Data Management and Statistical Analysis Plan (DMSAP)
Version 1.0

Molecular Module

WorldWide Antimalarial Resistance Network (WWARN)

Suggested citation: Molecular Module, WWARN, 2011. Data Management and Statistical Analysis Plan.

Version History

Version number	Revision(s) & reason for amendment	Release date
v1.0	Creation of document	05/19/2011

For more information, contact:
molecular@wwarn.org

WorldWide Antimalarial Resistance Network (WWARN)
www.wwarn.org

Contents

1. Scope

A key aim of WWARN is to monitor geospatial and temporal trends in antimalarial drug resistance. To achieve this, molecular marker data need to be collated at an individual patient level so that standardized methodologies and definitions can be applied[1].

The purpose of the Molecular Module **Data Management and Statistical Analytical Plan** (DMSAP) is to present a clear and transparent methodology by which WWARN handles and analyses molecular data. Documenting the entire process by which data are uploaded, transformed, analysed and presented ensures reproducibility. It also provides a framework for discussing and developing such methodology. WWARN hopes that this process will facilitate best practice, timely data publication and allow geospatial and temporal comparison of molecular data.

2. Introduction

Molecular markers of antimalarial resistance are a critical tool for timely and comprehensive surveillance of drug efficacy. The Molecular Module of the Worldwide Antimalarial Resistance Network (WWARN) is working to strengthen the practical value of resistance markers in public health by compiling molecular surveillance data and linking molecular markers with phenotypes of clinical and *in vitro* resistance. Linkage of genetic markers to clinical and in vitro phenotypes is particularly critical for validation of candidate resistance markers to artemisinins. Additionally, the WWARN Molecular Module aims to enhance scientific capacity in endemic countries and increase the amount of useful molecular data available.

The Molecular Module is working towards these goals as follows:

- Facilitating the inclusion in a Data Repository of results from molecular studies carried out by research groups, Non-Governmental Organisations (NGOs) or National Malaria Control Programmes (NMCPs). This will include comprehensive prevalence maps of antimalarial resistance markers in space and time as well as current data on the emergence of resistance markers by location.
- Creating standardised processes to facilitate the collection of diverse datasets from studies taking place around the world.

[1] Plowe *et al* 2007. World Antimalarial Resistance Network (WARN) III: molecular markers for drug resistant malaria. *Malar J.* 2007 6:121.

- Optimising analytical tools to increase comparability of results between heterogeneous studies and to aid in marker validation by testing the association between molecular markers and in vivo and in vitro phenotypes of resistance.

2.1 Compiling data on molecular markers of antimalarial resistance

High quality data on molecular markers are collected by many research groups, NGOs and NMCPs around the world. However, differences in malaria epidemiology, patient populations, specific research questions and logistical constraints contribute to variation in study design, and data are recorded in diverse formats. In addition, analysis may vary depending on genotyping methods and the treatment of mixed genotype samples. The interpretation of aggregated data from published results of molecular studies is therefore fraught with confounding factors[2], making it difficult to reliably assess geographical or temporal trends.

2.2 WWARN's process of data collation

WWARN aims to facilitate two processes. Firstly to give researchers the tools to collect, clean and analyse their own data, and secondly to transform molecular data from a diverse range of studies into a common format that can be derived from almost all database structures, so that data from different studies can be pooled and analysed collectively in a standardised manner.

To achieve these goals, a series of steps are followed.

I. Upload individual data from anonymised patients or parasite isolates from a clinical study or community survey.
II. Transform uploaded data to a common format.
III. Check and identify unexpected data points.
IV. Data revision - either by the data submitter, or automatically using a set of rules to autocorrect unexpected data points and neutralise their impact on the analysis.
V. Data analysis, applying uniform analytical methodology and reporting to provide consistent estimates of marker prevalence. These are provided as an automated report to the data submitter. If the data submitter agrees, the data may be presented on WWARN Explorer, an interactive, online tool which allows users to view studies filtered by treatment, time of sampling, molecular marker and geographic location.

WWARN notes that other researchers may take different approaches to data management, particularly with regard to analysing mixed infections and haplotypes. It is very important to stress that WWARN-derived prevalence estimates may vary to

[2] Picot *et al* 2009. A systematic review and meta-analysis of evidence for correlation between molecular markers of parasite resistance and treatment outcome in falciparum malaria. *Malar J.* 8:89

some degree from analyses performed by the data submitter. These differences do not reflect a value judgment as to which analytical approach is correct. The decisions are made only to apply standardised methodologies and minimise bias on geospatial and temporal trends derived from the many studies compiled in the Data Repository.

3. Dataset submission process

The Molecular Module accepts datasets from cross-sectional surveys, antimalarial clinical drug studies, *in vitro* samples and other studies with molecular data that have been obtained in accordance with any laws and ethical approvals applicable in the country of origin.

