CHAPTER 2 LITERATURE REVIEWS

2.1 Clinical Pharmacology

Pharmacology is the science of drugs and their effects on biological systems. Drug, the recreational substance in a medicine, can be defined as a chemical that can cause a change in a biological system (Shile and Stöppler, 2008). Modern pharmacology owes part of its developments to William Withering, who published a book entitled *An Account of the Foxglove and Some of Its Medical Uses* in 1785. He noticed the power of using digitalis as medicine for congestive heart failure which constructed a way of rationally approaching a therapeutic problem. Several botanical drugs were therefore extracted, purified, and concentrated into fractionated components to be used as drugs (Burger, 1995; Hollinger, 2002). Subsequently, clinical pharmacology which often differed from basic pharmacology has been developed. Clinical pharmacology can be specifically characterized as the study of drugs and the exertion of their effects in human body (Atkinson, et al., 2006). This knowledge involves the complex interaction between the patient and the drug. Over recent years clinical pharmacology has undergone great expansion resulting that allows the understanding of molecular action and the capacity to exploit the enormous therapeutics potential.

2.1.1 Principles of Clinical Pharmacology

Drugs are chemical compounds that naturally apply their structure to modify or influence the way the body works in biological activities. Drugs can also mimic, facilitate, or antagonize a normally occurring phenomenon. An understanding of pharmacological concepts is important for drug evaluation and individualized therapy management. The discipline that drives the pharmaceutical industry is described in further detail below.

2.1.1.1 Pharmacokinetics

If drug has a therapeutic effect, it first has to be administered and absorbed in some way to reach its site of action. Later administered and absorbed, distribution to different parts of the body follows. The passage necessarily gets through the liver that almost metabolizes some drugs into totally inactivated, and then generally excretes via the kidney. The properties of the movement of drugs into, within, and out of the body are known as *pharmacokinetic (PK) characteristics*. The components of PK are ADME; administration or absorption, distribution, metabolism, and elimination or excretion (Thorp, 2008).

Administration or Absorption of drugs

In order to get the response, drugs have to be administered in some route. There are two major routes of drug administration: enteral and parenteral. Enteral route means to be permeable within the gastrointestinal tract that starts from oral and ends at rectal. By

contrast, parenteral route includes all other means of enteral route, for example, intravenous injection, inhalation, and skin.

Drugs distribution

When drugs are absorbed from site of administration, then blood circulation transports the free drugs to the tissues within plasma. The body can be considered to be assembled of aqueous and lipid compartments. The distribution of drugs into different compartments depends on aqueous solubility, blood flow, plasma protein binding, lipid solubility, acid dissociation, tissue sequestration, metabolism and excretion, and volume of distribution.

Drugs metabolism

Most drugs are chemically modified or metabolized in the body which determine the duration of drugs in action, elimination, and toxicity. Drug metabolism may provide an active compound to inactive, or activate an inactive precursor, or produce a toxic by-product. However most tissues in the body have the enzymes for metabolizing a variety of substances, the majority of drug metabolism takes place in the liver. There are two general types of metabolic reactions, Phase I and Phase II, based on chemical nature of drug. Phase I is biotransformation that cytochrome P450 monooxygenases (CYP) catalyze drugs and endogenous compounds by oxidation in the liver to a more water-soluble metabolites. Phase II is the reactions that conjugate drugs or metabolites of Phase I into more hydrophilic, less toxic substances by adding polar product to endogenous species. The most important conjugation reaction is glucuronidation, a major portion of drug metabolites in urine excretion. Other conjugations occur with sulfate, acetyl, methyl, and glycine groups. Additional effects on drug metabolism are enzyme induction and inhibition, species, sex, age, and individual variation (Atkinson, et al., 2006; Thorp, 2008). Some drugs experience both Phase I and Phase II reactions while others are possibly metabolized by either type of reaction only as shown in Figure 2.1.

