# Oncology In Vivo Data Integration for Hypothesis Generation

**Jia Wei[1*], Ming Chen[1]**

[1]AstraZeneca Innovation Center China, R&D Information China, Building 7, No.898 Halei Road, Zhangjiang Hi-Tech Park, Shanghai 201203, China

**Summary**

AstraZeneca's Oncology in vivo data integration platform brings multidimensional data from animal model efficacy, pharmacokinetic and pharmacodynamic data to animal model profiling data and public in vivo studies. Using this platform, scientists can cluster model efficacy and model profiling data together, quickly identify responder profiles and correlate molecular characteristics to pharmacological response. Through meta-analysis, scientists can compare pharmacology between single and combination treatments, between different drug scheduling and administration routes.

## 1 Introduction

Historically, oncology in vivo data across different AstraZeneca (AZ) R&D sites are managed differently. Scientists at one site use excel files to record efficacy data, while scientists at other sites use standalone applications to capture efficacy data. Genomics and genetics data for animal models are also scattered in file systems or different databases. To better share in vivo data across different sites, integrate and analyze these data in a consistent way to help define characteristics of compound sensitivity, we developed an in vivo data integration platform which provides a single access point for diverse data types and helps scientists to determine response profiles and generate hypothesis to predict clinical efficacy from preclinical studies.

## 2 InVivoDB Overview

InVivoDB integrates diverse data types, including animal model information, model genetics and genomics data, in vivo study efficacy data, in vivo pharmacokinetic (PK) and pharmacodynamic (PD) data. For standard xenograft models, their implanted cell line information is accessible from InVivoDB. For primary explant models, their implanted tissue sample information is also accessible from InVivoDB.

InVivoDB provides a set of features from in vivo study design, efficacy data collection and PK/PD sample collection, to data analysis, data visualization and study report generation. Efficacy of treatment agents is measured by tumor growth inhibition. Gene expression in animal models is measured by Affymetrix microarrays or immunohistochemistry (IHC) using tissue microarray. Gene copy number is analyzed by Agilent aCGH arrays or fluorescence in situ hybridization (FISH). Gene mutation is analyzed by amplification refractory mutation system (ARMS) and direct sequencing. The raw profiling data are summarized and uploaded into InVivoDB. PK/PD results are loaded into InVivoDB after analysis.

* To whom correspondence should be addressed. Email: jenny.wei@astrazeneca.com

InVivoDB provides a number of search functionalities which facilitate data exploitation. Faceted browsing can quickly help a new user to find interested information. By using crowd intelligence techniques, the system can recommend in vivo data in related animal models, in related disease areas or with related treatment agents based on a user's current search.

## 2.1 Animal study data workflow

InVivoDB captures in vivo study flow. It supports different types of in vivo studies including antitumor, growth curve, PK/PD and tolerability.

In study design, a user can record animal models used, randomize animal groups, record mono or combo therapies used and treatment scheme including dose, route and scheduling. The system automatically creates tasks for tumor implantation and animal dosing. It adds scheduled tasks to a user's Outlook calendar.

During in vivo data collection, data from measuring devices (caliper, balance etc) can be automatically uploaded into InVivoDB. The system generates an alert if measured tumor volume or body weight has reached predefined thresholds. Clinical observations, animal mortality and necropsy can also be recorded. For end point assays, samples of different types, for different assay purposes, and in different storage formats are collected and recorded in the system. Samples collected by time points can also be handled. Sample information and study protocol can be exported for downstream assays, including PK, PD and sample profiling.

In data analysis, InVivoDB computes anti-tumor activity parameters (inhibition%, regression) using different calculation methods (geometric mean vs arithmetic mean, tumor volume vs relative tumor volume). The system supports statistical analyses for group pair wise comparison (one tailed and two tailed student t-test). Analysis options provide flexibility for a user to explore efficacy data. A user can choose linear or log scales to visualize data, view data for specific animal groups, configure reference day for baseline tumor volume. A user can also view raw data grouped by treatment groups and animals in Spotfire.

To control data quality, InVivoDB provides mechanism for a study director to QC a study. To correct data collection mistakes, a user can modify raw data through electronic approval process built in the system. The system has audit trail to track data modification and verify data integrity. Data modification history can be viewed by a user. To monitor study data variations, a user can use Manhattan analysis which calculates standard deviation of tumor volume changes of control animal groups in a selected pool of in vivo studies.

