

Pooling strategy of clinical data

Abraham Yeh, Xiaohong (Grace) Zhang, Shin-Ru Wang, Novartis Pharmaceuticals Corporation, East Hanover, NJ

ABSTRACT

Pooling of clinical data is used by all pharmaceutical companies. Submission of a new compound to Health Authorities is one of the main reasons to pool data. However, there are multiple other needs for pooling data such as Development Safety Update Report (DSR), Risk Management Plan (RMP), publications, etc.

The concept of pooling data is neither new nor defined by a pre-determined set of rules or methodologies. Inevitably pooling requirements are project/compound specific and are driven by the objective(s) of the pooled data.

There is no obvious right or wrong way to pool data. The intention of this paper is to provide guidance on the creation of pooled datasets with general information, typical pitfalls and key points in order to obtain a robust pool. It will cover topics such as planning, data standards, programming strategies, validation and documentation.

1 INTRODUCTION

The concept of integrating or pooling data is neither new nor defined by a pre-determined set of rules or methodologies. Even with a well-defined data standard in place a certain amount of variability is to be expected. This is compounded if the integration includes studies of special interest, studies run and reported by local country and studies developed across multiple phases of the clinical development plan.

As the process of pooling data from multiple clinical trials is often labor intensive and can be error prone, a clear strategy with clear goals defined as early as possible in the process is essential.

The pooling of data can have a number of benefits

- Pooling data from different studies can improve the precision of an incidence estimate (i.e., narrow the confidence intervals by enlarging the sample size).
- Pooling can be done across all phases, within a compound, or across compounds and provide the larger database that will allow explorations of possible drug analysis.

For these reasons health authorities (HA) and Pharmaceutical Industries can be very interested in pooling, especially in terms of safety where pooled data may allow the identification of rare and uncommon safety signals. The quality of the pool is however dependent on the individual trials used as pooling may obscure real potentially meaningful differences between studies.

When preparing pooling strategy, it is important to understand the rationale of pooling data for specific study sets.

- It is often most appropriate to combine data from studies that are of similar design i.e. dose, duration, control and population (could also be studies with same concomitant medication use, same visits intervals).
- If pooling data from studies it may be critical to account for exposure duration and to look at time dependent events in relation to this.

There are a number of HA/ICH guidances that discuss integration of data across trials but there are rarely any fixed rules. Therefore, for each integrated analysis it is important to ask the question first of all - "do we need to pool at all?" As guidances change over time, it is important that teams keep up to date with the latest industry information.

2 PLANNING

A plan should be developed as early as possible and is needed to focus the purpose, objectives and requirements of any integration (not just for a submission).

The plan should be used to define which studies and which data should be pool to add value in a pooled analysis for any given deliverable from both an efficacy and a safety perspective.

Key team members (Drug Regulatory Affairs, clinical, Drug Safety and Stats) should be involved in planning.

Timelines, resources and responsibilities need to be discussed and agreed to ensure a successful execution.

Data standards should be used within projects (and across projects where feasible) to enable more efficient pooling.

2.1 DO WE NEED TO POOL?

It is important that any strategy focuses on the purpose, objectives and requirements of the integration and therefore this is a crucial question and one that should not be overlooked. For example, if the study characteristics are very different it will not be appropriate to pool data for safety or efficacy and side-by-side presentations could be considered. In terms of efficacy pooling, it can obscure a meaningful effect if the studies to be pooled have markedly different results. In this situation pooled data will not add value to the information about the compound.

2.2 ASSESSING THE PURPOSE OF THE POOL

When developing a plan or strategy for a given data pool it is important to consider how the pool will be used not only for the current deliverable but also for any future reporting. The assessment may help to determine if:

- The pool is developed as a stand alone pool to meet the needs of a specific deliverable
- The pool will be periodically updated (appended to) and re-run – e.g. to support DSURs

It is common for a project to develop a number of data pools to support a variety of activities that require pooled analysis. It may be that a publication is developed to look at a data pool that relates to a specific data cut-off within a program (e.g. locked studies of certain length at a certain time point). In this case a static data pool may be desired that will not be updated. This pooled database may however form the base of a subsequent wider ranging pool. Where multiple pools have been developed, centralized tracking of the data pools available is essential to allow for easy location

2.3 WHEN SHOULD PLANNING START?

Planning should start early to understand the likely future pooling needs and should be considered in both the Clinical Development Plan and Safety Profile Plan.