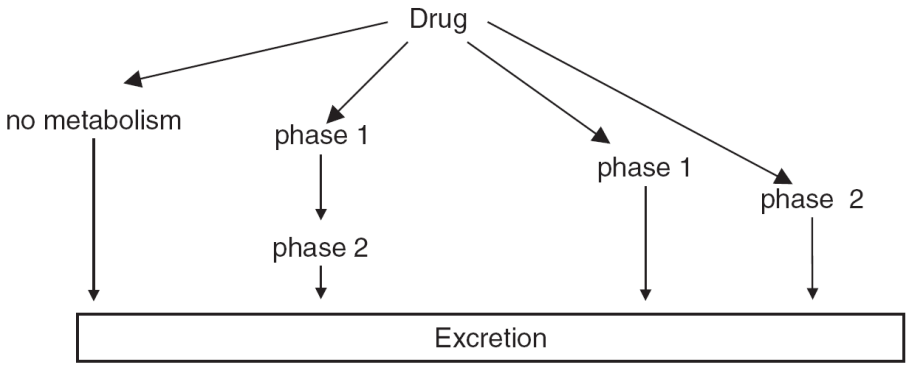


Figure 2.1 Possible routes for metabolism of drugs (Thorp, 2008)

Elimination/Excretion of drugs

Even though drugs intend to be therapeutic use, they also have the potential as unknown substances that should be rapidly eliminated from the body. The main route of excretion of drugs and drug metabolites is via the kidneys, particularly nephron. Additionally, biliary, pulmonary, sweat, salivary, and mammary glands are also elimination routes in order of decreasing importance (Hollinger, 2002; Thorp, 2008).

2.1.1.2 Pharmacodynamics (PD)

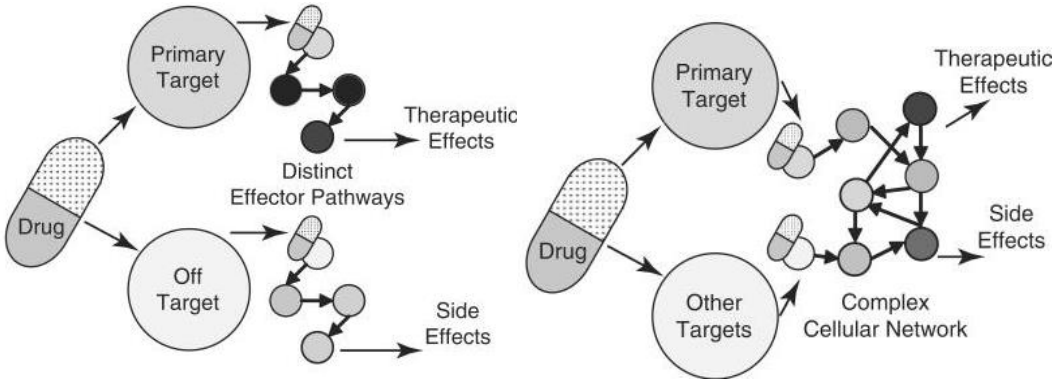
Using of medications, drugs are usually distributed through the target site that profit on therapeutic effect to cause the treatment. However, the discipline that quantifies the relationship between drug concentration at the site of drug action and the drug's pharmacological effect is essential. Body functions are mediated by the way of control systems that required chemotransmitters or local hormones, receptors, enzymes, carrier molecules and other specialized macromolecules such as DNA. Drugs act by binding to some specialized component of the cell to alter their function. Understanding of drugs that can have and produce the effects is known as *pharmacodynamics (PD)*. Hence, PK, what the body does to drugs, and PD, what drugs do to the body, comprise two major subdivisions of clinical pharmacology (Bennett and Brown, 2003; Atkinson, et al., 2006; Thorp, 2008)