To support research externalization, InVivoDB provides mechanism to export study protocol and data collection template to external partners. After study data are collected at external sites, they can be easily imported back to the system.

After a study is completed and validated by a study director, study reports can be exported into word, excel and PDF format.

## 2.2 Connection with other AZ systems

InVivoDB integrates with a number of AZ internal applications, including Projects database, Compound bank, Cell bank, Tissue bank, Animal inventory, Tissue microarray database, Genomic and genetics data repository, Lab notebook and People's directory.

## 2.3 In vivo data integration

By integrating different types of data, InVivoDB system provides a unified access point for in vivo studies, animal model characteristics, animal model profiling and in vivo PK/PD data.

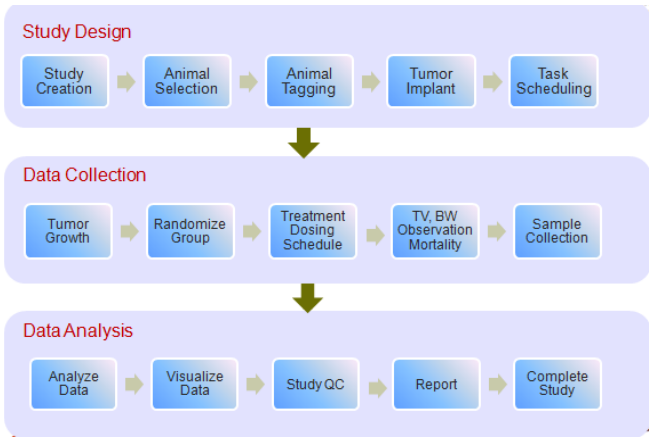


**Figure 1: Workflows for in vivo study design, data collection and data analysis.**

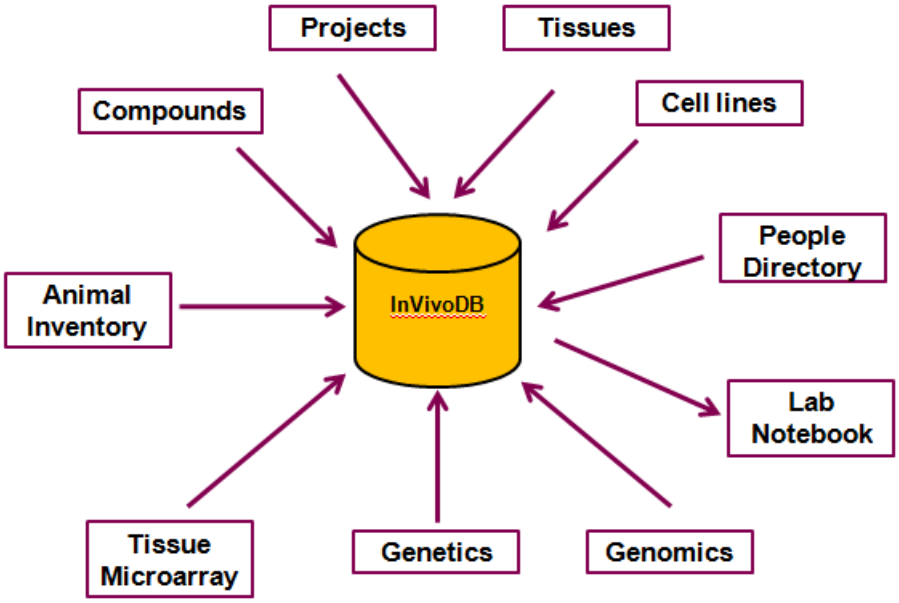


**Figure 2: Interactions of InVivoDB with other AZ systems**

### 2.3.1 In vivo data exploitation

Faceted navigation is a major data exploitation feature in InVivoDB. All in vivo data are categorized by four top level entities, namely Study, Animal Model, Gene and Treatment Agent. Under Study, data are further grouped by projects and research sites. Under Animal Model, data are further grouped by model types, disease areas and animal species. Under Gene, profiling data are further grouped by different data types. Under Treatment Agent, data are further grouped by small and large molecules.

