

# DATA MINING

## In Drug Development and Translational Medicine

July 2009

Hermann A.M. Mucke, PhD

**Transforming Data Into Knowledge**

The biopharmaceutical industry is grappling not only with sheer data volume but with the ability of researchers to extract information through identification and contextual analysis of those data that are relevant to a particular set of investigations. This report examines:

- **Techniques, technology, and software used in life science data mining**
- **Data mining for early preclinical safety assessments**
- **Data mining in clinical trials**
- **Data mining in pharmacovigilance**
- **Business models and solutions in drug development bioinformatics**

## Overview

The mountain of data generated and stored is growing ever-higher. The information content of life science data is multidimensional and not readily accessible by merely looking at the output. Unless such data can be put into proper context and interpreted—*i.e.*, mined—their value is only in their potential. **Data Mining in Drug Development and Translational Medicine** examines data mining challenges and approaches in pharmaceutical R&D.

The pharmaceutical industry has made decisive moves to improve the predictiveness of early-stage drug safety testing. These efforts generate large amounts of data, in which the clue to safety-related, potential “red flags” can be buried. In this context we examine options for mining types of text data, “pathway mining” for pathway-related effects of a compound, and the multidimensional output of high-content screening methods. Also examined are approaches to mining data generated in preclinical trials for identification of toxicity signatures.

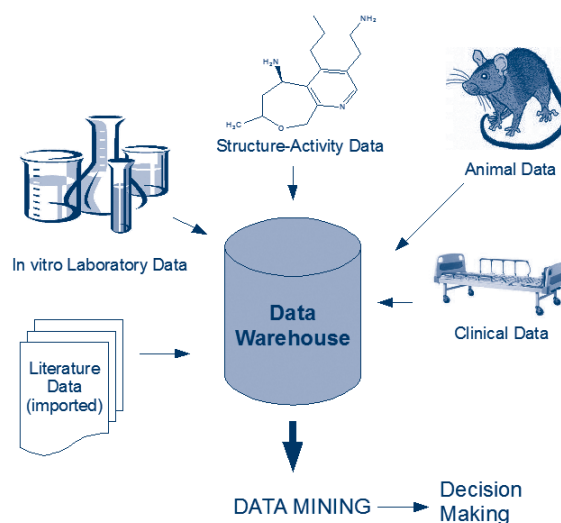
Much more clinical trial data are captured than are actually analyzed to build the regulatory data file. Clinical databases can thus be mined for information that the respective study was not explicitly designed to provide. **Data Mining in Drug Development and Translational Medicine** describes how data mining from investigational human trials can reveal hidden information that has the potential to massively improve the understanding of drug mechanisms, the efficacy and side effect behavior of drug candidates in various patient subpopulations, and even the integrity of clinical investigators. We look at text mining of literature and patent databases, which offers the possibility for knowledge discovery concerning activity in a particular field of therapeutic development from many different angles.

Pharmacovigilance is a field where large volumes of interconnected data have to be analyzed in many dimensions. We describe various databases used in support of post-market drug safety evaluation, including those maintained by the FDA, WHO,

and EMEA. Data mining algorithms applied to pharmacovigilance databases and efforts to bring separate databases into full compatibility with one another are described. Case studies illustrating the use of data mining and analysis to investigate relationships between marketed drugs and adverse events are presented.

**Data Mining in Drug Development and Translational Medicine** concludes by profiling the most significant vendors that either offer dedicated solutions for data mining in drug development and pharmacovigilance, or provide more general commercial data mining solutions that have been successfully adapted and applied to these endeavors.

## The Data Warehouse as a Hub in Translational Drug Research and Development



Source: H.M. Pharma Consultancy

**About the Author:** *Hermann A.M. Mucke, PhD*, spent 17 years in academia and industry before he founded H.M. Pharma Consultancy ([www.hmpharmacon.com](http://www.hmpharmacon.com)) in 2000 to become an independent pharmaceutical consultant, analyst, and science author. His last industry position was Vice President R&D in a European pharmaceutical company, which he helped to take public on the Frankfurt Stock Exchange in 1999. Since then, Dr. Mucke, who holds a PhD in biochemistry from the University of Vienna (Austria), became a consultant and advisory board member for several European and American pharmaceutical companies and a regular reviewer of drugs and patents for Thomson Current Drugs and Ashley Publications. Dr. Mucke is based in Vienna.

## Tables and Figures

### FIGURES

The Knowledge Extraction Pyramid  
Development of Searches in the PubMed Internet Database, 1997–2007  
The Data Warehouse as a Hub in Translational Drug Research and Development  
Visualization of a Mining Query of the PubMed Literature Database  
Representation of a Typical Clinical Data Mining Workflow  
The Data Integration Challenge in Clinical Data Mining  
Workflow Schematic for the Data Mining System Described by Cao *et al.* [*Immunome Research*. 2008;4:7.]

