105% Registration

Findings and Actions

November, 2019 – Draft

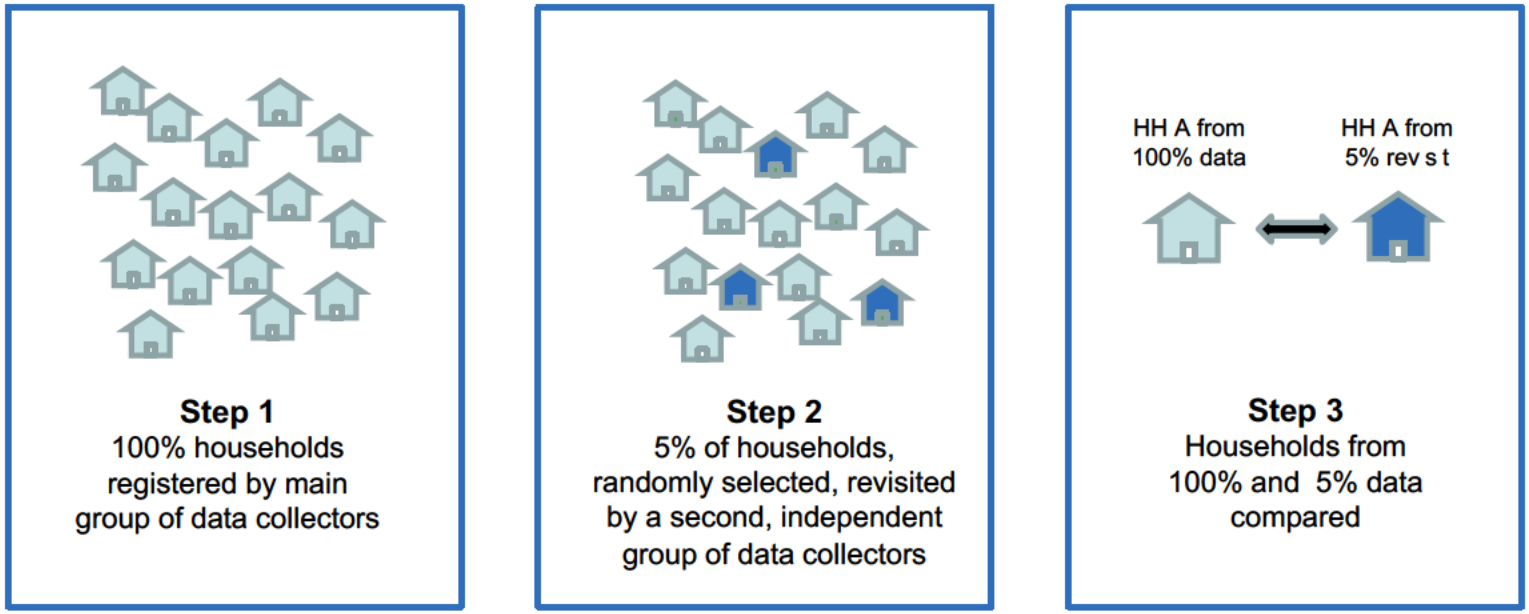
105% Registration - Process

105% registration is a process where 5% of all households registered (the 100%) are randomly selected and re-visited.

Objectives

1. To increase the quality of the 100% registration by informing data collectors ahead of time that a random sample of their work will be checked
2. To assess the quality of the 100% registration by comparing the 100% (registration, or main) and 5% (revisit) datasets

Process



Implementation

AMF has implemented the 105% registration process in every distribution since it was introduced in 2017. Differences in data collection and transfer (outlined below) affect how we match and compare the 100% and 5% household data.