2.4 THE “CORE” TEAM

It is likely that many of the milestones relevant to integrated analysis require input from a wide scope of team members.

2.5 OPERATIONAL CONSIDERATIONS

When planning the details of how and when to execute an integration/pooling exercise there are a number of logistic and operational factors that should be considered:

- Timelines – The timelines for the key milestones are often defined by the wider project team and depend on the deliverable.

2.6 WHAT SHOULD AN INTEGRATED DATABASE CONTAIN?

Studies may have to be pooled by indication, development program or across projects or therapeutic areas. They may be pooled by treatment comparator. These and which data (domains) are required to be pooled are primarily dependent on the deliverable and objectives of any analysis. Below is a list of data which are commonly pooled, however, some deviations may be required or reasonable depending on the focus of the analysis.

2.6.1 SAFETY

As mentioned earlier it is quite common for safety data to be pooled. However, this should not be done without some thoughts. The first step is to decide what studies the integrated database should contain as outlined above. The second step is to decide the relevant information to pool, this will commonly include:

- Duration of treatment exposure.
- Adverse events, SAEs (including deaths) and ADRs.
- Specific defined risks of interest which are often defined based on adverse event SMQs or NMQs but could also contain data other than AEs. These are often especially relevant for the RMP.
- Lab values, often with a focus on parameters that represent a specific safety marker significant to the indication or a drug class. There is often special interest in lab abnormalities in relation to clinically significant abnormality ranges.

Further areas of potential interest, depending on indication, compound characteristics and data captured may include:

- Vital signs.
- Specifically solicited events captured outside of standard panels (e.g. immunogenicity problems).
- ECG/EKG, x-ray, other relevant tests/assessments (e.g. MRI or neurological examinations).
- Adjudicated endpoints such as deaths, hospitalizations, cardiovascular events, liver abnormalities or other endpoints that may be specific to an indication or drug class.
- Concomitant medications: all or only selected subset, for example indication-related concomitant medications.
- Patient disposition.
- Medical history – in particular may be relevant to identify defined risks of interest (Risk History).

The default should not be to pool everything but to do so when it adds value. For example in a small project if there are only 2 studies - 1 small proof of concept and 1 relatively small phase III study, there may be limited worth in expending the time and effort to pool one or more data domains.

2.6.2 EFFICACY

The selection of which efficacy data to be pooled is very much dependent on the project. Selection should primarily focus on key endpoints that support the overall objectives of a program or submission or that may be used to

make label or commercial claims or add clarity or additional weight to results seen across individual studies of a development program.

Frequently one of the objectives of pooling is to facilitate the identification of subgroups of patients who may get the most benefit. It often requires pooling of this data to increase the patient numbers in order to make any subgroup analysis feasible. Subgroups for investigation should be pre-planned. With the increased business focus on specialist care, interest in pooling to analyze patient subgroups within a broader population is increasing.

2.7 DATA STANDARDS

Whilst pooling may be driven by a specific need, the implementation of a strong data standard is one of the key elements that can drive efficiencies when pooling.

2.8 KEY POINTS FOR EXECUTION

- Execution of pooling should be early enough to allow the programming of datasets and any analysis that is required to take place.
- Documentation is important and should be detailed and stored in the document management system. In addition to CSPD or MAP modules specific tracking sheets may be used.
- Validation is important and may be based on reports or by study checks.
- Pooled data sets can be created from existing raw or derived datasets. The decision to use one or other will depend on the availability of standard data, availability of derived data, common endpoints, etc.

2.9 MAPPING STUDY LEVEL DATA TO POOL LEVEL

The development of a specification to define how data should be mapped from the study to pooled level can be a complex exercise, especially if the pool includes a large number of studies reported using a variety of data standards.