3.1 Variables

The dataset and/or accompanying documents (e.g. protocols, publications) must contain the following information:

I. Unique sample or patient identifier
II. Date of inclusion or sample collection
III. Days or dates of follow up visits (if applicable)
IV. Genotypes of molecular markers of resistance, which may include:
    - Single nucleotide polymorphisms (SNPs) – expressed as nucleic acids or amino acids
    - Variable length repeats
    - Copy number polymorphisms

If these essential variables are not provided or immediately apparent, additional information may be required prior to data processing and output generation. In these situations, WWARN Data Managers will contact the data submitter for clarification.

The Molecular Module also requests additional variables, where available. These will be used to generate additional outputs in current and future outputs:

I. Patient age or date of birth (high priority, if available)
II. Genotyped microsatellites in regions flanking resistance markers (future outputs)
III. Complexity or multiplicity of infection (future outputs)

The full molecular dictionary is available in Annex A.

3.2 Online submission system

Datasets are submitted using the WWARN online submission system. Data contributors must accept the Terms of Submission [available at http://www.wwarn.org/data/usage] at the beginning of the submission process.

The data submission steps are:

I. **Register a study**: contributors enter their study title. Each study is assigned a unique identifier.
II. **Permissions**: each study may have any number of administrators, assigned by the data submitter who created the study title and accepted the Terms of Submission. Any designated administrator can access the study, upload files and edit supplied information.
III. **Files**: contributors are asked to submit data files and supporting documentation, such as a data dictionary.
IV. **Acknowledgements**: the names of acknowledged individuals and institutions will appear in the study details displayed in WWARN Explorer.
V. **Study info**: data submitters are asked to provide information on the study site and study design. They may enter this information themselves or provide relevant protocols and publications which will be used by Data Managers to extract the relevant data.

4. Data extraction to a standardised format

Submitted data are extracted and transformed into a standardised format, allowing single study analysis, the generation of a study report and, with permission, visualisation of summary study information on WWARN Explorer. The transformed data are stored in a secure Data Repository.

4.1 WWARN Molecular Dictionary

The WWARN Molecular Dictionary contains molecular variables required for generation of prevalence estimates. A copy of the Molecular Dictionary is attached in Annex A.

4.2 Database structure

The WWARN Molecular Dictionary is arranged in six tables linked by a unique patient identifier and date of sample collection. The tables, with examples of component variables, are:

I. **Study**, e.g. investigator, study label
II. **Location**, e.g. country, site
III. **Subject**, e.g. patient age (if available), date of inclusion
IV. **Sample**, e.g. date of sample collection
V. **Genotype**, e.g. resistance status, molecule type, genotype value
VI. **Marker**, e.g. locus name, locus position