Country	Year	Data collection	Data transfer
Uganda	2017	Paper	100% and 5% data received in a single Excel file
Togo	2017	Paper	
Zambia	2018	Paper	100% and 5% data received as separate Excel files
Malawi	2018	Paper	Input to the DES
Ghana	2018	Electronic	100% and 5% data received in a single Excel file
Guinea	2019	Paper	Input to the DES

Re-visit Completion

Country	Distribution	Revisit % Required	Actual Revisit %
Uganda	2017	5%	3.0%
Togo	2017	5%	4.5%
Zambia	2018	5%	10.0%
Malawi	2018	5%	3.5%
Ghana	2018	1.5%	1.4%
Guinea	2019	5%	5.0%

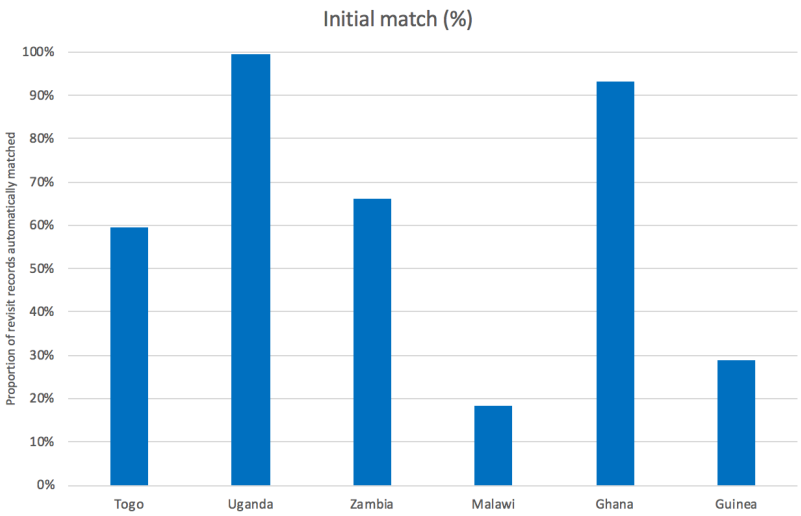
Completion rates are generally close to, or greater than, required.

Automatic Matching – Results

The percentages of 5% records that we can automatically match to a 100% record in the DES varies by country, between 18% and 99%. Some of this variation can be explained by differences in data collection and transfer.

At present, automatic matching takes place only within a village, and generally includes:

- Matching on coupon/serial number (a unique ID, where relevant)
- Matching on phone number
- Matching on first and last name of household head



Note: In Ghana the matching percentage is high because re-visit data was captured electronically using the household's unique ID. In Uganda, the registration and revisit data were received in a single file so matching was effectively already done.

In Guinea and Malawi, the automatic matching percentage is low because of issues with the location hierarchy. In Guinea, a number of health centres were erroneously created during data entry. These were later corrected, but this led to an inflated number of villages. In Malawi, the use of 'distribution site' as a location likewise inflated the number of villages.

Assessing Data Quality

For registration analysis, one current AMF standard is that when matching pairs from the 100% and 5% datasets are compared, data quality is acceptable given:

- the proportion of pairs with household populations within +/- 1 people is >= 75%,
- the proportion of pairs with LLIN quantities within +/-1 LLINs is >= 75%.

These standards (thresholds) are not set in stone, but reflect the realities of working in challenging environments where:

- Individuals hold different understandings of the definition of a household, and of a sleeping space
- 100% and 5% data collectors may not speak with the same “household head”, meaning they may collect data from individuals with differing definitions
- Despite consistent national training, 100% and 5% data collectors may still apply different definitions of households and implement net allocation differently.
- While revisit data collection should be carried out shortly after the main data collection, high levels of movement may mean that household composition changes even in this time.

These thresholds assess discrepancies between 100% and 5% data from the same household. We will continue to monitor these thresholds as we receive and review data from new distributions.

Data Quality Results

Across all distributions, the 105% results pass the threshold for data quality.

