

**Figure 1.2. Development of Searches in the PubMed Internet Database, 1997–2007\***



\* 1997 = baseline

*Source: National Center for Biotechnology Information*

Patent offices are drowning in applications. According to statistics and estimates published by the World Intellectual Property Organization, the average number of pharmaceutical and biotechnology patent applications filed in the world's patent offices was about 115,000 per year for the period 2001–2005.<sup>3</sup> (This figure, which is the most recent one available, does not include filings from among the 60,000 applications in organic chemistry that subsequently find pharmaceutical use.)

In another more specialized example, the number of published international patent applications concerning drugs, diagnostics, and tissue-based technologies for eye medicine (excluding equipment unless it is intended for drug delivery) was 119 in the year 1998. In 2008, the figure had grown to 420 applications and the growth had been consistent during the ten-year period.<sup>4</sup> During the same ten-year period, the average length of published PCT patents in this field grew by almost 40%.



A team working at Toray Industries (Kanagawa, Japan) has developed a gene expression chip that, according to the authors, can work directly with 10–100 pg of messenger RNA.<sup>8</sup> This is roughly two orders more sensitive than reference systems. Such a microarray could allow an analysis of toxicity-related transcriptome elements without prior amplifications, much closer to real-time mode than is possible today. Such a system could feed its data directly to a data mining system, allowing the results to be correlated with more directly observable parameters with minimal delays. This would not resolve the problem of tissue sampling, but blood samples that are routinely obtained from “satellite animals” (which are routinely added to the standard animal cohort to investigate toxicokinetics) could serve as an RNA source, however, one that is not tied to a specific organ.

### Seeking Out and Interpreting Digital Pathology Data

An old joke says that pathologists know everything about diseases a doctor could wish to know—it’s just that their knowledge comes too late. This is, of course, no longer fully true in human medicine. But in preclinical drug investigations, histopathology essentially remains limited to post-life assessment of organ preparations and microscopy slices. As has been mentioned before, histopathology is not only a bottleneck in preclinical trials but it also remains a highly subjective assessment by a specialized expert who excels at diagnosis based on a trained capability of specific pattern recognition in microscopy images.

There have been many attempts to automate cytopathology and histopathology and (at least ultimately) to remove the pathologist from the assessment by using image processing and analysis software. AQUA (Automated Quantitative Analysis technology; from HistoRx, a Yale University spinoff originally known as Histometrix) is an example of such a software package, which attempts to define morphologic features in a sample based on the expression of a biomarker (*e.g.*, keratin for delineating cells in a stained tissue slice). The software creates a binary mask that is used to automatically delineate regions of interest in the sample (*e.g.*, a tumor) using subtraction algorithms. One published study compared estrogen receptor expression ratings of primary breast cancer by pathologists and by the AQUA automated analysis system and found excellent concordance between scores from each source (mean linear regression R value 0.89).<sup>8a</sup> Such digital pathology can achieve superiority, especially with high-expressing tumors. Other software packages (*e.g.*, from market leader Aperio Technologies in Vasta, California [www.aperio.com] and those from other vendors) perform

This chapter 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 the clinical investigators. The discussion mainly deals with Phase III programs because these provide a sufficiently large data pool for mining. However, data from larger Phase II trials can also be meaningfully mined, especially those with rich data capture designs. There are two principal modalities: near-real time and retrospective.

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

Clinical studies are strictly formalized investigations in hospitalized or out-patients that are designed along strict rules and with very specific goals in mind. The rules are those set by regulatory authorities, and the goals are defined by the nature of the claims for which the sponsors intend to secure regulatory approval. In the case of a Phase III program, there is really only one question to answer: Namely, if the drug is safe and effective for the intended purpose. It is definitely not good statistical practice to analyze a trial for something it has not been intended to demonstrate. Could clinical databases nevertheless be mined to obtain information that the respective study was not explicitly designed to provide?

The answer is a resounding yes. The information that can be uncovered and used in hypothesis generation (*i.e.*, to elucidate which further systematic questions might be the most important to ask) includes the following:

- Identification and characterization of patient subgroups (responders, non-responders, patients who are particularly prone to side effects)
- Uncovering of previously unknown relationships between safety and efficacy aspects of the drug
- Early identification of rare adverse drug effects (of the type that could precipitate drug withdrawal after approval and launch)
- Better (multidimensional) characterization of the cost-efficacy relationship
- Better comparative profiling against existing therapeutic alternatives

confirmation, rather than signal detection. From a technical standpoint, we believe that the methodologies around signal confirmation or hypothesis strengthening may be much more feasible to implement in the near term. After these approaches have been validated, additional attention should be turned to signal detection.”

### **PROTECT – Method Development for Pharmacovigilance in Europe**

The European Medicines Agency has established an approximate equivalent to the Reagan-Udall Foundation: a multinational consortium of 29 public and private partners that will manage PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium), a collaborative European project aiming to develop innovative methods in pharmacoepidemiology and pharmacovigilance. PROTECT will run over five years, with a total funding of EUR 20 million. Half of the funding will be in-kind contributions from the participating companies who are members of the EFPIA organization (the approximate European equivalent of BIO in the United States).