2.1.1.3 Pharmacogenetics (PGt) and Pharmacogenomics (PGx)

During the clinical use of a drug, a prescribing physician has no intends to the individual response because of the broad assumption that all patients are homogeneous group, little or no interindividual variation. The finding that severe anemia caused by an inherited deficiency of glucose-6-phosphate dehydrogenase (G6PD) in 1950s is the first example that genetics can determine apparently unpredictable drug toxicity (Meyer, 2000; Giacomini, et al., 2007). It sets out that genetic factors (genotype) of an individual significantly govern the safety and efficacy outcomes in drug responses (phenotype). The scientific field that studies the role of genetics and the possible relationship to medication therapy is referred to as *pharmacogenetics (PGt)* or *pharmacogenomics (PGx)*. Even they are used interchangeably, they are still unlike (Lee, et al., 2010). Based on FDA definitions, PGx is the general study of all of the many different genes that can determine new drug behavior while PGt is defined as the study of variations in DNA sequence, individual gene variants, as related to drug metabolism and response (Roden, et al., 2006; Al-Ghoul and Valdes, 2008). PGt influences PD in pharmacological targets of drugs and drug metabolizing enzymes by presenting genetic polymorphisms that affect drug response. PGt influences PK in most drug metabolizing enzymes that are expressed in genetically variant forms. As a consequence, PGt grounds the functional basis of PK and PD. Methodologies of PGx and PGt can personally lead to their ultimate goal that optimize drug treatments in both terms of efficacy and safety (Lesko and Woodcock, 2004; Al-Ghoul and Valdes, 2008).

2.1.1.4 Systems Pharmacology

During the latter half of the 20th century, biology is strongly influenced by approaches that focus on the generation of information. The advent of high-throughput (HT) experimental technologies is forcing biologist to view cells as systems, rather than fixing their attention on individual cellular components. Analysis biology in systems is the discipline that academics and also pharmaceutical industries are rapidly accepting. Systems pharmacology research changes the way believes in drug actions. As shown in Figure 2.2, the classic view of drug actions, therapeutic effects and side effects are modulated though different pathways, has been changed to systems pharmacology view. It considers the nature of the links that connect components and expresses the functional states of the networks which leading to both therapeutic and adverse reactions (Berger and Iyengar, 2009; Csermely, et al., 2013).



(A) (B)

Figure 2.2 Changing view of drug actions (Berger and Iyengar, 2009)

(A) The classic view of drug action. (B) The systems pharmacology view of drug action

Network analysis in clinical pharmacology not only aims to describe and to understand the operation of complex biological in systems, but also helps to identify new drug targets, leads to optimize drug efficacy or other potentially interesting properties of the drugs, and minimizes unpleasant effect of the drugs (Butcher, et al., 2004; Ekins, et al., 2005; Tatonetti, et al., 2009; Pujol, et al., 2010). For all of advantages, systems pharmacology, if absolutely successful, should receive the privilege to play a central part in the development of novel polypharmacology strategies.

2.1.2 Drugs Development

The rational of drug discovery has relied upon increasing numbers of known natural chemical mediators. In the past decade, PGt and systems pharmacology have led to a new pathway from previous mysterious mechanism. Drug development, therefore, is a complex process that is divided into preclinical research and development (R&D) and

clinical development phase (Roses, 2004). For having a potential target, usually protein, drug candidate is identified and optimized, then put through *in vitro* screens and animal testing in preclinical phase as illustrated in Figure 2.3.