The exercise of mapping becomes more important as studies switch between data standards. It is likely that at some stage any legacy data will need to be mapped to the latest data standard to allow for future reporting.

Though this may begin with a review of the analysis plans developed for each study, there may be existing SAS programs available to help expedite this process.

Developing SAS reports that compare the raw and analysis data and metadata for a given study with the final pooled dataset structure may be a more efficient and less error prone approach than performing this activity by manual review. These reports can be developed to compare:

- The variables required in pooled dataset vs those available in the study level datasets.
- The attributes of variables required in a pooled dataset vs those available in the study level datasets e.g. label, length, type (numeric value, character, date or time, numeric or character codelists) etc. If done in a not thoroughly planned manner, integrating variables of different lengths, format can lead to issues such as truncation, mis-coding of categorical fields etc.
- The values contained in variables or respective code lists applied to a given variable. In particular this may highlight where the study level variable contains common codes but where the codes correspond to a different value when decoded.
- Lab parameters and respective codes collected across studies. Some lab parameter codes may need to be

mapped to a common standard.

References to some useful papers discussing these issues and possible solutions to facilitate mapping are included in the Useful References section later.

2.10 DISCUSSION FOR DIFFERENT APPROACHES: RAW VS. DERIVED

There are 2 main sources for pooling data, raw and derived (analysis) datasets. The decision to use either raw or derived datasets can be based on a variety of factors (see Sections 2.10.1-2.10.2). A model that mixes data sources may be deemed appropriate once these factors are taken into consideration (see Section 2.10.3).

The following sections discuss some of considerations that should be taken when deciding on a strategy.

2.10.1 POOLING FROM DERIVED/ANALYSIS DATA

In many cases teams may prefer to use derived data if these are already available at the study level when pooling activities plan to begin relative to each study lock date and the submission date, and provided the studies share a common data standard. This is of particular relevance for pooling of efficacy data which often includes a large number of derived endpoints.

Utilizing derived (analysis) datasets generated for CSR reporting can reduce work in re-deriving common endpoints or data points and help ensure consistency in derivations with the CSR.

However care should be taken to ensure that all derivations and algorithms are consistent across individual studies and with those needed for any pooled analysis. This assessment should not only focus on algorithms defining the derivation rules of given endpoints, for example AUC (area under the curve), but also on windowing or imputation rules for endpoints that handle missing data, such as LOCF.

In addition, analysis datasets may be pooled to include only data that is relevant e.g. treatment emergent.

Pooling from derived data requires the availability of analysis datasets (either final or dry run) to facilitate execution.

2.10.2 POOLING FROM RAW DATA

Pooling from raw data may be a suitable option in cases where there is significant variability in the data structures, standards or endpoint derivations across studies.

The use of raw data for a pooled database will inevitably require the reprogramming of endpoints, variables and derivations already programmed at a study level, however, this may be a necessity if endpoints in the final pooling differ from those defined for one or multiple studies.

One distinct advantage of this method is that if endpoints, baseline definitions etc. are recreated this will maximize the consistency across trials.

2.10.3 POOLING FROM RAW AND DERIVED DATA

In special situations it might happen that raw and derived datasets together need to be pooled.

Such situations could arise when:

- some 'standard' studies have the required endpoints available in derived datasets but some 'non-standard' studies do not: we could use the derived datasets for 'standard' studies and used the raw datasets for 'non-standard' studies.

- for an ongoing pool raw data might be used for the most recent study as the corresponding validated derived datasets are not yet available.

- in derived datasets not all patients are included, but for specific analyses we need all patients; in this case data for those missing patients could be taken from raw dataset.

2.11 VALIDATION

Validation is a critical element in the execution of any pooling exercise and should be based on a risk assessment and according to a validation strategy. As such, teams should prepare for this well in advance and this aspect should be covered in any execution strategy. It is vital that validation efforts are made according to the pooling strategy and carried out thoroughly.

Whilst the basic concepts for validation of programming are no different to any other programming or reporting activity, the pooling of data from multiple studies can be complex and error prone. Verification in addition to standard programming validation is strongly advised. The sections below provide examples of such checks.