*PROTECT, which has a target start date for the project of September 2009, is more geared toward methodology development than the FDA's Sentinel Initiative.*

PROTECT, which has a target start date for the project of September 2009, is more geared toward methodology development than the FDA's Sentinel Initiative. It will look at limitations of current methods used in pharmacovigilance and pharmacoepidemiology in order to strengthen the monitoring of the benefit/risk balance of medicines marketed in Europe. A set of innovative tools and methods will be developed and validated in specialized Work Packages, including (i) modern tools of communication to directly collect data from consumers of medicines; (ii) improved tools for early and proactive detection of signals; (iii) a framework for the design, conduct, and analysis of pharmacoepidemiological studies; and (iv) methods for continuous benefit/risk monitoring.

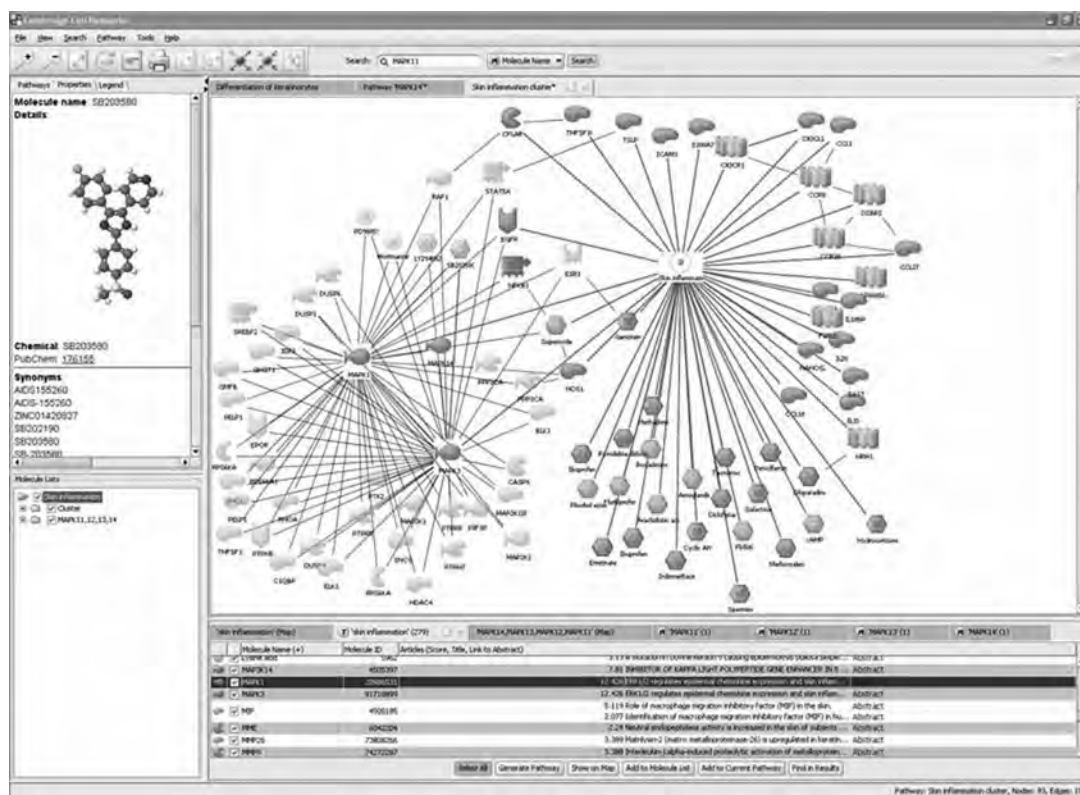
### **Electronic Health Records: A Future Key Factor for Data Collection**

The American Recovery and Reinvestment Act of 2009 (ARRA) has set itself the goal to stimulate the use of healthcare information technology through an ambitious \$19-billion package. It has a particular focus on electronic health records (EHRs), which so far have seen only very limited implementation even on their most basic level: While 17% of US hospitals and only an estimated 10% of doctors keep EHRs today, The Office of the National Coordinator for Health Information Technology (ONC), a leadership structure to guide federal healthcare

CCNet is a toxicology-centered computational systems biology company founded in 2002 as a spinoff from the University of Cambridge. It develops tools for scientists who need to better understand on- and off-target mechanisms of action of pharmacologically active compounds and the links between genes, targets, chemicals, and pathologies.

The central software is ToxWiz (currently in version 2.2), which is essentially a mining tool for toxicology-related biological pathways. ToxWiz has a database of about 35,000 bioactive compounds and hundreds of thousands of datapoints (molecules, pathologies, and their interactions) integrated. ToxWiz is designed to predict the possible cellular mechanisms of any indicated toxicity. It features a very fast fingerprint search algorithm coupled to a Tanimoto score, as well as multiple searches to better serve the metabolomics community.

**Figure 6.1. Screenshot of an Analysis with Cambridge Cell Networks' ToxWiz Software**



Source: Image © [www.camcellnet.com](http://www.camcellnet.com); Apic G, Ignjatovic T, Boyer S, Russell RB. Illuminating drug discovery with biological pathways. *FEBS Lett.* 2005;579:1872–7.