InVivoDB also provides simple and advanced search. In simple search, a user can specify any term and the system will filter the navigation tree. In advanced search, a user can search Study, Animal Model, Gene and Treatment agents separately using different attributes for these entities.

Association is another data exploitation feature in InVivoDB. If a user selects animal models of certain type from the navigation tree, he can use Association to find all the in vivo studies using these animal models, all the treatment agents tested in these models, and all the genes

with profiling data for these models. A user can further filter the association results before visualizing them.

### 2.3.2 In vivo data integrated view

After a user selects interested animal models, genes, studies and treatment agents, InVivoDB will display an integrated view with animal model efficacy data and profiling data.

Through the integrated view, a user can compare efficacy data under different treatment agents, or with the same treatment agent but using different dose and scheduling methods. By comparing profiling data between models with and without efficacy, responder profile can be identified.

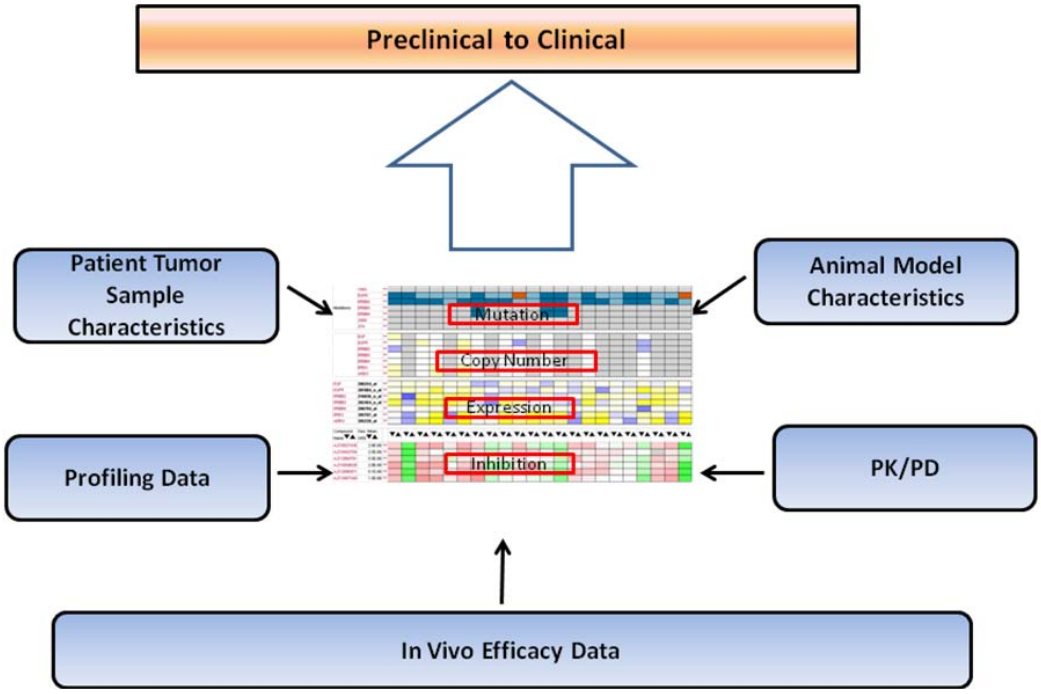


**Figure 3: In vivo data clustering for hypothesis generation.**

### 2.3.3 Recommendations by crowd intelligence

Based on a user's search query, InVivoDB recommends relevant in vivo data to view. The recommendation is performed in the following areas:

- Given a user searched animal model type, recommend data on other model types tested with the same set of treatment agents
- Given a user searched treatment agents, recommend data on other treatment agents targeting the same gene
- Given a user searched disease areas, recommend data on other disease areas using the same animal model sets and treatment agents
- Given a user searched AZ internal studies, recommend related in vivo data published

# 3 Design and Methods

InVivoDB is designed and developed as a web based application using J2EE technologies. The system is a typical three tier application: web tier, service layer and data access layer. All three layers are localized on the same machine in order to achieve high performance.

Struts2, a MVC web framework from Apache, lies in web tier to accept requests from browser clients and dispatch them to associated server-side handler, then send proper responses back to clients.

Spring lies in business layer and persistent layer. Spring is a sophisticated framework for developing enterprise-level applications.