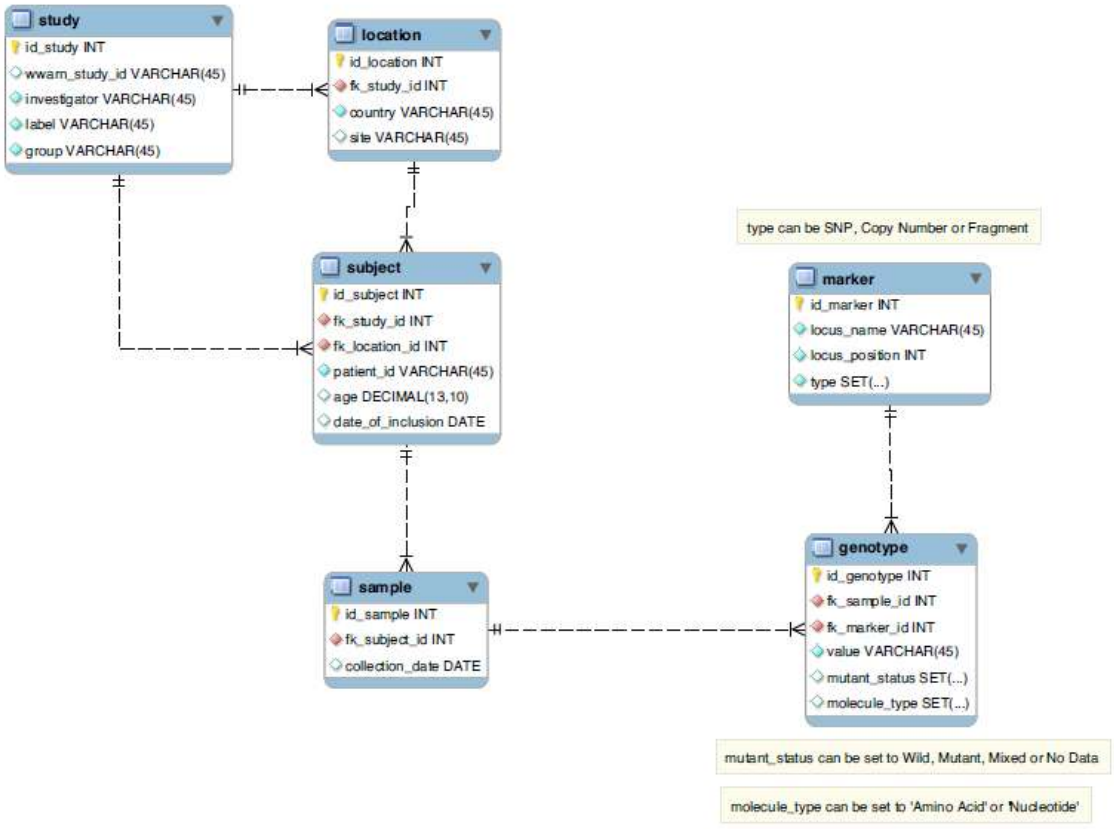


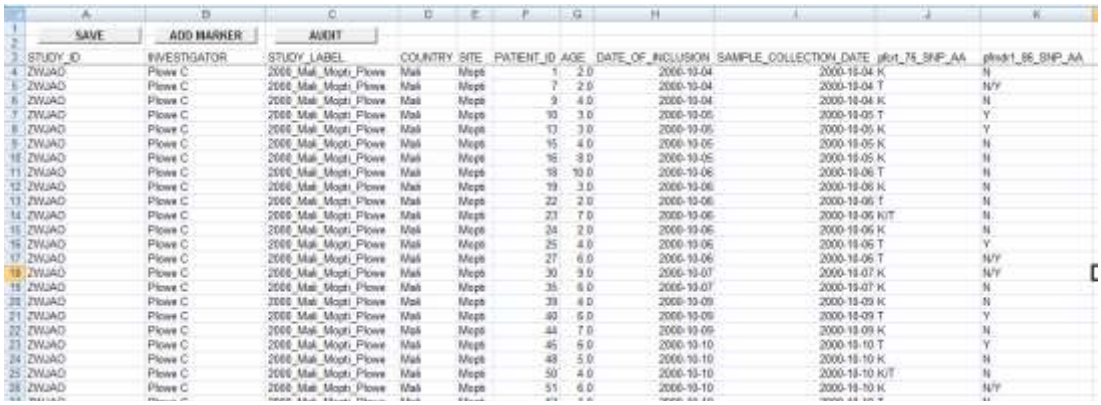
Figure 1. WWARN Molecular database structure

Each patient may be linked with several samples, collected at different times or days. Each sample is linked with corresponding genotypes, marker names and positions.

Submitted data are stored in the Data Repository within these six tables. Data within the Molecular Module can be linked to associated clinical, pharmacology and *in vitro* data using the study and patient identifiers.

4.3 Data transformation and extraction

Source data may be presented as flat files, with one line per patient or sample, or multiple relational databases. The extraction process transforms all source datasets into a common standard format. Variables from the source dataset, equivalent to variables in the WWARN Molecular Dictionary, are extracted and transformed into a flat intermediate template file. Study and location information are extracted from the Study Site Questionnaire. From the template, data are automatically imported into one of the six tables. An audit trail records and saves the complete data extraction and transformation process. Once transformed, the submitter or their designees may download the derived dataset; this can either be in set of a relational databases or the flat intermediate template file, both amenable for offline analysis. Some important considerations in the data transformation process are described below.



	A	B	C	D	E	F	G	H	I	J	K
1	SAVE	ADD MARKER	AUDIT								
2											
3	STUDY_ID	INVESTIGATOR	STUDY_LABEL	COUNTRY	SITE	PATIENT_ID	AGE	DATE_OF_INCLUSION	SAMPLE_COLLECTION_DATE	pfcrt_76_SNP_AA	pfmdr1_86_SNP_AA
4	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	1	2.0	2008-10-04	2008-10-04	K	N
5	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	7	2.0	2008-10-04	2008-10-04	T	N/Y
6	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	9	4.0	2008-10-04	2008-10-04	K	N
7	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	10	3.0	2008-10-05	2008-10-05	T	Y
8	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	13	3.0	2008-10-05	2008-10-05	K	Y
9	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	15	4.0	2008-10-05	2008-10-05	K	N
10	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	16	8.0	2008-10-05	2008-10-05	K	N
11	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	18	10.0	2008-10-06	2008-10-06	T	N
12	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	19	3.0	2008-10-06	2008-10-06	K	N
13	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	22	2.0	2008-10-06	2008-10-06	T	N
14	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	23	7.0	2008-10-06	2008-10-06 K/T		N
15	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	24	2.0	2008-10-06	2008-10-06 K		N
16	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	25	4.0	2008-10-06	2008-10-06 T		Y
17	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	27	6.0	2008-10-06	2008-10-06 T		N/Y
18	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	30	9.0	2008-10-07	2008-10-07 K		N/Y
19	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	35	6.0	2008-10-07	2008-10-07 K		N
20	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	39	4.0	2008-10-09	2008-10-09 K		N
21	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	40	6.0	2008-10-09	2008-10-09 T		Y
22	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	44	7.0	2008-10-09	2008-10-09 K		N
23	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	45	6.0	2008-10-10	2008-10-10 T		Y
24	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	48	5.0	2008-10-10	2008-10-10 K		N
25	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	50	4.0	2008-10-10	2008-10-10 K/T		N
26	ZWJAD	Plowe C	2008_Mali_Mopti_Plowe	Mali	Mopti	51	6.0	2008-10-10	2008-10-10 K		N/Y

Figure 2. Example of intermediate template file used in molecular data transformation.