2.11.1 REPORTS TO SUPPORT REVIEW

As pooling often requires the derivation of common derived variables to enable the mapping of data of similar times (visits), time points etc to a common reportable way, validation may also require a formal review to ensure that data is mapped as expected and required.

A simple example is the mapping of study visits and assessment time points. This would be traditionally handled by creating an analysis visit variable. In some cases the programming definition of the derivation may contain a more general rule stating the final analysis visit values and stating that study level visit and time points should be mapped to this.

In an example like this it is advised that the source visit and time point variables are retained in the final dataset. This would support traceability as well as allow a simple report to be written to allow a review of the mapping from source to the pooled analysis variable across all studies. This mapping matrix type report is a useful way to review that mapping has been implemented as expected. This may also help flag where mappings represent an equivalence rather than an exact mapping, for example, where a 15 min pre-dose, 30 min pre-dose, non-time specific pre-dose time points collected at the study level are mapped to "pre-dose" for the purpose of pooled data reporting.

The same principle can be applied to allow a review of the mapping of code lists from the source data to the equivalent code list for pooled reporting. Whilst it may not always be efficient or useful to retain the source variable in these cases, producing a reporting that summarizes source code list value and the eventual mapped value as part of the dataset programming can be useful to allow a review of the mapping for correctness and appropriateness.

The consideration of developing programmed reports to facilitate validation and review is an important one and can help increase quality and efficiency of review.

2.11.2 CROSS CHECKS BETWEEN POOLED AND STUDY LEVEL DATA

Once data are pooled it is important to ensure that the pooled data represent the data from individual studies consistently where expected. Use of tools to track the population counts and treatment exposure from individual trials (e.g. CSPD Section 8.2 Appendix 1) on an ongoing basis as trial databases lock is essential to support this cross check.

Inevitably pooling from raw data will result in reprogramming of the same data and endpoints produced for CSR analysis.

In these cases teams may consider performing suitable consistency checks to ensure that any re-derivations yield the same results as the study or that any discrepancies are due to known differences in algorithms.

Differences that may be identified could be documented in case questions arise regarding any inconsistencies between a pooled level report and the individual CSR summaries so this information is readily available.

One mechanism to support this additional level of QC is to develop programs that will produce pooled analysis in such a way that they can be easily subsetted to report the data from the pooled database for a single study only. These subsetted reports generated from the pooled database can then be compared to the equivalent trial level outputs.

3 CONSIDERATIONS AND POSSIBLE PITFALLS

We need to ensure formats and units are consistent across all data and that missing data is handling consistently. When pooling data care should be taken with the mapping of treatment data and visits and the use of consistent algorithm definitions. Consider specific situations like the issue of pooling cross-over studies and parallel group studies or including extension data. Adverse events and concomitant medications should be reported under the latest version of the relevant dictionary (MedDRA, WHO DRL). All mapping, specific algorithms, study specific information should be clearly documented in programming specifications documents.

3.1 GENERAL CONSIDERATIONS

As stated previously it is essential to have consistent data; all formats and units should be checked. In addition missing data should be handled consistently across all studies pooled.

3.1.1 MAPPING OF TREATMENTS

It is often the case in pooled data presentations that treatments are presented differently to those in individual studies. For example, the key treatment arms and doses are likely to be presented individually but treatment arms and regimens investigated for dose ranging, or exploratory studies may be grouped into an “other active treatment” type group.

Special care should be taken for studies with multiple periods of treatment where up-/down titrations as well as add-on therapies might have to be taken into account.

It is important that the required mappings are well documented.