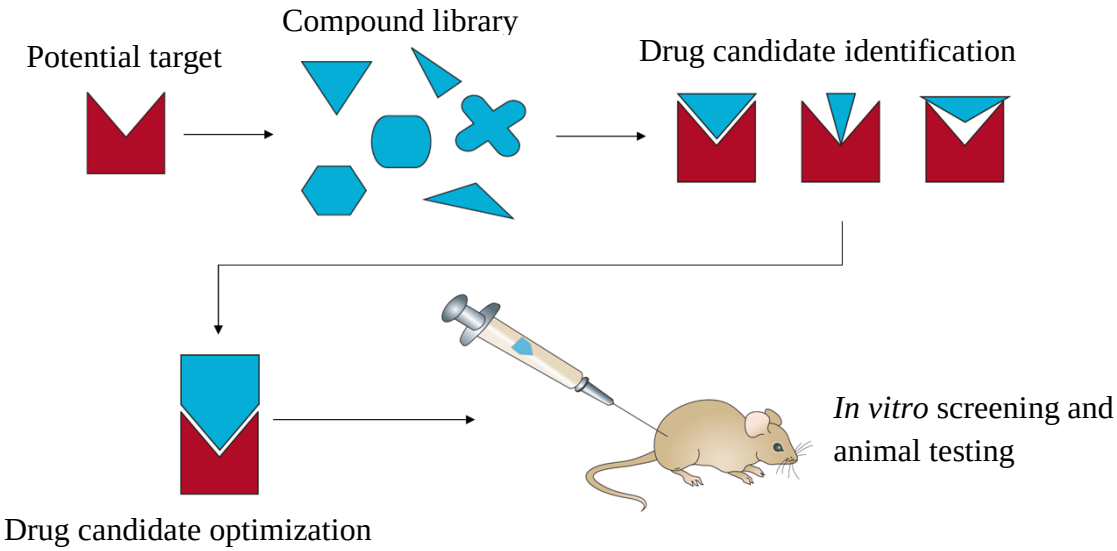


Figure 2.3 Preclinical phase of pharmaceutical pipeline, modified from Roses (2004)

After preclinical, Phase I clinical development begins with a limited number of studies in healthy volunteers or patients to test for safety and dosage in human. Phase II studies in a selected group of patients. They are conducted to obtain therapeutic efficacy and drug toxic responses. Phase III trials confirm therapeutic efficacy and document safety of drug in a larger patient population before accreditation. Clinical phase of drug development is exhibited in Figure 2.4.

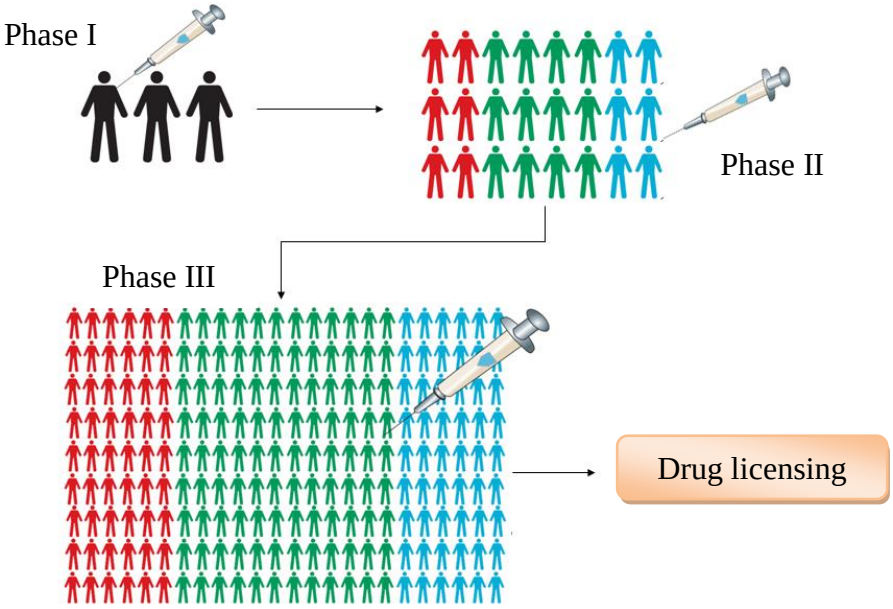


Figure 2.4 Clinical phase of pharmaceutical pipeline, modified from Roses (2004)

These pharmaceutical pipelines usually require several years and huge cost of development, approximately 1 billion US dollars, to advance a new successful drug (Roses, 2004; Atkinson, et al., 2006). Nevertheless, tiny investments are successful (Csermely, et al., 2013). One new molecular entity (NME) launched to the market requires approximately 24 development candidates to enter into the development pipeline. Attrition of Phase II studies is the key challenge, where only 25% of drug-candidates survive. The success rate of NMEs for 14 large pharmaceutical companies is demonstrated in Figure 2.5. Because of the failures during development or especially after launching, the approach to minimize these downfalls is still required.