4.3.1 Dates of inclusion and sample collection

If the date of inclusion or date of sample collection per patient is not specified, a theoretical date of inclusion of January 1 [study year] is used. This date is used to tag the sample to a specific year and to derive the subsequent dates of follow up or any further samples collected from the same patient.

4.3.2 Marker names

In the intermediate template, molecular marker names are converted to a standard format of locusname_locusposition_markertype_moleculetype (e.g. pfcrt_76_SNP_AA). These marker names are then extracted into the genotype and marker tables of the database (Figure 1).

4.3.3 SNPs

The values of SNPs are stored as either nucleic acids or amino acids in the molecular database. Values marked ‘mutant’, ‘mixed’ and ‘wild type’ (or other variations) are converted to amino acids in the intermediate template using a frequently updated lookup table of values. Mixed resistance genotypes are converted to a value of sensitive genotype/resistant genotype (e.g. K/T for pfcrt_76_SNP_AA).

4.3.4 Copy number polymorphism

One copy number value per sample is stored in the database. If multiple copy number estimates per sample (replicates) are presented in the dataset, the mean is calculated in the intermediate template file for export to the database.

5. Data cleaning

Data are checked for **inconsistencies**, **unexpected values** and **missing values**. If identified, a query is returned to the data contributor for clarification. Resubmitted, corrected, and/or transformed missing values will be used to update the Data

Repository. If corrections cannot be made, the inconsistencies and unexpected results are transformed to missing values.

In order to transform all data sets into a consistent format for potential combined analysis, variables are constrained within limits set by the analysis program.

5.1 Inconsistencies and unexpected values

The following data checks are conducted on all studies and a list of inconsistencies and unexpected values is generated for the submitter.

I. Subject age < 0.08 years old
II. Subject age > 120 years old
III. Date of inclusion or sample collection < 1980
IV. Date of inclusion or sample collection > current year
V. Date of sample collection < date of inclusion
VI. Multiple countries per site
VII. Unexpected genotypes
    - Unexpected genotypes vary by marker and are based on a lookup table of expected values and ranges. These will be regularly updated to include new information.

5.2 Missing values

The following variables are checked for missing values and a query generated for the submitter.

I. Patient or sample identifier
II. Date or year of inclusion
III. Date or year of sample collection
IV. Age (optional)
V. Missing genotypes
    - Missing genotypes are designated as either 'Not genotyped' or 'Genotyping failure' in the Molecular database. Genotype fields left blank will generate a query for the submitter to choose one of the two missing options.

6. Methods of prevalence analysis

6.1 Analysis population

Prevalence estimates are calculated using genotyped samples from cross-sectional surveys or day 0 samples from clinical trials. Samples taken after initiation of treatment in a clinical trial are not used to calculate prevalence. These data are also stored in the Data Repository, and will be used in future analyses comparing pre- and post-treatment genotypes and marker selection.

Samples within a study are grouped into a geographical plus temporal unit for prevalence analysis. The geographical unit is the **study site**, as defined by the contributor. This definition includes the number of sites within a study, the geographic locations of these sites, and the assignment of samples to each site. Site groupings should represent the smallest biologically relevant geographic unit that comprises the area (s) sampled.

For each site, samples collected **within two years** are grouped to form the unit of prevalence analysis. Samples from sites spanning more than two years will be split into two-year groups for analysis and display. This temporal grouping is intended to combine samples taken during two consecutive seasons in many clinical trials, while differentiating samples collected over many years in a longitudinal trial or survey.

Although all resistance markers in a dataset are imported into the Data Repository, analysis is performed on a subset of markers and their expected genotypes. Variable length repeat data are stored but not currently analysed for reporting and display. The list of markers and genotypes analysed in the Study Report is given in Annex B and the list of markers, genotypes and resistance status presented in WWARN Explorer is given in Annex C. **These lists are preliminary, and will expand shortly** to include other *P. falciparum* and *P. vivax markers.* They will also be continuously updated to incorporate new resistance markers and genotypes as necessary.

6.2 Calculation of sample size

Sample size is calculated as the number of unique samples successfully genotyped for any particular marker. Samples without a genotype are labelled as either 'Not genotyped' or 'Genotyping Failure' and are excluded from sample size.

6.3 Calculation of SNP prevalence

Prevalence of a particular molecular resistance genotype is calculated as the number of samples with the resistance genotype divided by the sample size for that marker (see Section 6.2). Prevalence is calculated for 'pure' resistance genotypes such as *pfcrt* 76T and for 'mixed' resistant and sensitive genotype infections such as *pfcrt* 76K/T.

6.4 Calculation of SNP haplotype prevalence

Haplotypes consisting of several SNPs, such as the *pfdhps* double mutant (*pfdhps* 437G + 540E), are considered 'pure' resistant if all SNPs in the combination are recorded as 'pure' resistant genotypes. Haplotypes are considered 'mixed' resistant if up to one SNP in the combination is recorded as 'mixed'. This is a conservative designation, because when more than one SNP in a combination is recorded as 'mixed', it is impossible to know the 'linkage phase', that is, whether the full purely resistant haplotype occurs in the infection or whether multiple clones, some of each

with a different SNPs in the haplotype but none with the full resistant haplotype, are present[3].

