Infrastructure development for building, maintaining and modeling indication-specific summary-level literature databases to support model-based drug development

**McDevitt H[1], Pillai G[1], Wu K[1], Ramakrishna R[1], Flesch GJP[1], Ette G[1],
Devidas S[2], Thatavarti R[2], and Buchheit V[1]**

[1]Novartis Pharma AG, Basel, Switzerland; and [2]GVK Biosciences Private Ltd, Hyderabad, India.

Background

- Several Pharma Modeling and Simulation (M&S) departments and M&S consulting companies have demonstrated the value of using models developed from summary level literature data in the public domain.[1–3]
- Health authorities and Payors (such as NICE) are also very interested in the comparison of new and existing therapies for healthcare cost containment.
- Any organization that intends using this kind of data systematically by multiple users needs to take time to carefully consider the processes and infrastructure to ensure reliability and quality.
- Wherever possible automation should be used to increase efficiency, especially in loading the data.
- At Novartis, an IT infrastructure that uses publicly available open source software is planned and will be presented in this poster.

Objectives

- To solicit feedback and discussion on the establishment of an infrastructure for building, maintaining and modeling summary-level literature data from the published literature publications.
- The value of establishing a common industry standard for recording such literature-based data with the potential for exchange is of particular interest.
- As the majority of the input data (*i.e.* published manuscripts) is in the public domain, it makes sense from an efficiency and economics perspective for interested parties to exchange databases built in particular disease areas rather than everyone recreating the same databases on an ongoing basis.