To achieve high performance and high extensibility, full-text search engine and multidimensional data query are introduced to the system.

## 3.1 System architecture

Web tier receives requests from clients and invokes service layer interface to generate responses. Service layer contains core business logics, which invokes data access interface in data access layer. Data access layer uses Spring JDBC to access oracle database, SolrJ to access Solr indexed data, and Mondrian to query multidimensional data.

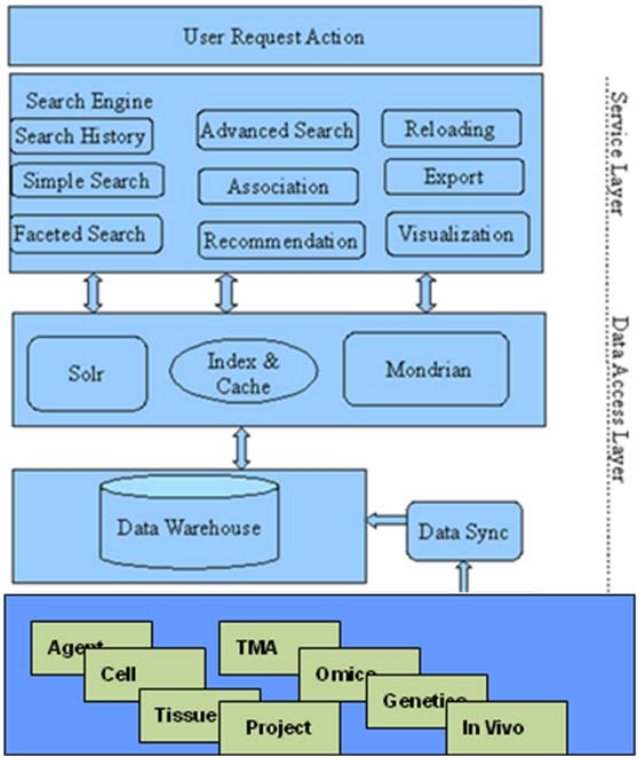


**Figure 4: Data integration and query system architecture.**

### 3.1.1 Full-text search

To achieve rapid data update, scalability, and high performance during searching, a full-text search engine-Solr is introduced to InVivoDB. Solr from Apache is a popular enterprise search platform with powerful full-text search, faceted search, database integration, rich

document (e.g., Word, PDF) handling functions. InVivoDB uses Solr as shown in Figure 5.

### 3.1.2 Data synchronization

To deliver high data accessibility, a reliable and efficient synchronization solution to ensure the consistency between original data sources and data warehouses is needed.

Solr uses a typical entity relation schema which contains four entity tables and six inter-relation tables as shown in Figure 6.

Mondrian uses a classic snowflake schema which includes two cubes, two measure tables, five dimension tables as shown in Figure 7.