A list of haplotypes and classifications in the current Data Repository are given in the following table. This list will be updated as new resistance haplotypes are identified.

Table 1. Examples of resistance haplotypes and classification as 'mixed' or 'pure' included in the WWARN Data Repository. For a full list see Annex C.

Haplotype combination	Classification
***pfdhfr* triple mutant**	
pfdhfr 108N + 51I + 59R	Pure triple mutant
pfdhfr 108N + 51I + 59C/R	Mixed triple mutant
pfdhfr 108N + 51N/I + 59C/R	Not a triple mutant
***pfdhps* double mutant**	
pfdhps 437G + 540E	Pure double mutant
pfdhps 437G + 540K/E	Mixed double mutant
pfdhps 437A/G + 540K/E	Not a double mutant
***pfdhfr/pfdhps* quintuple mutant**	
pfdhfr 108N + 51I + 59R + *pfdhps* 437G + 540E	Pure quintuple mutant
pfdhfr 108N + 51I + 59C/R + *pfdhps* 437G + 540E	Mixed quintuple mutant
pfdhfr 108N + 51I + 59R + *pfdhps* 437A/G + 540E	Mixed quintuple mutant
pfdhfr 108N + 51I + 59C/R + *pfdhps* 437A/G + 540E	Not a quintuple mutant
pfdhfr 108N + 51I + 59R + *pfdhps* 437A/G + 540K/E	Not a quintuple mutant

6.4 Calculation of prevalence in copy number polymorphisms

Raw copy number estimates are stored in the WWARN Molecular Data Repository (see Section 4.3.4 for treatment of datasets providing replicate values). Criteria for acceptance of copy number estimates varies among investigators[4,5,6]. These cutoffs are currently not implemented in the WWARN Data Repository. Rather, all estimates reported in the dataset are assumed to have met the submitter's quality criteria.

[3] Hastings *et al* 2010. A comparison of methods to detect and quantify the markers of antimalarial drug resistance. *Am J Trop Med Hyg.* 83:489-495.

[4] Price *et al* 2004. Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. *Lancet.* 364:438-447.

[5] Nair *et al* 2008. Adaptive Copy Number Evolution in Malaria Parasites. *PLoS Genet.* 4: e1000243. doi:10.1371/journal.pgen.1000243

[6] Griffing *et al* 2010. pfmdr1 amplification and fixation of pfcrt chloroquine resistance alleles in Plasmodium falciparum in Venezuela. *Antimicrob Agents Chemother.* 54:1572-1579.

For the purpose of reporting and analysis of prevalence, copy number polymorphisms are grouped as follows. These groupings may be subject to change as necessary for newly identified loci exhibiting copy number polymorphism.

Table 2. Copy number polymorphism categories for analysis in WWARN Data Repository and corresponding raw values

Category	Range of values
Copy number = 1	0.50 – 1.49
Copy number = 2	1.50 – 2.49
Copy number > 2	> 2.50
No data	< 0.49

6.5 Sub-category calculation of resistance marker prevalence

6.5.1 Age groups

When patient ages are provided in a dataset, marker prevalence is calculated within age groups for reporting and display, according to the following categories:

< 1 year
1-4 years
5-12 years
> 12 years

6.5.2 Clinical outcome

Future versions of the Molecular DMSAP will include calculations of marker prevalence at day 0 in subjects exhibiting variation in clinical efficacy including categories of parasite clearance (to be determined) and treatment outcome (treatment failure versus adequate clinical and parasitological response; see Clinical DMSAP).

7. Study Report

7.1 List of components in the Molecular Study Report

After a dataset has been uploaded and transformed, an automated Study Report is produced using the standardized data format. An example will soon be available on the WWARN website.

The Study Report contains the following:

1. **A basic description of the study** including the total number of participants, sites, and list of reported molecular markers.
2. **Information on which source variables** were extracted.

3. **Data transformation:** transformations of data (Section 4.3) to fit the WWARN format of coding variables are provided in an annexed list of original and transformed values.
4. **Systematic audits**
   4.1. **Data consistency**: the numbers and frequencies of unexpected results (Section 5.1) are listed in a table, with an annexed, detailed list of cases (patient or sample ID, date, patient- or sample-specific data).
   4.2. **Missing data**: a table presenting the number and frequency of missing values with an annexed, detailed list of cases (patient or sample ID, date, missing data).
5. **Prevalence results**
   5.1. **Sample profile**: displaying the total number of included patients and samples, by study site plus time window.
   5.2. **Baseline characteristics**: displaying the median age and age range (if included) and date ranges for inclusion and sample collection.
   5.3. **Prevalence of resistance genotypes (summary graphs):** percentage of pure and mixed resistance genotypes (for SNPs) or percentage of multi-copy samples (for copy number polymorphisms) at each site are displayed in graphs.
   5.4. **Prevalence of resistance genotypes (detailed tables):** sample sizes and proportions of all reported genotypes for each marker at each site and two-year time window are presented in tables. Breakdown of sample size and proportion by age group is included when ages are present in the dataset.