3.1.2 MAPPING OF VISITS AND TIME POINTS

During the planning of mapping it is essential that any plans define how time points and visits will be pooled. Some considerations may include:

- Are pre-dose assessments in all pooled studies consistent? Different protocols may define pre-dose measures at different time points (e.g. 15 minutes pre-dose, 30 minutes pre-dose, pre-dose – with no specific time point). It may be useful to consider if there is a rationale to present all pre-dose assessments under one single pre-dose time point or if there is something to consider in the trial design that may mean that a pre-dose value at a given visit is influenced by the previous dose that means these values may not be directly comparable
- Are key reporting visits consistent across trials? Would it make sense to pool a week 24 visit with a week 26 visit, if planned visit differs across studies but essentially represent the same time point in the study (e.g. 6 months)?
- Is there value in pooling visits or time points that occur in only 1 or 2 of 20 pooled studies? This may be a necessity if endpoints are derived that utilize all data e.g. minimum or maximum post baseline value or change,

however, the inclusion of these values in a pooled dataset does not always mean that these infrequently occurring time points need to be reported, e.g. in by visit summaries

3.1.3 ALGORITHM DEFINITIONS

Though endpoints may be consistent across studies, the algorithms used to derive these may be dependent on a number of factors such as the length of the trial.

For example, a given endpoint may be analyzed using LOCF. Part of the algorithm for selection of the LOCF value may include a rule to cut-off data based on a number of weeks or days prior to the time point of interest e.g. LOCF is derived as the last non-missing value within X weeks of the primary end point. As the primary endpoint may vary between studies depending on the length, objectives and visit schedule of the trial, X may be variable across trials/studies. Once the data are pooled, does the LOCF need to be revised so a consistent algorithm is implemented?

Other common algorithm issues may include baseline definitions and other key covariates. These should all be considered.

It may be of value to note key algorithmic differences across studies while working on the study reporting to facilitate later planning for pooling.

3.1.4 CROSS-OVER DESIGN VS PARALLEL DESIGN

For many data domains it may not be appropriate to pool data from studies with cross-over designs with parallel design studies. This kind of pooling may be limited to exposure, disposition, AEs and some other key safety parameters.

Essentially a single period parallel group study can be viewed as one period cross over study. Variables to indicate period may be a critical for pooling, especially to facilitate the development of generating reporting programs.

Special attention is often needed if events happened in wash-out periods for cross over studies. For example:

- The number of days during the washout between periods will not be included in calculating the duration of exposure to study medication. For subjects exposed to more than one treatment, subjects are counted in each respective treatment to which they are exposed.
- AEs occurring within a treatment period (say period X) and during the washout prior to the start of the next treatment period (say period X+1) will be assigned to the treatment taken in that period (i.e. X).
- Some baseline values for cross over trials may be re-calculated prior to each period if there is a washout between each period.

3.1.5 EXTENSION AND INTERIM ANALYSIS

If extension data need to be pooled special attention to the related data should be considered; for examples:

(1)Treatment: on which treatment should patients who switch treatment from core to extension be considered?

(2)Baseline: for patients who switch treatment from core to extension, should their core baseline or end of core study as baseline be used?

(3)Handling of duplicate records (events starting in a core study and continuing to its extension) need careful consideration (like AEs, Concomitant medications...).

If interim analysis data is included in the pool, we should also think at the specific points like how to determine the cut-off point (include all data up to a specific date? a specific visit? include only patients who completed the study? etc.)

3.1.6 POOLING OF STUDY SITE/CENTER

Often statistical analyses may investigate the site/center as an interaction and covariate. Whilst at a study level it is relatively simple to identify unique centers/sites, once data are pooled this may not be such a trivial task. It is unlikely that site 1 in study A is the same as site 1 in study B. A unique center number is needed to identify the same center in various studies if site/center effects are to be analyzed with the pooled data. Some exploratory work may be needed to ensure these assignments are correct and consistent.

Additionally it may be possible to use different country/region splits in pooled data due to the larger subject numbers across studies. This should be considered when specifying the datasets.

3.1.7 CODING

When pooling it is often the case that studies in the pool were coded using different versions of a coding dictionary or even different coding dictionaries. This is especially relevant to MedDRA and WHO DRL dictionaries.

In general, for pooled analyses data should be reported under a consistent coding dictionary with the latest version available at the time of reporting by remapping based on lowest level code. For AE data using latest SMQ terms is required for RMP update.