Vaccine Trials Activity Relative to Cancer Prevalence and Survival Reports Received and Entered Into the AERS Database by Type of Report, 1999–2008  
Adverse Event Reports for Oseltamivir vs. Unexpectedness, 1997–Q1/2008  
Specific Symptoms In Influenza Patients Treated with Oseltamivir for Whom “Abnormal Behavior” Had Been Reported  
Screenshot of an Analysis with Cambridge Cell Networks’ ToxWiz Software  
TIBCO Spotfire DecisionSite Software for Preclinical Research  
A Window from the TIBCO Spotfire Clinical Trials Analysis Software

To order a report, e-mail [rlaraia@healthtech.com](mailto:rlaraia@healthtech.com), call Rose LaRaia at 781-972-5444, or order on-line

# Table of Contents

## CHAPTER 1: THE NEED FOR DATA MINING IN DRUG DEVELOPMENT: NATURE AND OBJECTIVES

- 1.1. The Exponential Growth of Humankind's Data Volume
- 1.2. Making Sense of Data: Ascent to the "Grand Picture"

Learning About the Unexpected: Exploratory Data Analyses for Hypothesis Generation

Seeking Specific Signatures: Data Mining for Hypothesis Testing

- 1.3. Who Mines Data Today...And For What?

Strategic Marketing

Financial Services and Tax Offices

Military and Security Assessments

Other Users of Data Mining Solutions

- 1.4. The Challenge of Life Science's Own Data Avalanche

Literature and Patent Texts

Cheminformatics

Sequence and Biomarker Information

Modeling of Market Dynamics and Competitor Behavior

## CHAPTER 2: TECHNIQUES, TECHNOLOGY, AND SOFTWARE

- 2.1. Capturing Data and Knowing Their Bias

Experimental and External Data

- 2.2. Building Data Warehouses from Disparate Sources

- 2.3. Text Mining: Semantics and Artificial Intelligence

- 2.4. Structure Searches in Digital Chemical Libraries

- 2.5. Image Mining: The Greatest Challenge

- 2.6. Machine Learning with Pharmaceutical and Biological Data

- 2.7. Visualization of Results: The Challenge of Meaningful Reporting

- 2.8. Standardization and Regulatory Compliance: CDISC's SDTM and SEND

## CHAPTER 3: DATA MINING FOR EARLY PRECLINICAL SAFETY ASSESSMENTS

- 3.1. A Close Look at Text Data: Literature, Patents, and Databases

- 3.2. "Pathway Mining" for Model Building and Matching

- 3.3. High-Content Screening as a Data Feed

- 3.4. Seeking Signatures of Toxicity in Animal Data

Behavioral Data: From Automated Counts to Video Mining

Biomarker Response Assessments in Animal Studies

Seeking Out and Interpreting Digital Pathology Data

## CHAPTER 4: DATA MINING IN CLINICAL TRIALS

- 4.1. The Clinical Trial Database: Much More Than Meets the Eye

The "E-Trial": The Key to Patient Record Mining in Near-Real Time

Retrospective Mining of Completed Trials: The "Paper Legacy"

*Case Study: Statins and Amyotrophic Lateral Sclerosis*

- 4.2. Mining for Safety Signals in Clinical Trials

Premarket Safety Data Mining by Regulatory Agencies

Hepatotoxicity

QT Interval Prolongation

- 4.3. Clinical Trial Data Mining for Drug Response Signatures

Genotype versus Phenotype: Identifying Potential Responders

*Image Registration: Mining Imaging Data for Response Signatures*

- 4.4. Detection of Data Bias and Fraud

- 4.5. Correcting for Non-Compliance in Outpatient Trials

- 4.6. Mining the Clinical Literature for Optimizing Scientific Approaches and Business Development

## CHAPTER 5: DATA MINING IN PHARMACOVIGILANCE

- 5.1. The Challenges of Assessing Post-Marketing Drug Performance

- 5.2. Databases Supporting the Push for Post-Market Safety Evaluation

AERS and VAERS: The FDA Adverse Event Reporting System

VigiBase: The WHO Drug Safety Database

The EudraVigilance Post-Authorization Module

Prescription-Event Monitoring Databases

Corporate Pharmacovigilance Databases

## 5.3. Mining Adverse Event Databases

Basic Types of Mining Algorithms

The Influence of Coding Terms and Direct Patient Reporting

Case Studies and Promising Objectives

*Oseltamivir and Hallucinations*

*Antipsychotics and Diabetic*

*Events: An Effect of Chemical Structure?*

*Statins and Psychiatry: A*

*Confusing Story with a Long History*

*Biphosphonate Drugs and*

*Osteonecrosis of the Jaw*

- 5.4. Developments Shaping the Data Mining Environment in Pharmacovigilance

The FDA's Sentinel Initiative and the Reagan-Udall Foundation

PROTECT – Method Development for Pharmacovigilance in Europe

Electronic Health Records: A Future Key Factor for Data Collection

## CHAPTER 6: BUSINESS MODELS AND SOLUTIONS IN DRUG DEVELOPMENT BIOINFORMATICS

- 6.1. Phase Forward

- 6.2. ProSanos

- 6.3. AltraBio

- 6.4. ID Business Solutions (IDBS)