   See Annex B for a list of markers and genotypes currently included in prevalence calculations. SNP haplotype prevalence is not displayed in the Study Report.

7.2 Justification of data correction

Sample profiles, baseline characteristics, and prevalence tables are all presented in the report using both auto-corrected data and the original non-modified data. Prevalence graphs are presented using only auto-corrected data. Auto-correction transforms the unexpected values described in Section 5.1 into missing values. In the unchanged mode, all source values sent by the data submitter, which might include a patient age or date out of range, are unchanged and included in the analysis and display.

The submitter will receive a spreadsheet to allow correction of unexpected data points and missing values if desired. Displaying outputs with both auto-corrected and unchanged data allows the data submitter to readily judge the impact of auto-correction. If auto-correction causes limited or no changes to the prevalence results or baseline characteristics the submitter may choose not to correct the data with patient files. Although the report will contain both sets of data, the WWARN Data

Repository will utilise only the submitter-corrected data and/or auto-corrected data for future study analyses.

8. WWARN Explorer

The WWARN Explorer is an online, open-access tool which allows users to perform custom queries of summarised studies and visualise the results using dynamic interactive maps.

8.1 Selected data for display in WWARN Explorer

Imported variables from each study are used to derive a set of standardised calculations that define sample size and prevalence estimates as described in Section 6. See Annex C for a list of markers that are currently available for display in WWARN Explorer.

8.2 WWARN Explorer outputs

WWARN Explorer will display a brief overview (small window) and a detailed summary (large window) for each site plus time window in a study.

The small window displays:

I. Basic study information (study title, country, year)
II. Sample size (n) and prevalence (%) of main molecular resistance markers
III. Prevalence graph of main molecular resistance markers/haplotypes

NOTE: Because of the potentially large range of markers available for display, a small subset of 'main' resistance markers and haplotypes from each study will be automatically selected from a pre-set list for presentation in the summary.

The large window displays:

I. Basic study information (study title, location, country, year)
II. Additional study information (investigator, acknowledgments, link to publication when available)
III. Inclusion criteria (age range, sampling scheme, symptomatic status)
IV. Sample source (clinical trial, community survey)
V. Detailed location map
VI. Interactive prevalence graphs of all resistance markers and haplotypes including display by age category, if available
VII. Sample size per resistance marker/haplotype

8.3 Approval of results

Data submitters are able to review results obtained following the WWARN Molecular DMSAP in a comprehensive Study Report. **Only after the submitter's approval is given will the study outputs be made available on WWARN Explorer.** Data submitters may request screenshots of their data, as they will appear on WWARN Explorer, before approval.

8.4 Marker selector toolbar

The current version of WWARN Explorer includes a Treatment Selector toolbar for filtering of displayed studies. A similar Marker Selector option will be included in future versions to display studies containing data on particular resistance markers.

9. DMSAP versioning

The Molecular DMSAP has four main elements that are subject to change in future versions: a set of variables, lists of resistance markers, a data management plan, and a description of outputs.

9.1 A set of variables (the Molecular Dictionary)

The set of variables imported from the data submitters - the molecular dictionary - define the types of analysis that will be possible with the dataset. The current DMSAP version 1.0 focuses on the essential variables necessary to define prevalence of resistance markers. It does not currently capture additional variables, including:

- complexity or multiplicity of infection;
- allele frequencies or major versus minor alleles; or
- genotyped microsatellites in regions flanking resistance markers.

Additional variables may be included in future version of the Molecular Dictionary and DMSAP.

9.2 Resistance markers for prevalence analysis

The lists of resistance markers used in prevalence analysis in the Study Report and in WWARN Explorer will be updated regularly to remain current, at intervals more frequent than DMSAP versioning. Up-to-date versions of these lists will be available on the WWARN website.

9.2 Management of missing and unexpected values

In the current version, missing and unexpected values are omitted from analysis. In future versions, multiple imputations or other strategies may be applied to minimise data loss.

9.3 Outputs

The outputs from version 1.0 are described in the Study Report (Section 7) and the WWARN Explorer (Section 8). New outputs under consideration for future versions include non-map based visualisations, "cross-modular" analyses of studies where molecular along with *in vivo*, PK, and/or *in vitro* data have been documented for individual patients, as well as outputs using other marker types.

- "Sequence view" of SNPs in genetic loci associated with resistance
- Comparison of resistance marker prevalence by clinical outcomes (see Section 6.5.2)
- Comparison of parasite clearance and treatment efficacy among infections with sensitive versus resistant markers at day 0
- Effect of treatment on selection of resistance markers in recurrent infections
- Relationship between resistance markers and *in vitro* drug sensitivity
- Variable length repeats associated with resistance
- Multiplicity or complexity of infection

9.4 Reporting and pooled analysis of individual patient data

Reports and analyses created by WWARN may be published or otherwise made publicly accessible, including through the WWARN website. Such reports will not focus on individual studies, instead aggregating data from many studies to produce the WWARN summary reports. WWARN reports will acknowledge all contributing studies without identifying individual authors.