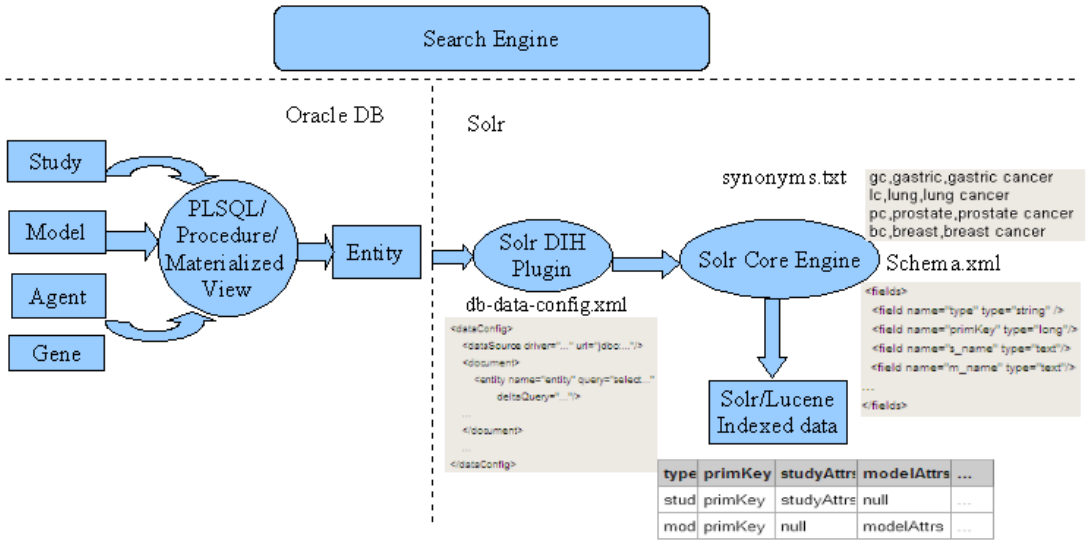


**Figure 5: Full-text search data architecture**

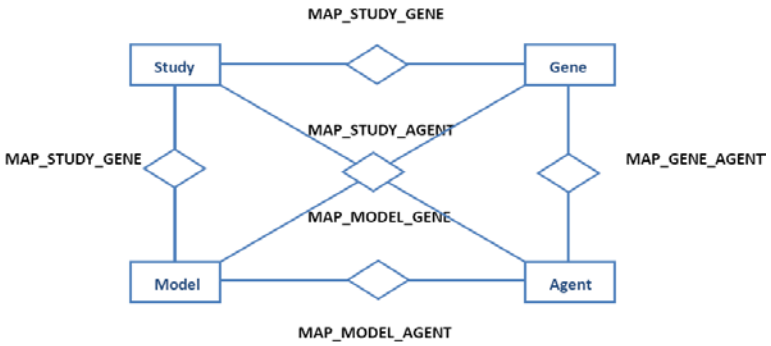


**Figure 6: Solr used entities**

Data synchronization solution uses Oracle built-in ETL functions like merge or materialized view architecture as major data replication mechanism. The major synchronization strategies used in the system are as following:

- Full table refresh
- Full table merge
- Fast refreshable materialized view
- Incremental merge

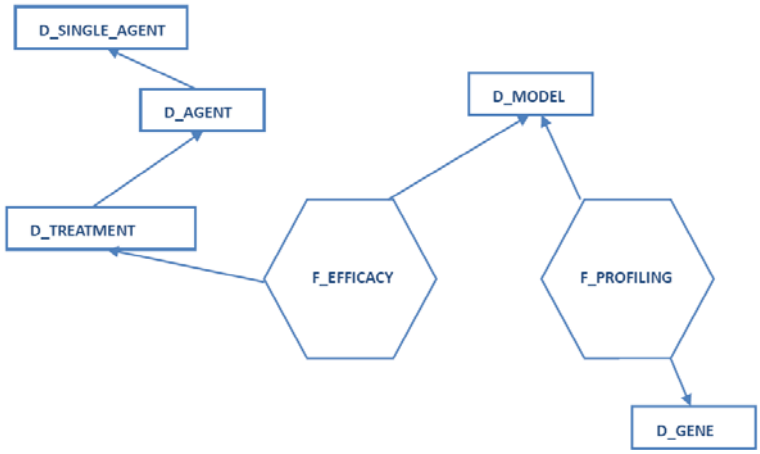


**Figure 7: Mondrian used entities**

## 3.2 Data analysis methods

### 3.2.1 Efficacy calculated as tumor growth inhibition

There are a number of methods to determine in vivo efficacy [1][2]. InVivoDB uses tumor growth inhibition to compare efficacy. %inhibition is calculated using Relative tumor volume (RTV).

% Inhibition = (Geometric Mean(RTV(Control)) – Geometric Mean(RTV(Treatment)))*100 / (Geometric Mean(RTV(Control)) – 1)

Relative tumor volume (RTV) = final tumor volume/initial tumor volume

### 3.2.2 Tumor regression analysis

When tumor inhibition is more than 100%, tumor regression is calculated as 1-Mean(RTV(treatment)) for each measurement day.

### 3.2.3 Group comparisons

InVivoDB provides statistical analysis (student t-test) for group comparisons (treated vs control, combination vs mono therapy etc).

P-value calculation is based on the logarithmic value of Relative tumor volume and uses one tailed student t-test.