		Household Population		LLIN Quantity		
Country	Distribution	Exact	≤ ±1 person	Exact	≤ ±1 LLIN	Key data match
Togo	2017	67%	86%	76%	94%	60%
Uganda	2017	86%	90%	100%	100%	85%
Zambia	2018	96%	98%	97%	99%	96%
Malawi	2018	91%	96%	95%	99%	62%
Ghana	2018	89%	93%	91%	97%	89%
Guinea	2019	64%	76%	68%	84%	60%

This table also shows the results for key data matches, i.e. the proportion that matches on both household population and LLIN quantity.

Key questions

These findings prompt two key questions about using the 105% process to measure the quality of registration data:

1. **Confidence about past distributions:** While it is encouraging that data quality *within* the automatically matched samples holds to AMF standards, do we have confidence that the quality of key data (i.e. household population and LLIN quantity) is similarly high for the unmatched records?

2. **Improved matching going forwards:** How can we improve the initial match percentage, such that measures of data quality are based on a larger number of matched pairs?

Approach to answering the questions

To answer these questions, we conducted manual matching for a representative, randomly selected sample of re-visit records that could not be matched automatically. Further details on the method and full results can be shared on request.

The records were selected such that the proportion for each country in the sample reflected the proportion for each country across all unmatched re-visit records. Each re-visit record in the sample was manually compared to registration records.* Where a match existed, this was noted so that 100% and 5% records could be compared. Issues that prevented matching were also recorded.

Country	Distribution	Random Sample (#)
Togo	2017	73
Uganda	2017	Not required**
Zambia	2018	64
Malawi	2018	126
Ghana	2018	20
Guinea	2019	113
Total		**396**

*During comparison, the number of people and nets were withheld to prevent bias.

**Originally, 104 records for Uganda were also selected, for a total of 500 records. During the matching exercise, it became clear that there were duplicates in the re-visit data. Once these were removed, automatic matching increased such that manual matching was no longer required.

Q1: Confidence in quality

These findings prompt two key questions about using the 105% process to measure the quality of registration data:

1. **Confidence about past distributions:** While it is encouraging that data quality *within* the automatically matched samples holds to AMF standards, do we have confidence that the quality of key data (i.e. household population and LLIN quantity) is similarly high for the unmatched records?

2. **Improved matching going forwards:** How can we improve the initial match percentage, such that measures of data quality are based on a larger number of matched pairs?

Data quality – manual matching

The results for the matching pairs from the manual sample show that the quality of key data (i.e. household population and LLIN quantity) passes the current threshold.

As a reminder, the thresholds are:

- the proportion of pairs with household populations within +/- 1 people is >= 75%,
- the proportion of pairs with LLIN quantities within +/-1 LLINs is >= 75%.

		Household Population		LLIN Quantity		
Country	Distribution	Exact match	≤ ±1 person	Exact match	≤ ±1 LLIN	Key data match
Togo	2017	62%	81%	70%	94%	53%
Zambia	2018	78%	94%	67%	89%	61%
Malawi	2018	77%	93%	85%	96%	74%
Ghana	2018	94%	100%	94%	100%	88%
Guinea	2019	63%	75%	63%	80%	56%

Exact matches on both household population and LLIN quantity (i.e. key data match) vary by country.

Data quality results – all matching

Overall, the data quality for the automatically matched and manually matched data is similar. Combining the results, the data passes the thresholds for data quality.

Comparing the manual matching results to the automatically matched results, we see:

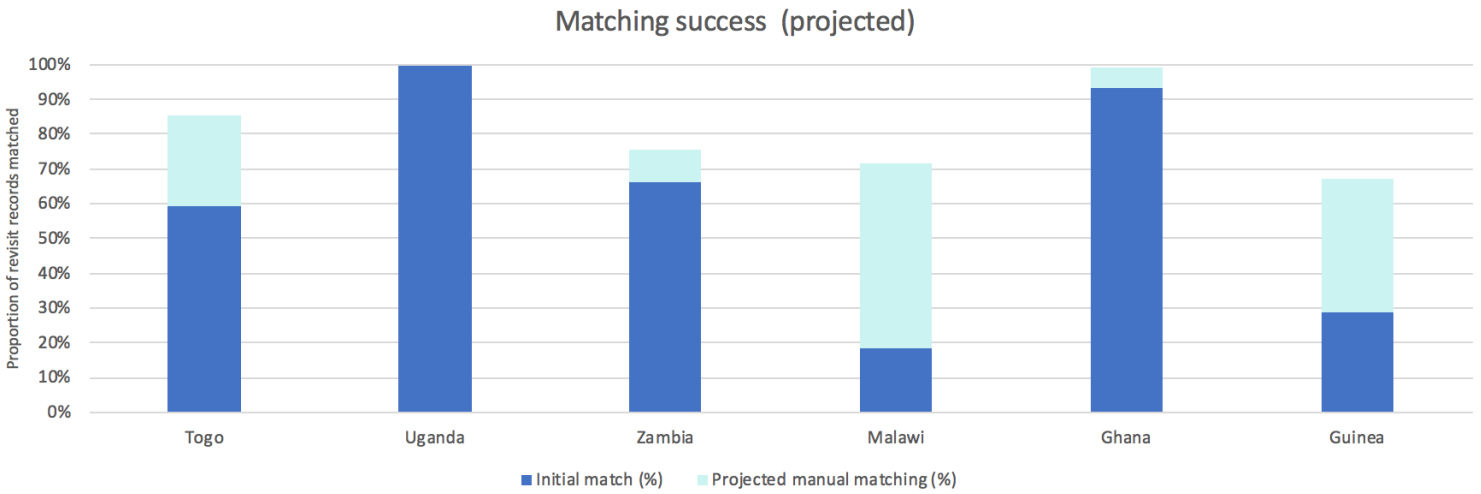
- similar data quality for household population in all distributions, with the key measure (i.e. < +/- 1 person) always within 5 percentage points.
- similar data quality for LLIN quantity in most distributions, with the key measure (i.e. < +/- 1 LLIN) within 5 percentage points. The exception is Zambia, where the key measure differs by 10 percentage points.

Extrapolating from the manually matched sample, and calculating a blended average, we can see that in all cases the data meets the AMF standards.

		From automatic matching			From manual matching			Blended rates		
Country	Distribution	≤ ±1 person	≤ ±1 LLIN	Key data match	≤ ±1 person	≤ ±1 LLIN	Key data match	≤ ±1 person	≤ ±1 LLIN	Key data match
Togo	2017	86%	94%	60%	81%	94%	53%	84%	94%	58%
Uganda	2017	90%	100%	85%	-	-	-	90%	100%	85%
Zambia	2018	98%	99%	96%	94%	89%	61%	98%	97%	92%
Malawi	2018	96%	99%	62%	93%	96%	74%	94%	97%	71%
Ghana	2018	93%	97%	89%	100%	100%	88%	94%	97%	89%
Guinea	2019	76%	84%	60%	75%	80%	57%	75%	82%	59%

Data quality - overall

We can have confidence that our data passes the quality threshold for all results that we can match (either automatically or manually), which varies from 67% to 100% of the data.



The remaining re-visit data that remains unmatched must have a location and/or household head name that differs from the registration record. The key question is then whether we think this remaining data therefore also has inaccurate household populations/ LLIN quantities. Need a sentence here, like: **On balance, we take the view that inaccuracies in recording/ inputting this information do not necessarily mean that household populations and LLIN quantities are incorrect.**

Q1: Summary answer

These findings prompt two key questions about using the 105% process to measure the quality of registration data:

1. **Confidence about past distributions:** While it is encouraging that data quality *within* the automatically matched samples holds to AMF standards, do we have confidence that the quality of key data (i.e. household population and LLIN quantity) is similarly high for the unmatched records?