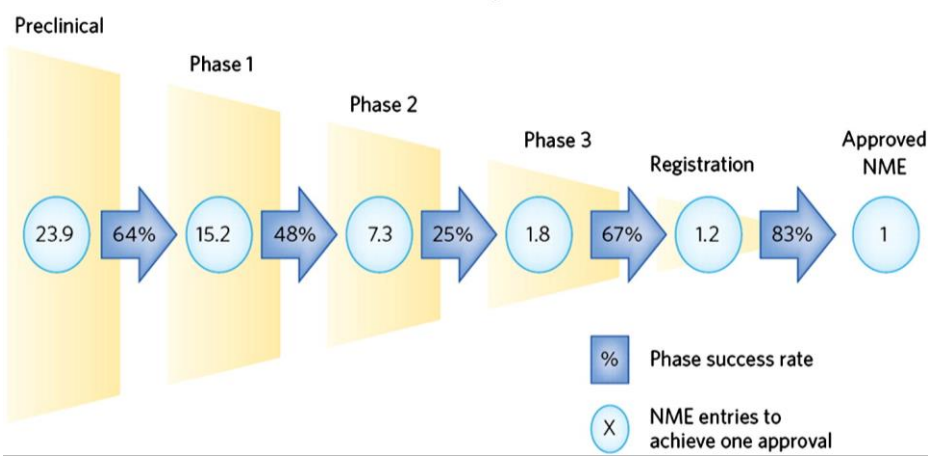


Figure 2.5 The success rate of new molecular entities (NMEs) through preclinical research and development (R&D) and clinical development phase (Csermely, et al., 2013)

2.2 Adverse Drug Reactions (ADRs)

Safety issues are necessarily emerged throughout the history of drug. From preclinical screening through clinical trials and, significantly, after the marketed for a myriad of populations, drug safety is investigated. Even though ADRs are relatively rare once a drug is marketed, they can lead to drug withdrawals. Furthermore, ADRs increase unnecessary hospital stay and are numbered to be the fourth leading cause of death in the United States, not far behind cancer and heart disease (Giacomini, et al., 2007; Wu, et al., 2010). ADRs are acknowledged to shorten not only the treatment failure, but also advance quality of life or medication confidences. It would be worthwhile to pharmaceutical industry if ADRs can be prevented.

2.2.1 Definitions

Mostly, the context of adverse drug reactions is defined ambiguous. The fundamental definition is “An injury resulting from medical intervention related to a drug” which unclear for the word “injury” and “medical”. The standard statement of meaning, WHO’s definition which has been used more than 30 years, is “A response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological

function". Even though, the word "noxious" is loose explanations that include all minor reactions, for instance, slight dryness of the mouth. This global statement crushes the surveillance systems that WHO currently operate. Thus, Edwards and Aronson (2000) characterize ADRs:

> "The appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product".

The terms 'adverse drug reaction' and 'adverse drug effect' are interchangeable. Adverse drug reaction is seen from the point of view of patient, whereas adverse drug effect is seen from point of view of drug. Moreover, the terms adverse drug reaction and adverse drug effect are supported to 'toxic effect' and 'side effect'. Toxic effect is an overestimation of the desired therapeutic effect at normal dose and the same mechanism as the therapeutic effect. An example of toxic effect is a calcium antagonist caused headache by vasodilatation that is the same mechanism as the therapeutic effect. Otherwise, side effect occurs via some other mechanism of therapy and may be dose-related or not. The example of side effect is anaphylaxis from penicillin usage.

2.2.2 Epidemiology

ADRs are the major cause of morbidity and mortality worldwide. They occur in everyday medical practice. For example, drug induced morbilliform rash is the commonest adverse reaction (Riedl and Casillas, 2003). 14.7% of French and 17% of Swiss hospital inpatients have given reliable histories of systemic adverse reactions to one or more drugs (Vervloet and Durham, 1998). From US hospitals, 6.7% of inpatients have serious ADRs and 0.32% of these have fatal reactions that cause about 100,000 deaths per year. This number drives ADRs between the fourth and sixth leading cause of death in hospital inpatients (Meyer, 2000). In addition, the estimated 10,000 Canadians die annually from ADRs (Kondro, 2005). Wu, et al. (2010) summarized the annual number of total hospital admissions in ten years. They found 557,978 (0.9%) from 59,718,694 emergency hospital admissions were diagnosed code indicative of ADRs. From these incidents, 26,399 (4.7%) died. The details of Wu, et al. study are shown in Table 2.1.