In addition, WWARN may convene ad hoc, collaborative working groups of data contributors to conduct pooled analyses of specific scientific questions. Pooling individual datasets from several molecular studies conducted in different locations and/or at different times would yield useful indicators of temporal or spatial trends as the basis for an effective, early warning system of emerging or spreading resistance. Pooled analyses and linkage between molecular and clinical data will be particularly important for validating relatively low frequency candidate resistance markers as predictors of clinical measures of resistance. Moreover, these analyses would also address broader questions such as how well a particular marker can predict treatment efficacy on a population level and how marker prevalence varies

by age group, symptomatic status or other epidemiological parameters. All contributors whose studies are included would have the option to join the working group and if publication is anticipated, the working group would determine authorship.

A separate, future *Pooled Dataset Statistical Analysis Plan* will describe the processes and methods for such analyses, although some elements may be included in future Molecular DMSAP versions.

10. Conclusion

This data management and statistical analytical plan is an evolving document that aims to stay current as new information on molecular markers of drug resistance are identified and validated. WWARN encourages feedback on these issues and will endeavour to incorporate suggestions into future versions or bring major issues into a wider forum for open discussion. Comments should be directed to molecular@wwarn.org.

11. References

1. Plowe CV, Roper C, Barnwell JW, Happi CT, Joshi HH, Mbacham W, Meshnick SR, Mugittu K, Naidoo I, Price RN, Shafer RW, Sibley CH, Sutherland CJ, Zimmerman PA, Rosenthal PJ. 2007. World Antimalarial Resistance Network (WARN) III: molecular markers for drug resistant malaria. *Malar J.* 2007 6:121.

2. Picot S, Olliaro P, de Monbrison F, Bienvenu AL, Price RN, Ringwald P. 2009. A systematic review and meta-analysis of evidence for correlation between molecular markers of parasite resistance and treatment outcome in falciparum malaria. *Malar J.* 8:89

3. Hastings IM, Nsanzabana C, Smith TA. 2010. A comparison of methods to detect and quantify the markers of antimalarial drug resistance. *Am J Trop Med Hyg.* 83:489-495.

4. Price RN, Uhlemann AC, Brockman A, McGready R, Ashley E, Phaipun L, Patel R, Laing K, Looareesuwan S, White NJ, Nosten F, Krishna S. 2004. Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. *Lancet.* 364:438-447.

5. Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, Newton P, Nosten F, Ferdig MT, Anderson TJ. 2008. Adaptive Copy Number Evolution in Malaria Parasites. *PLoS Genet.* 4: e1000243. doi:10.1371/journal.pgen.1000243

6. Griffing S, Syphard L, Sridaran S, McCollum AM, Mixson-Hayden T, Vinayak S, Villegas L, Barnwell JW, Escalante AA, Udhayakumar V. 2010. pfmdr1 amplification and fixation of pfcrt chloroquine resistance alleles in Plasmodium falciparum in Venezuela. *Antimicrob Agents Chemother.* 54:1572-1579.

12. Glossary

Download the supporting glossary from the WWARN website.

Annex A: Molecular dictionary

Study and location data (extracted from Study Site Questionnaire)

- Study ID
- Investigator
- Study Label
- Countries
- Sites

Baseline patient data

- Unique patient identifier
- Date of inclusion and days or dates of follow up visits
- Patient age or date of birth (optional)

Molecular marker data

- Locus name
- Locus position
- Marker type
- Molecule type (if applicable)
- Value (amino acid, nucleic acid, copy number or fragment size)

Annex B: Resistance markers and genotypes analysed in the Study Report

Resistance Marker	Genotypes					
pfcrt72	C	S	C/S			
pfcrt73	V					
pfcrt74	I	M	I/M			
pfcrt75	N	E	N/E			
pfcrt76	K	T	K/T			
pfcrt220	A	S	A/S			
pfdhfr16	A	V	A/V			
pfdhfr50	C	R	C/R			
pfdhfr51	N	I	N/I			
pfdhfr59	C	R	C/R			
pfdhfr108	S	N	S/N			
pfdhfr164	I	L	I/L			
pfdhps436	S	A	F	S/A	S/F	A/F
pfdhps437	A	G	A/G			
pfdhps540	K	E	K/E			
pfdhps581	A	G	A/G			
pfdhps613	A	S	T			
pfmdr186	N	Y	N/Y			
pfmdr1184	Y	F	Y/F			
pfmdr11034	S	C	S/C			
pfmdr11042	N	D	N/D			
pfmdr11246	D	Y	D/Y			
pfmdr1 CN	1	2	> 2			

Annex C: Resistance markers, genotypes and resistance status for display in WWARN Explorer