Given these results, yes, we can have confidence as the results of manual matching show that the data quality of the records is similar and passes the thresholds. This means that while we may not be able to automatically match a 'high proportion' of records, when the matched data passes the threshold we can be more confident that the results represent the data as a whole.

2. **Improved matching going forwards:** How can we improve the initial match percentage, such that measures of data quality are based on a larger number of matched pairs?

Q2: Improving automatic matching

These findings prompt two key questions about using the 105% process to measure the quality of registration data:

1. **Confidence about past distributions:** While it is encouraging that data quality *within* the automatically matched samples holds to AMF standards, do we have confidence that the quality of key data (i.e. household population and LLIN quantity) is similarly high for the unmatched records?

Given these results, yes, we can have confidence as the results of manual matching show that the data quality of the records is similar and passes the thresholds. This means that while we may not be able to automatically match a 'high proportion' of records, when the matched data passes the threshold we can be more confident that the results represent the data as a whole.

2. **Improved matching going forwards:** How can we improve the initial match percentage, such that measures of data quality are based on a larger number of matched pairs?

Improving matching - methods

To improve automatic matching, we need to incorporate some of the elements of the manual matching.

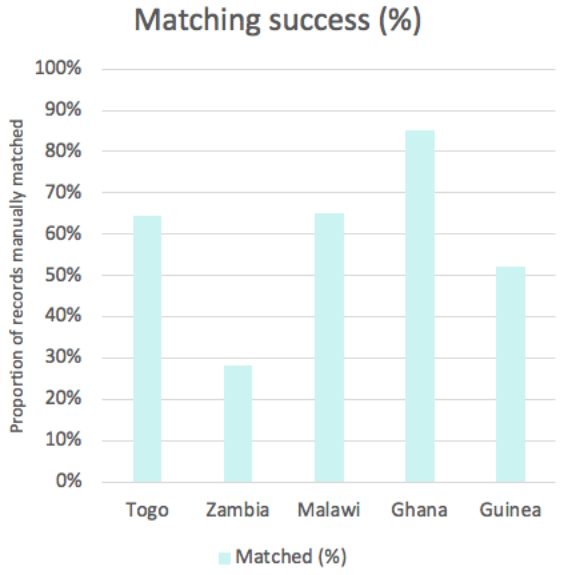
The key differences between the methods are:

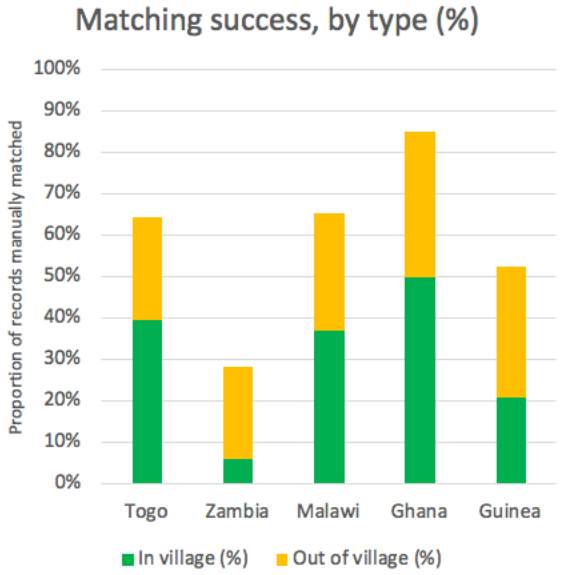
Matching element	Automatic	Manual
Location	Within village	Within village and outside the village (e.g. in health centre)
"Sounds like" matching	No	Yes
Sensitive to spelling errors	No	Yes
Able to differentiate between records with the same household head name (e.g. based on mother's name)	No	Yes