### 3.2.4 Estimate group size

In study design, InVivoDB provides modules to estimate number of animals needed for each group [3]. To calculate group size, Pooled Standard Deviation (SD) is calculated from control (vehicle group) from previous 12 months worth of data as default. A user can select SD from previous 6 month worth of data, from all historical data, or from Manhattan plot, or specify SD value as input. SD is calculated within the same research site and the same animal strain. Control group should have tumor volume (TV) data for each animal from start to end of treatment. Difference of log transformed TV between start and end of the treatment is calculated and used as the input for SD estimation. SD within a study is calculated in the usual way. To pool the SD,

$SD(p) = \Sigma(n(i) -1)*SD(i) / \Sigma(n(i) -1)$, i = study 1,2,3…, n(i) = sample size of study i

InVivoDB performs sample size calculation based on normal distribution approximation in the background first using the following formula,

$$SS = 2*[\Phi(1-\alpha)+\Phi(1-\beta)]^2 * SD(p)^2 / \log(E)^2$$

α and β are the input. The default is 0.05 for α and 0.2 for β.Φ is the inverse standard normal distribution.

Power calculation uses the above SS as a starting point. With an incremental of 1 at a time, the system loops through a given SS to a calculated power until the power is over 80%. The SS where power reaches over 80% is the required sample size. In some programming tools, there is non-central T distribution whereas others do not have such distribution readily available. So an approximation is implemented.

When non-central T distribution is available,

Power(i) = 1-prob(t>TINV(0.05, df) | t~NCT(γ= log(E)/(SD(p)*sqrt(2/SS(i))), ν= SS(i) -2))

Where df = (SS(i) - 1) + n(t)*(SS(i) – 1)

When non-central T distribution is not available,

γ(i) = log(E)/(SD(p)*sqrt(2/SS(i)))

$$N(crit) = TINV(0.05, df) * (1-1/(4*df)) - \gamma(i)) / \sqrt{(1+TINV(0.05, df)^2/(2*df)}$$

Power(i) = 1-prob(x>N(crit) | x ~N(0,1))

Where df = (SS(i) - 1) + n(t)*(SS(i) – 1)

### 3.2.5 Manhattan analysis

Manhattan analysis calculates Standard Deviation of tumor volume from control group for each study for a given model within a research site and an animal strain. Standard Deviation is compared using Manhattan plot. A user can determine study data variation from Manhattan plot and perform data quality control.

### 3.2.6 Affymetrix microarray data analysis

CEL files from Affymetrix HG-U133 Plus2 arrays are first processed in Affymetrix Expression Console using MAS5 algorithm. After data QCed in Expression Console, gene based intensity data are summarized using best probeset algorithm [4].

### 3.2.7 aCGH data analysis

FE files from Agilent Human 44K aCGH arrays are processed in Nexus (www.biodiscovery.com). Gene based CNV data are summarized from interval based CNV aberrations exported from Nexus.

## Acknowledgements

We sincerely thank AstraZeneca in vivo scientists Jingchuan Zhang, Steve Wedge, Sharon Pearsall and Maureen Hattersley for helping us with application design and system review.

We thank AstraZeneca statistics scientists Ping Zhan and Robert Shaw for statistical analysis support. We also gratefully acknowledge the support of Augmentum for system development.

## References

[1] J. Wu. Statistical Inference for Tumor Growth Inhibition T/C Ratio. *Journal of Biopharmaceutical Statistics*, 20(5):954-964, 2010.

[2] K. Fluiter, A. L..M..A. ten Asbroek, M. B. de Wissel, M. E. Jakobs, M. Wissenbach, H. Olsson, O. Olsen, H. Oerum and F. Bass. In Vivo Tumor Growth Inhibition and Biodistribution Studies of Locked Nucleic Acid (LNA) Antisense Oligonucleotides. *Nucleic Acids Res.*, 31(3):953-962, 2003.

[3] D. Hussey, J. N. DaSilva, E. Greenwald, K. Cheung, S. Kapur, A. A. Wilson and S. Houle. Statistical Power Analysis of In Vivo Studies in rat Brain Using PET Radiotracers. *In* R. E. Carson, M. E. Daube-Witherspoon and P. Herscovitch (editors). Quantitative Functional Brain Imaging with Positron Emission Tomography, Academic Press, 273-277. 1998.

[4] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass and H. W. H. Göhlmann. I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, 23(21):2897-2902, 2007.