Locus Name	Locus Position	Marker Type	Genotype	Label	Status
Pfcrt	76	SNP	T	pfcrt 76T	Pure
Pfcrt	76	SNP	K/T	pfcrt 76T	Mixed
Pfcrt	220	SNP	S	pfcrt 220S	Pure
Pfcrt	220	SNP	A/S	pfcrt 220S	Mixed
Pfdhfr	16	SNP	V	pfdhfr 16V	Pure
Pfdhfr	16	SNP	A/V	pfdhfr 16V	Mixed
Pfdhfr	50	SNP	R	pfdhfr 50R	Pure
Pfdhfr	50	SNP	C/R	pfdhfr 50R	Mixed
Pfdhfr	51	SNP	I	pfdhfr 51I	Pure
Pfdhfr	51	SNP	N/I	pfdhfr 51I	Mixed
Pfdhfr	59	SNP	R	pfdhfr 59R	Pure
Pfdhfr	59	SNP	C/R	pfdhfr 59R	Mixed
Pfdhfr	108	SNP	N	pfdhfr 108N	Pure
Pfdhfr	108	SNP	S/N	pfdhfr 108N	Mixed
Pfdhfr	164	SNP	L	pfdhfr 164L	Pure
Pfdhfr	164	SNP	I/L	pfdhfr 164L	Mixed
Pfdhps	436	SNP	C	pfdhps 436C	Pure
Pfdhps	436	SNP	S/C	pfdhps 436C	Mixed
Pfdhps	436	SNP	A/C	pfdhps 436C	Mixed
Pfdhps	436	SNP	F	pfdhps 436F	Pure
Pfdhps	436	SNP	S/F	pfdhps 436F	Mixed
Pfdhps	436	SNP	A/F	pfdhps 436F	Mixed
Pfdhps	436	SNP	C/F	pfdhps 436C/F	Pure
Pfdhps	437	SNP	G	pfdhps 437G	Pure
Pfdhps	437	SNP	A/G	pfdhps 437G	Mixed
Pfdhps	540	SNP	E	pfdhps 540E	Pure
Pfdhps	540	SNP	K/E	pfdhps 540E	Mixed
Pfdhps	581	SNP	G	pfdhps 581G	Pure
Pfdhps	581	SNP	A/G	pfdhps 581G	Mixed
pfmdr1	86	SNP	Y	pfmdr1 86Y	Pure
pfmdr1	86	SNP	N/Y	pfmdr1 86Y	Mixed
pfmdr1	184	SNP	F	pfmdr1 184F	Pure
pfmdr1	184	SNP	Y/F	pfmdr1 184F	Mixed
pfmdr1	1034	SNP	C	pfmdr1 1034C	Pure
pfmdr1	1034	SNP	S/C	pfmdr1 1034C	Mixed
pfmdr1	1042	SNP	D	pfmdr1 1042D	Pure
pfmdr1	1042	SNP	N/D	pfmdr1 1042D	Mixed
pfmdr1	1246	SNP	Y	pfmdr1 1246Y	Pure
pfmdr1	1246	SNP	D/Y	pfmdr1 1246Y	Mixed
Pfdhps	613	SNP	S	pfdhps 613S	Pure
Pfdhps	613	SNP	A/S	pfdhps 613S	Mixed
Pfdhps	613	SNP	T	pfdhps 613T	Pure
Pfdhps	613	SNP	A/T	pfdhps 613T	Mixed
Pfdhps	613	SNP	S/T	pfdhps 613S/T	Pure
pfmdr1		CN	1	pfmdr1 CN	pfmdr1 CN=1
pfmdr1		CN	2	pfmdr1 CN	pfmdr1 CN=2
pfmdr1		CN	> 2	pfmdr1 CN	pfmdr1 CN>2
pfdhps,pfdhps	437,540	SNP	G,E	dhps double	Pure
pfdhps,pfdhps	437,540	SNP	G,K/E	dhps double	Mixed

pfdhps,pfdhps	437,540	SNP	A/G,E	dhps double	Mixed
pfdhfr,pfdhfr,pfdhfr	108,51,59	SNP	N,I,R	dhfr triple	Pure
pfdhfr,pfdhfr,pfdhfr	108,51,59	SNP	S/N,I,R	dhfr triple	Mixed
pfdhfr,pfdhfr,pfdhfr	108,51,59	SNP	N,I,C/R	dhfr triple	Mixed
pfdhfr,pfdhfr,pfdhfr	108,51,59	SNP	N,N/I,R	dhfr triple	Mixed
pfdhfr,pfdhfr,pfdhfr,pfdhps,pfdhps	108,51,59,437,540	SNP	N,I,R,G,E	dhfr/dhps quintuple	Pure
pfdhfr,pfdhfr,pfdhfr,pfdhps,pfdhps	108,51,59,437,540	SNP	N,I,C/R,G,E	dhfr/dhps quintuple	Mixed
pfdhfr,pfdhfr,pfdhfr,pfdhps,pfdhps	108,51,59,437,540	SNP	N,I,R,A/G,E	dhfr/dhps quintuple	Mixed
pfdhfr,pfdhfr,pfdhfr,pfdhps,pfdhps	108,51,59,437,540	SNP	S/N,I,R,G,E	dhfr/dhps quintuple	Mixed
pfdhfr,pfdhfr,pfdhfr,pfdhps,pfdhps	108,51,59,437,540	SNP	N,N/I,R,G,E	dhfr/dhps quintuple	Mixed
pfdhfr,pfdhfr,pfdhfr,pfdhps,pfdhps	108,51,59,437,540	SNP	N,I,R,G,K/E	dhfr/dhps quintuple	Mixed