**Identifying outliers and errors; and analyzing outliers, errors, and missing values for the t-HFA**

| **Question no.** | **Brief description of question** | **Comments on identifying outliers and errors; and analyzing outliers, errors, and missing values** |
| --- | --- | --- |
| A002 | Facility’s catchment population | * Identifying outliers and errors. Look for values that are either implausibly large or small. First consider the facility type (question A001: primary, secondary, or tertiary), as the order of magnitude of the catchment population will depend on facility type.   + For primary and secondary facilities, outliers might be implausibly large values (e.g., populations that are larger than the district or province where the facility is located).   + For tertiary facilities, outliers might be implausibly large values (e.g., populations that are larger than the country). Note that the concept of a catchment population for a tertiary facility might be ambiguous: a respondent might interpret catchment population to mean the nearby area (e.g., the city in which the hospital is located) where most clients/patients live, or catchment population might be interpreted as being very large (e.g., an entire state or country) because the hospital is a referral center for a very large area.   + For any facility type, look for implausibly small values with 1 or 2 digits (e.g., a surveyor who intended to enter a “don’t know” code of 9 incorrectly enters another single digit, such as 6 or 8, or a similarly small value, such as 99).   + For any facility type, look for large values that are repeated digits (e.g., 111111 or 999999), which could be keypunching errors. * Analyzing missing values and outliers that must be removed. The best approach is to obtain this information from the country’s statistics/demographics bureau. An alternative approach is to use the mean value for the facility type (question A001: primary, secondary, or tertiary) that corresponds to the facility with the missing value. For example, if a primary care facility has a missing value (and you can’t find this value from the country’s statistics/demographics bureau), then let the missing value equal the mean value of catchment population for all primary care facilities in the country (with non-missing values). |
| B025, B026, B046 to B057 | No. of health worker posts, and posts not filled for at least 6 months. | * Identifying outliers and errors. Look for values that are either implausibly large or small. First consider the facility type (question A001: primary, secondary, or tertiary), as the order of magnitude of the number of health worker posts will depend on facility type.   + For primary facilities, implausibly large numbers might be more than 50 health worker posts or more than 10 physicians. However, as some countries have unusually large staff numbers, any potential outlier should be checked against the overall distribution of staff numbers for the country.   + For secondary and tertiary facilities, implausibly small numbers might be less than 10 health worker posts.   + For any facility type, the value of B046 (total full-time posts) should be greater than or equal to the sum of B048 (physicians), B050 (nurses & midwives), B052 (lab technicians), B054 (pharmacists), and B056 (community health workers [CHWs]). Similarly, the value of B047 (full-time posts not filled for 6+ months) should be greater than or equal to the sum of B049 (physicians), B051 (nurses & midwives), B053 (lab technicians), B055 (pharmacists), and B057 (CHWs). If either sum is greater than B046 or B047, then try to validate all health worker post data with another data source (e.g., from a country’s human resources department). If this validation is not possible, then the sums should be used in place of B046 or B047. Similarly, the value of B046 should be greater than B047.   + For any facility type, look for large values that are repeated digits (e.g., 666 or 888), which could be keypunching errors. * Analyzing missing values and outliers that must be removed.   + The best approach is to obtain this information from the country’s human resources department. An alternative approach is to use the mean value for the facility type (question A001: primary, secondary, or tertiary) that corresponds to the facility with the missing value. For example, if a primary care facility has a missing value of total full-time posts (and you can’t find this value from the country’s human resources department), then let the missing value equal the mean value of total full-time posts for all primary care facilities in the country (with non-missing values).   + If B052 and B053 (lab technicians) are missing or don’t know: if the facility has no lab or performs no lab tests (i.e., question 2900 = 2 or 3), the value for B052 and B053 might be zero.   + If B056 and B057 (CHWs) are missing or don’t know: if the facility has no CHWs (i.e., question C010 = 2), then the value for B056 and B057 might be zero. |
| B034 and B058A to B058E | Outpatient caseload | * Identifying outliers and errors. Look for values that are either implausibly large or small. First consider the facility type (question A001: primary, secondary, or tertiary), as the order of magnitude of caseload will depend on facility type.   + For primary facilities, implausibly large numbers might be more than 200 outpatients. However, as some countries have unusually large caseloads, any potential outlier should be checked against the overall distribution of caseloads for the country.   + For secondary and tertiary facilities, implausibly small numbers might be less than 10 outpatients.   + For any facility type, if a caseload is zero, check if the date corresponds to a day that the facility might normally have been closed (e.g., a weekend or holiday). In such cases, a zero value is plausible.   + For any facility type, look for large values that are repeated digits (e.g., 666 or 888), which could be keypunching errors. * Analyzing missing values and outliers that must be removed.   + If any of the caseload variables have a missing value and the date corresponds to a day when the facility would normally be closed (e.g., weekend or holiday), then the value can be assumed to be zero.   + For days when the facility would normally be open and for facilities with at least one non-missing, non-outlier caseload value, any variables that are missing or that have an excluded outlier can be assigned a value equal to the mean of the remaining non-missing, non-outlier caseload values for that facility.   + For days when the facility would normally be open and for facilities with all missing caseload values, then the missing values can be imputed using the mean value for the same facility type in the country. |