers to the goal number) are grouped and standardised. These groups are described in Table 4 and are mutually exclusive.

The baseline level (GX.8b) and target level (G1.8d) of goals categorised as 'Impact' commitments in the classification system are standardised to remove the unit of measurement (e.g., %), resulting in only numeric data that remains in the same unit as the original value. For example, a prevalence value of '8%', which is presented as a whole number becomes '8.0' in the standardised version of this variable. This standardisation allows for the calculation of intended change between baseline and target values. Impact commitments result in changes in population and the types of indicators and units are repeated, allowing for comparisons to be made. Goals classified as 'Enabling' and 'Policy' commitments have much greater breadth of variation and are standardised in a similar manner during the verification process.

Table 4: Standardisation groups for indicator names

Indicator name group	Description
Beneficiaries	Number of people benefited, percentage of people benefited, decrease or increase in the number of people, specific population groups. Example: Number of children treated for malnutrition, coverage of children with iron supplementation.
External	Already established metrics, evaluated against some standard, establishes new standards, evaluated by committees or groups. Example: Establishment of a monitoring system, percentage of school-age children who study food education in primary schools.
Financial	Budget allocation, costs, cost of products, sales, market share, marketing, household income change. Example: Annual US$ disbursement, percentage increase in household income.
Food	Eating habits, specific food, nutrition programmes, supplements, vitamins, food production, crop diversion. Example: Adequate food for three meals for a family of five members, sodium content of foods.
Market	Specific products or product sales, market engagement, companies. Example: Percentage of sales volume meeting the criteria for highest nutritional standards, marketing of highly processed foods high in sugar, salt and fat outlawed.
Prevalence	Everything that mentions prevalence or gives a value for target prevalence.

Indicator name group	Description
	Example: Prevalence of stunted children under five years of age, prevalence of low birth weight.
Results	Change in government, policy, decisions, training of people; change in production; change in nutrition habits; change in services, awareness, numerical values. Example: Number of cases of maternal anaemia averted, number of enacted and enforced mandatory legislations for food fortification of selected staples.
Unknown	Indicator name described as 'unknown', 'TBD', 'TBC', or text indicating this has not been decided yet; indicator name described as 'not applicable'.

Action area

Standardisation for geographic area (GX.4 and GX.5 on the Commitment Registration Form where X refers to the goal number) only occurs for commitments reporting the geographical area of their commitment to be 'National', 'Sub-national' or 'Local'. If a country is not specified in the goal description (Q13), geographic area (GX.4) or additional information about the geographic area (GX.5), the country of the organisation, as specified in the Sign up Form, is used. If a country different to the organisation's country is specified in Q13, GX.4 or GX.5, the alternative country is used instead. This results in a country name as the action area. These countries are further grouped into global regions (e.g., Asia), and sub-regions (e.g., Central Asia), as per the GNR's standard groups.

The income level for each country action area is calculated using the World Bank development indicators.[2] Each country action area is also assessed to determine whether it experiences high levels of three types of malnutrition. The thresholds for assessing whether a country is burdened or not are based on the following prevalence: stunting in children aged under five years ≥20%; anaemia in women of reproductive age ≥20%; overweight (BMI ≥25) in adult women aged ≥18 years ≥35%.[3] An action area is categorised as having 0–3 burdens of malnutrition, unless data for any indicator is not available, in which case the burden of malnutrition cannot be classified.[4]

Quality assurance

An internal and external quality assurance system was used to check the accuracy of the data cleaning and standardisation process. Internal quality assurance aimed to check the consistency of the outputs. Performed by an analyst not involved in the coding, it involved a series of checks on the statistical output generated:

- The aggregated frequencies for individual stakeholders' groups must be equal to the total number of commitments/goals in the sample.
- The aggregated breakdown frequencies within each group/sub-group must be equal to the total frequencies of the group (number of stakeholders/commitments/goals as relevant).
- The totals for each stakeholder must be consistent within each statistical output.

The external quality assurance process focused on validating the correctness of the Python code compiled to generate the master dataset used for the analysis. An independent data analyst used the raw data to recompile the data cleaning and standardisation code. The dataset generated was then compared with the original dataset.

Footnotes

1. World Bank. World Development Indicators: Exchange rates and prices. 2022. http://wdi.worldbank.org/table/4.16 [website accessed on August 31, 2022].

2. World Bank. World Bank Country and Lending Groups. 2022. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [Accessed 25 August 2022].

3. Sources: UNICEF/WHO/World Bank Group: Joint child malnutrition estimates: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/gho-jme-country-children-aged-5-years-stunted-(-height-for-age--2-sd); WHO Global Health Observatory: https://apps.who.int/gho/data/view.main.ANAEMIAWOMENREPRODUCTIVECOUNTRYv; NCD Risk Factor Collaboration. [Accessed 31 August 2022].

4. Data from 2019 was used for all three indicators. Data was available for all three indicators for 151 countries, therefore 44 countries were not classified with a level of burden of malnutrition.