Improving matching – by location

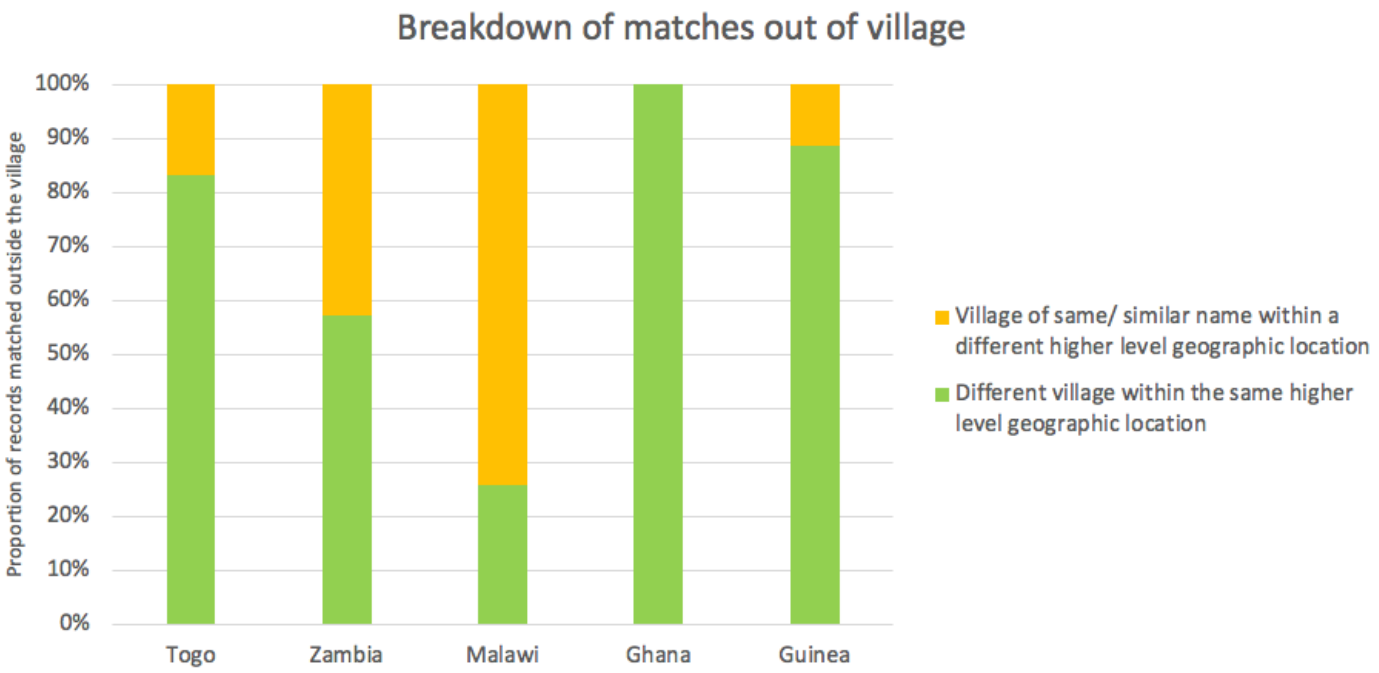
A significant proportion of the matches that were made manually were made by considering households from outside the village.

In Zambia, matching in the village is is low, as 40% of revisit data was allocated to villages without registration data. This is likely due to the fact that the re-visit and registration data was received in two separate files, with different location hierarchies.