- 6.5. Strand Life Sciences

- 6.6. SPSS

- 6.7. PointCross

- 6.8. Aperio Technologies

- 6.9. Molecular Devices

- 6.10. Cambridge Cell Networks (CCNet)

- 6.11. InforSense

- 6.12. SAS Institute

- 6.13. Temis

- 6.14. Search Technology

- 6.15. TIBCO Software

- 6.16. Salford Systems

## REFERENCES

## COMPANY INDEX WITH WEB ADDRESSES

## Related Conference

### CO-LOCATED SEQUENCING CONFERENCES

CHI's Second Annual

#### Next-Generation Sequencing Data Analysis

September 21-23, 2009

Rhode Island Convention Center | Providence, RI

and

CHI's Third Annual

#### Exploring Next-Generation Sequencing

September 21-23, 2009

For details and to register, visit: [www.healthtech.com](http://www.healthtech.com)

## Related Report

### Bioinformatics and Computational Biology: Bottlenecks and Options

The interdisciplinary fields of Bioinformatics and Computational Biology are locked in a high stakes race with analytical instrument developers and innovators. The pace and scope of change in many fields of biomedical research rivals what we once associated only with semiconductor devices. This report explores the interlocking challenges facing instrumentation advances, computational demands and our evolving systems biology knowledge. Key challenges presented in this report include:

- Instrumentation capable of generating terabytes of raw data daily
- Storage requirements for human gene sequences
- Need for cross-platform data analysis standards
- Appropriateness of analysis and modeling applications
- Database data quality and annotation protocols

Get **FREE** sample pages, view full details and order this report at [InsightPharmaReports.com](http://InsightPharmaReports.com)

## About Insight Pharma Reports

CHI Insight Pharma Reports are written by experts who collaborate with CHI to provide a series of reports that evaluate the salient trends in pharmaceutical technology, business, and therapy markets.

Insight Pharma Reports are used by senior decision makers at life science companies to keep abreast of the latest advances in pharmaceutical R&D, their potential applications and business impacts. Our clients include the top 50 pharmaceutical companies, top 100 biotechnology companies, and top 100 vendors of life science products and services. Typical purchasers are managers, directors, and VPs in business development, discovery research, clinical development, strategic planning, portfolio management, new product planning, and marketing.

#### Insight Pharma Reports offer:

- Current information and analysis of R&D technologies, therapeutic markets, and critical business issues.
- Analysis of the probability of success for various applications of each technology.
- Expert insight based on interviews with key personnel in companies at the forefront of technological advances who share their views on their technology's current status, applications, future direction, and market environment.

### Barnett Educational Services Leading Industry Publication

**New Drug Development: A Regulatory Overview:** A critical resource that addresses the most cutting-edge developments redefining how new drugs are developed and regulated today. For more information, visit [www.barnettinternational.com](http://www.barnettinternational.com) or call 800-856-2556

### Yes! I would like to receive a FREE subscription to:

- eCliniqua *Innovative management in clinical trials*
- Predictive Biomedicine *Informatics tools and strategies driving decisions*
- Weekly Update *The latest industry news, commentary, and highlights from Bio•IT World*

## Data Mining in Drug Development and Translational Medicine

Data Mining in Drug Devel. & Translational Medicine — July 2009 (114 pages)

Print  
 \$2,995.00

Single-Site/Operational Unit License\*

\$3,750.00

Bioinformatics and Computational Biology — May 2009 (134 pages)

\$2,995.00

\$3,750.00

#### Purchase both reports and receive a 10% discount

\*Single-site licenses are multi-user, searchable, cut-and-paste ready PDFs

Call for global license pricing; contact David Cunningham at 781-972-5472 or [cunningham@healthtech.com](mailto:cunningham@healthtech.com)

Total: \$ \_\_\_\_\_

Choose a payment option:

1.  Enclosed is a check order payable to Cambridge Healthtech Publishing, in U.S. currency. (In Massachusetts, add 5% sales tax.)

2.  Purchase order number \_\_\_\_\_

3. Credit card:  Amex  Visa  MC \_\_\_\_\_ Exp. Date: \_\_\_\_\_ Sec. Code: \_\_\_\_\_

Cardholder: \_\_\_\_\_ Signature \_\_\_\_\_

Mr.  Ms.  Mrs.  Dr. First Name: \_\_\_\_\_ Last Name \_\_\_\_\_

Job Title: \_\_\_\_\_ Div./Dept. \_\_\_\_\_ Company \_\_\_\_\_

Address (please include Mail Stop, Room or Bldg. #): \_\_\_\_\_

City/State/Postal Code: \_\_\_\_\_ Country: \_\_\_\_\_

Telephone: \_\_\_\_\_ Fax: \_\_\_\_\_ E-Mail: \_\_\_\_\_

Please refer to the key code below

#### TO ORDER:

Web: [InsightPharmaReports.com](http://InsightPharmaReports.com)

Phone: 781-972-5444

Fax: 781-972-5425

E-mail: [rlaraia@healthtech.com](mailto:rlaraia@healthtech.com)

Mail: Rose LaRaia  
250 First Avenue, Suite 300  
Needham, MA 02494