If the pool includes a significant legacy of older studies, some recoding of data may be required to ensure data can be reported consistently. For very old studies in the legacy this may require translation from a foreign language or recoding to a common dictionary however this should be done by a specialists in those areas, not by the statistical and programming team.

Retaining the study level coded values may help in easily verifying differences between the CSR and pool reported summaries, however, care should be taken to ensure that these variables are clearly differentiated from those to be used for reporting (i.e. clear variable names and labels)

3.1.8 SUMMARIZING DATA – SUMMARY STATS FOR LABS AND VITALS SIGNS

When planning the reporting requirements for any pooled data it is important to define appropriate analyses to ensure that the results can be interpreted and are not misleading.

For tables reporting summary statistics in particular, it is important to assess if the pooled summary is the most appropriate way to present the data and whether the results can allow for a solid and valid interpretation.

This is particularly relevant to observations such as labs, vitals signs, ECGs and any other that measure physiological parameters that are likely to be influenced by a patients age, sex, race, disease state or other characteristics.

Summarizing mean values for a given lab parameter across a large group of patients in a variable population where values would be expected to vary depending on age, sex or race may not be appropriate. In these cases data are likely to have limited or no value.

Similarly, if a number of lab vendors have been used across the studies pooled, the sensitivity and calibration of the equipment may vary, as would the normal ranges. Again, summary statistics of such data may have limited interpretability.

Some statistical methods may be applied to normalize the lab data to a standardized range based on the original normal range however many of these methods assume a normal distribution of the data which may not be an ap-

appropriate assumption for this kind of data.

The most useful approach may be to summarize pooled data in relation to shift in and out of normal ranges set by the lab vendor(s) (e.g. shift tables) rather than generating summary statistics.

With any summary of data of this type it is important to ensure that all values are converted to and reported in a standard unit. Whilst in general, data at the project level may be reported to a defined set of standard units, there may be variations for trials reported to region specific SI/US units (e.g. Japan: CSRs of Japan studies usually use local original values with preferred units, with adjusting unit conversion). When pooling the laboratory data, confirmation of the units used is necessary and additional adjustment might be required.

For lab test, there might exist variety of parameter names in the similar items and checking the N of observations by study and parameter name might be worthwhile.

3.1.9 “STANDARD” METHODS TO CALCULATE ENDPOINTS

Some endpoints or parameters may be calculated based on a standard method of calculation that has a number of clinically accepted standards.

When pooling and reporting this data it is important to ensure that data is reported using a consistent method.

3.2 DATA AVAILABILITY

The availability of data for pooling can be a significant issue when preparing a pooled database.

4 CONCLUSIONS

The pooling of clinical data can be complex at times. It is one of the most challenging works for the programming team. If teams can utilize the advance planning with properly discussion across many different functions, this could be done well in high quality. We hoped this paper gave you some good insights and we certainly welcome all valuable feedbacks.

REFERENCES

Guidances

FDA Manual of Policies and Procedures, MAPP 6010.3 Rev. 1

ICH Topic E 9 Statistical Principles for Clinical Trials

Data pooling and integration

http://www.scsug.org/SCSUGProceedings/2010/Gerlach%20Data_Integration.pdf

<http://www.lexjansen.com/pharmasug/2004/applicationsdevelopment/ad10.pdf>

ACKNOWLEDGMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

We would like to acknowledge Darren Weston, Simon Walsh and Florence Buchheit for the contribution of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Abraham Yeh
Novartis Pharma
G240A, 100 West Promenade Boulevard,
East Hanover, NJ 07936
Work Phone: 1-862-778-8465
Fax: 1-973-781-2320
Email: abraham.yeh@novartis.com

Xiaohong (Grace) Zhang
Novartis Pharma
G220A, 110 West Promenade Boulevard,
East Hanover, NJ 07936
Work Phone: 1-862-778-3557
Fax: 1-973-781-2320
Email: xiaohong.zhang@novartis.com

Shin-Ru Wang
Novartis Pharma
G240B, 100 West Promenade Boulevard,
East Hanover, NJ 07936
Work Phone: 1-862-778-4937
Fax: 1-973-781-2320
Email: shinru.wang@novartis.com