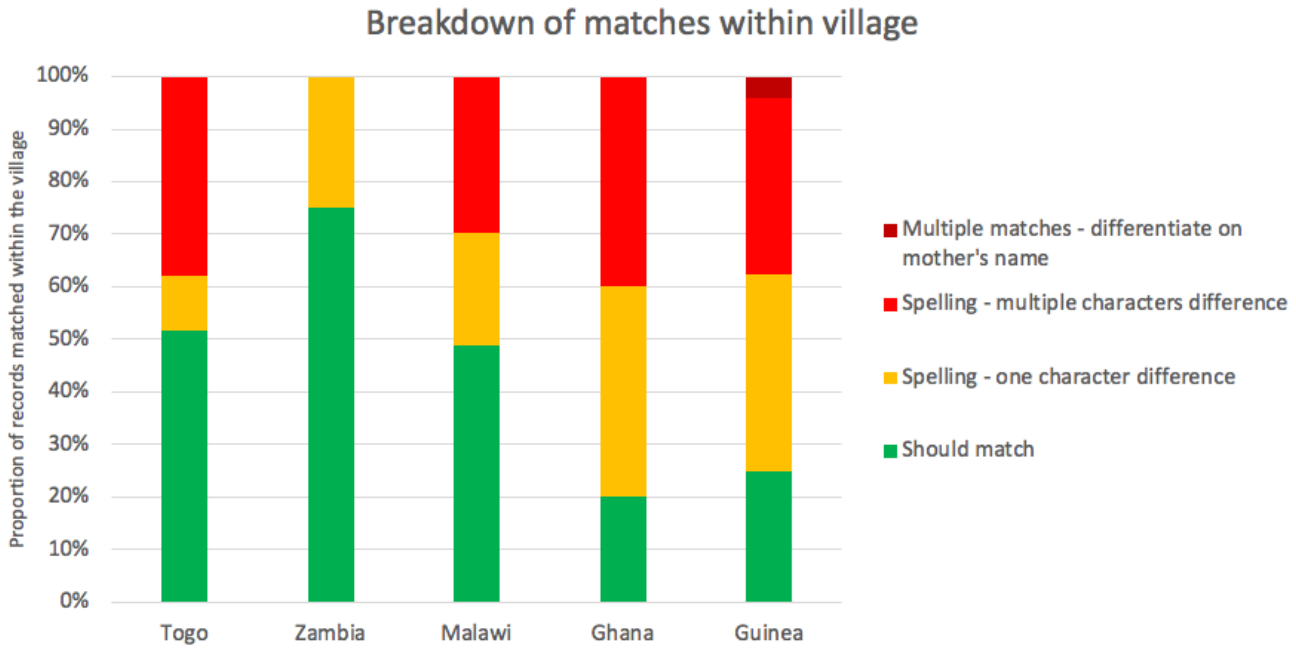
Improved matching – outside villages



Matching automatically outside the village could involve either a) matching in a different village in the same higher-level location (e.g. district), or b) matching in another village of a same or similar name.

An operational solution to improve the location hierarchy in data collection and input will be easier to implement than a technical solution to improve matching.