Table 2.1 Total number of emergency admissions which diagnosis of ADRs from 1999 to 2008, modified from Wu, et al. (2010)

Year	Total number of admissions	Number with drug-induced codes	Total admissions due to ADRs	In-hospital mortality of ADRs admissions	In-hospital mortality rate of ADRs admissions
1999-2000	5,321,796	16,434	42,453	1,816	4.3
2000-2001	5,319,791	16,237	43,288	1,879	4.3
2001-2002	5,337,034	17,151	45,518	2,259	5.0
2002-2003	5,494,066	16,645	47,455	2,480	5.2
2003-2004	5,818,057	18,307	53,218	2,672	5.0
2004-2005	6,116,507	19,652	56,629	2,679	4.7
2005-2006	6,423,646	20,465	61,931	2,878	4.6
2006-2007	6,496,945	20,009	64,536	3,013	4.7
2007-2008	6,560,785	20,685	67,874	3,189	4.7
2008-2009	6,830,067	22,439	75,076	3,534	4.7
Change (%) 1999-2008	28.3	36.5	76.8	94.6	10.0
Change (%) 2004-2008	11.7	14.2	32.6	31.9	-0.5

2.2.3 Classification

In modern pharmacology, classification of ADRs is primary differentiated from dose-related to non-dose-related reactions. Diagram of elementary ADRs differentiation is exemplified in Figure 2.6.The categorization was first designated as Type A and type B reactions, respectively (Vervloet and Durham, 1998; Riedl and Casillas, 2003; Thien, 2006).

Type A reactions are the most, about 80% of ADRs, which predictable and dose-dependent. This type of reactions includes toxic effects, side effects, and drug-drug interactions.

Type B reactions, comprise of 10–15% of ADRs, are hypersensitivity reactions that are unpredictable and non-dose-dependent. This type of reactions is divided into two subtypes: immune-mediated (5-10%) and non-immune-mediated (5-10%). Drug allergies are hypersensitivity reactions that involve an immune mechanism (IgE-mediated, T cell-mediated, or an immune complex or cytotoxic reaction). Unpredictable non-immune-mediated reactions can be classified as pseudoallergic, idiosyncratic, and intolerance. Pseudoallergic reactions are the result of direct mast cell activation and degranulation by drugs. They may indistinguishable from Type I hypersensitivity, but do not involve drug-specific IgE. Idiosyncratic reactions are pharmacologic reactions that occur only in a small percent of the population. Drug intolerance is a lower threshold to the normal pharmacologic action of a drug.