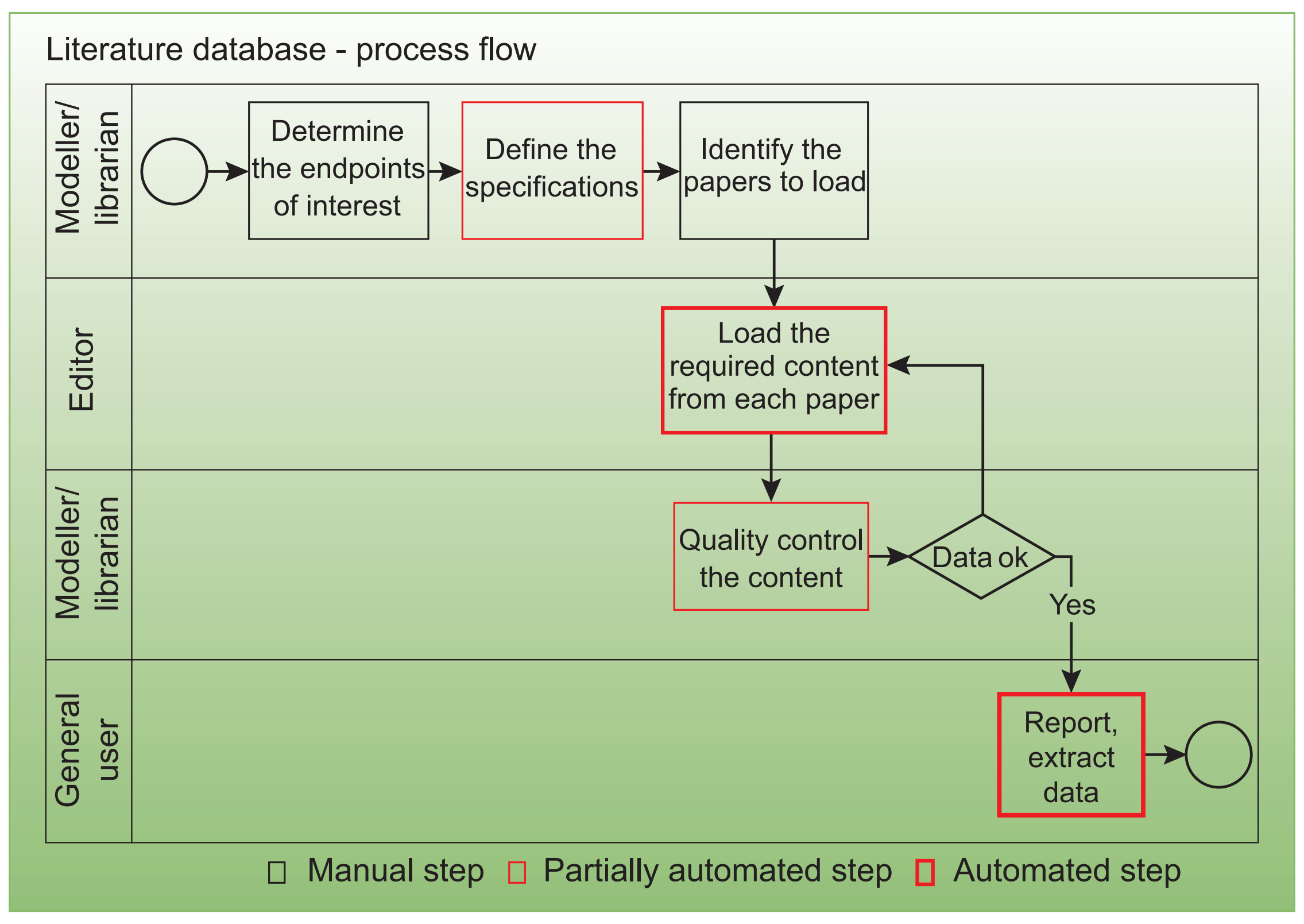
Methods

- To be successful three clear requirements can be defined:
  - The data need to be stored in a consistent standardized manner to make retrieval and further analysis possible in an intuitive way to the user.
  - A clear methodology needs to be established to be able to reliably record the literature data in a consistent manner.
  - The database needs to be updated on an ongoing basis to keep it current.
- These three steps together are necessary to ensure confidence in the database.
- **Figure 1** shows an overview of data acquisition and usage indicating where partial or complete automation is possible.

Data storage approaches

- There are different approaches to the way the storage can be handled.
  - A key requirement here is to recognize that the endpoints recorded for different indications will evolve over time.
  - The simplest approach is to record the data in an excel spreadsheet. While this is straight-forward to implement, there are issues with tracking versions of the data, querying the data, preventing duplication of data entry and automating the validation of the input data.
  - An alternative is to store the data in a relational database structure which can separate the published data from the data being reviewed while at the same time providing a more effective and extendible platform for querying and extracting the data.

Figure 1. *The process used to load and use literature data indicating their partial or complete automation might be possible*


Literature database - process flow
Modeller/ librarian
Editor
Modeller/ librarian
General user
Determine the endpoints of interest
Define the specifications
Identify the papers to load
Load the required content from each paper
Quality control the content
Data ok
Yes
Report, extract data
□ Manual step □ Partially automated step □ Automated step

Consistently recording the data

- To load the data six steps can be identified. Some of these steps lend themselves to partial or full automation in an IT system.

Step 1: *Identify the key assessment variables and endpoints for an indication*

- Each indication has its own characteristic endpoints, for example for type 2 diabetes the typical endpoints are fasting plasma glucose (FPG) and HbA_{1C}. The appropriate endpoints need to be selected. It is not always feasible to load all endpoints available.
- The endpoints selected should reflect the commonly available data across a wide variety of publications (the selected endpoints should be the ones which can be used in a model and which are the most relevant ones for each type of indication).
- Standards for each unique assessment (*i.e.* units and conversions) also need to be identified.

Step 2: *Define the data specification for the indication*

- A clear specification is required to indicate what data to load and in what format.
- Guidance on the normalization of data values needs to be defined.
- These specifications can be partially automated to conduct checking of loaded data.

Step 3: *Identify the relevant publications*

- Once the key endpoints and specifications have been identified the relevant publications can be identified and prioritized to be loaded into the database.

Step 4: *Load the relevant data from the identified publications*

- The process of loading the papers can then begin. As experience is gained the specifications can be adjusted.

Step 5: *Quality control the loading of the data*

- The key step in establishing a high-quality database is diligence in checking the data to ensure its quality.
- Using the data from the specifications and other metadata, the process of quality controlling the data can be made less time consuming by automating routine checks such as the range of values.

Step 6: *Publish the quality controlled data from the database for general use*

- The final step is to publish the checked data for general use.

Maintenance

- For the database to be successful in long-term, updating the content of each database on an ongoing basis as well as periodically considering the list of endpoints to be considered, is essential.

Results

- Several indication specific databases have been developed (diabetes, respiratory) storing the data in a spreadsheet following the process outlined above manually.
- Some databases have also been commercially acquired.
- The partnership with GVK Bio for data extraction and populating the spreadsheet has been established and has proved very effective.
- However, some issues with the manual process can be identified.
  - Time taken to load the data is substantial. There is much duplication of the same data. There is a high risk of errors.
  - Effort required to quality control a paper is substantial. Estimated at 4 hours per paper.
  - Searching and extracting data from the resultant database is not intuitive.
- With the commercial databases the costs are substantial and the databases have been found not to be up to date. They do not necessarily track all the endpoints of interest.
- For these reasons a relational database based system is planned.

Conclusion

- While the manual process can be made to work and could be used in the short term, it is felt that in the longer term a more robust solution is required that will reduce the overhead of loading and quality controlling the data as well as providing a much better platform for searching, extracting, exchanging and using the data.

The Future

- Can a common data standard be established in the industry so that indication specific databases can be cooperatively shared in the pre-competitive space to prevent duplication of effort? Will the regulators and payors be interested in partnering with the industry?

This study was supported by Novartis Pharma AG, Basel, Switzerland.

References

1. Corrigan B *et al*. Proceedings of the American Conference on Pharmacometrics. Available at http://tucson2008.go-acop.org/schedule.php.
2. Mandema J W *et al*. AAPS Journal 2005;7: E513–E522.
3. Ezzet F. 2008 AAPS National Biotechnology Conference, Toronto, ON, Canada.

Poster presented at the Population Approach Group Europe (PAGE), June 23-26, 2009, St. Petersburg, Russia. Support for poster preparation provided by MSCD India, Novartis Healthcare Private Limited.