Improved matching – other elements

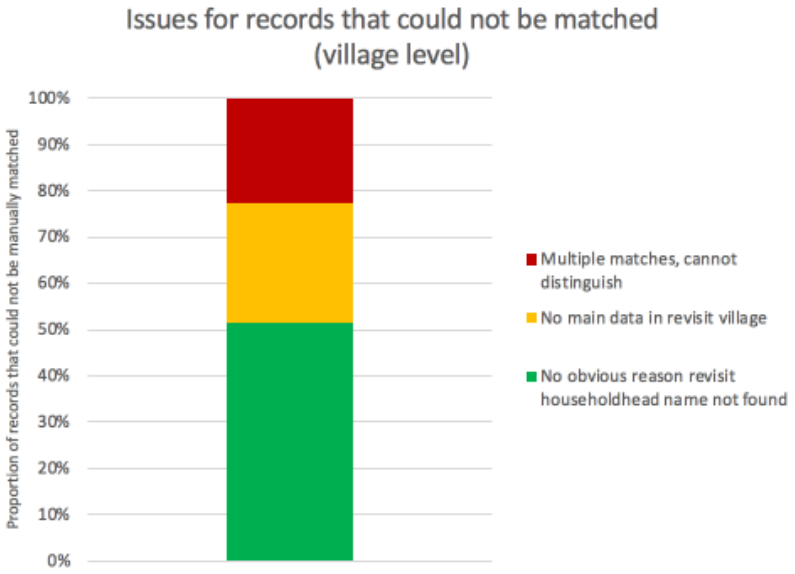


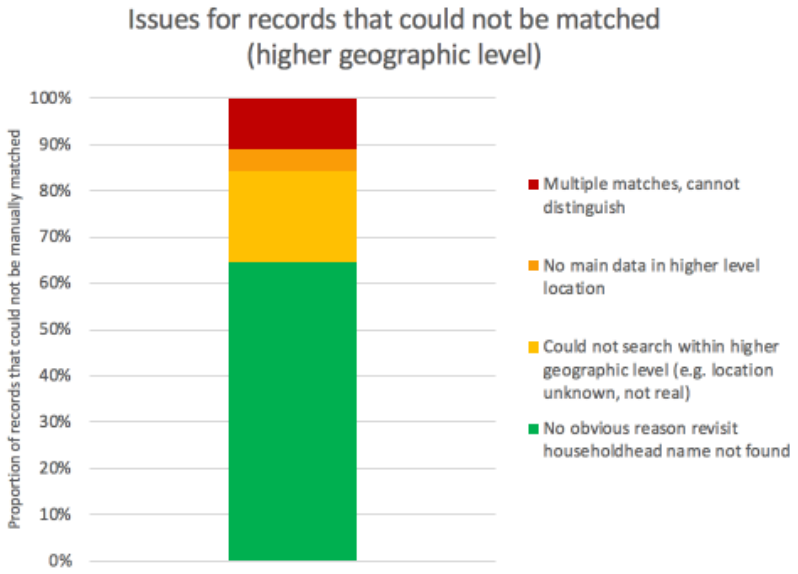
Incorporating other elements of manual matching would improve our matching within the village. However, implementation can be complex, as indicated by the traffic light system in the graph.

- Should match – a restriction in the code disallows the match, we are checking this to evaluate next steps.
- Spelling, one character difference – rewriting the code to find these is complicated, but possible.
- Spelling, multiple characters difference – the code is difficult to write, and likelihood of false matches is high.
- Multiple matches, differentiate on mother's name – this would significantly increase the time to run matching on the DES (already several days per country).

Improved matching – issues not addressed by manual matching

Considering the records that could not be manually matched, it is clear that even if we were able to incorporate all elements of manual matching into automatic matching, there would still be records that would remain unmatched. In almost half the cases, this would be due to either a) no main data in the re-visit village, or b) multiple matches that we cannot distinguish.





These issues will become less common as we move towards electronic data collection, but it will remain difficult to address the issue of multiple records with the same household head name.

Actions to improve matching

The results of this exercise suggest a number of key steps we can take to improve our matching in the future. In order of priority, these are:

Technical	Operational
Check our matching code to see what restrictions are disallowing matches on phone number and name within a village, and take appropriate action.	Ensure that the location hierarchy used for revisit data collection is the same as that used for registration.
Disable the creation of higher geographic levels (e.g. health centres) when data is being entered.	
Scope out the work involved in a technical solution that allows for matching where there are single character differences.	
Investigate matching using GPS coordinates, given move to electronic data collection.	

Conclusions

These findings prompt two key questions about using the 105% process to measure the quality of registration data:

1. **Confidence about past distributions:** While it is encouraging that data quality *within* the automatically matched samples holds to AMF standards, do we have confidence that the quality of key data (i.e. household population and LLIN quantity) is similarly high for the unmatched records?

Given these results, yes, we can have confidence as the results of manual matching show that the data quality of the records is similar and passes the thresholds. This means that while we may not be able to automatically match a 'high proportion' of records, when the matched data passes the threshold we can be more confident that the results represent the data as a whole.

2. **Improved matching going forwards:** How can we improve the initial match percentage, such that measures of data quality are based on a larger number of matched pairs?

We will take a number of steps, both operational and technical, to improve matching. In doing so, the initial matching rate should increase.