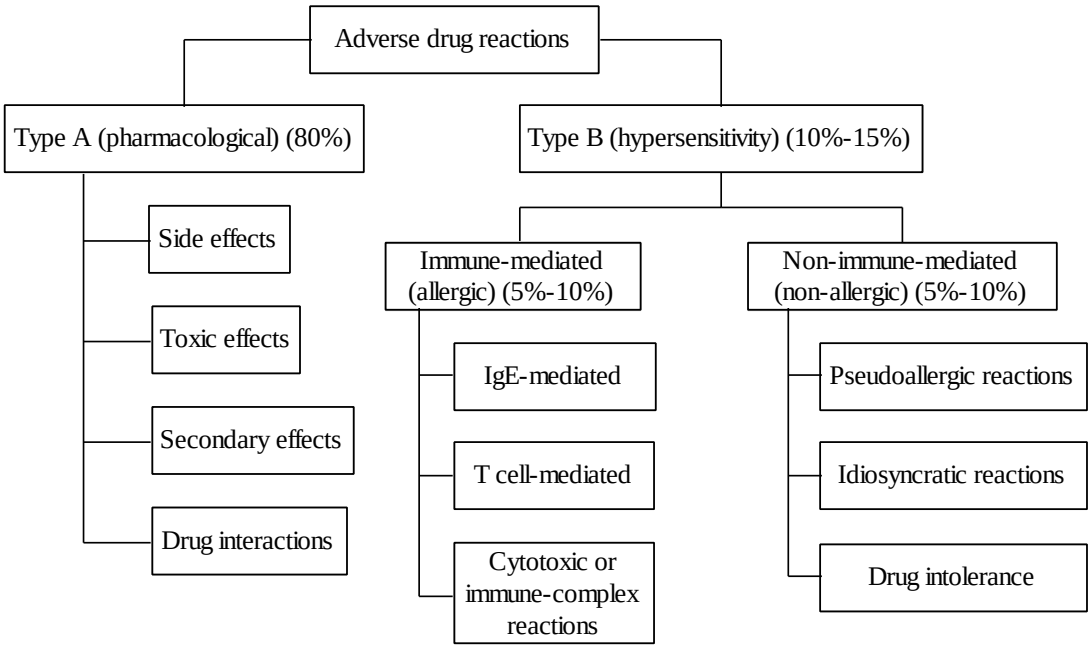


Figure 2.6 Type A and Type B classification of adverse drug reactions, modified from Thien, (2006)

Furthermore, two types of reaction were added: reactions related to both dose and time (type C) and delayed reactions (type D). The last category can be distributed into two: time-related reactions and withdrawal effects. In addition, a sixth category has been proposed: unexpected failure. Table 2.2 shows the complete ADRs classification with mnemonic purposes, features and examples in each category (Edwards and Aronson, 2000).

Table 2.2 Classification of adverse drug reactions, modified from Edwards and Aronson (2000)

Type of reaction	Mnemonic	Features	Examples
A: Dose-related	Augments	• Common • Related to a pharmacological action of the drug • Predictable • Low mortality	• Toxic effects: Digoxin toxicity; serotonin syndrome with SSRIs • Side effects: Anticholinergic effects of tricyclic antidepressants
B: Non-dose-related	Bizarre	• Uncommon • Not related to a pharmacological action of the drug • Unpredictable • High mortality	• Immunological reactions: Penicillin hypersensitivity • Idiosyncractic reactions: Acute porphyria, Malignant hyperthermia Pseudoallergy (eg, ampicillin rash)
C: Dose-related and time-related	Chronic	• Uncommon • Related to the cumulative dose	• Hypothalamic-pituitary-adrenal axis suppression by corticosteroids
D: Time-related	Delayed	• Uncommon • Usually dose-related • Occurs or becomes apparent some time after the use of the drug	• Teratogenesis • Carcinogenesis • Tardive dyskinesia
E: Withdrawal	End of use	• Uncommon • Occurs soon after withdrawal of the drug	• Opiate withdrawal syndrome • Myocardial ischaemia (β-blocker withdrawal)
F: Unexpected failure of therapy	Failure	• Common • Dose-related • Often caused by drug interactions	• Inadequate dosage of an oral contraceptive, particularly when used with specific enzyme inducers

2.2.4 Management

Rapid action is sometimes important because of the serious nature of suspected ADRs, such as anaphylactic shock (Edwards and Aronson, 2000; Riedl and Casillas, 2003; Thien, 2006). The most important and effective in managing ADRs is the discontinuation all of the offending medication, if possible. Alternative medications with unrelated chemical structures should be substituted when available, otherwise, medicine or medicines judgment using clinical benefit-risk should be considered when drug withdrawal as a trial.

If several medicines are introduced, the need for the drug, the severity of the reaction, and its potential for treatment have to be taken for a benefit-risk decision. The non-essential medicines should be withdrawn first, preferably one at a time. If the reaction is likely to be dose-related, dose reduction should be considered. During medication withdrawal, the patient should be observed. The monitored period could vary depending on the rate of elimination of the drug and the type of pathology.

If the patient cannot remain without a medicine that has caused an adverse reaction, continuing the essential treatment, meanwhile, provide symptomatic relief could be the option. Though when treating an adverse reaction, it is important not to submit more medicines than are essential.

2.3 Database in Pharmacology

A database is a collection of interrelated stored data (Teorey, et al., 2008). The integrating collection of many different types supplies the needs of multiple users within one or more groups. The reasons for using databases rather than files include greater availability to a diverse set of users, integration of data for easier access to and updating of complex transaction, and less redundancy of data (Lacroix, 2002; Hernandez and Kambhampati, 2004; Louie, et al., 2007). In pharmacology, numerous databases have been introduced with the same motivations as general usage in computer science. Each database in pharmacology accumulates various types of data depending on its purpose. In this review, databases in pharmacology were classified into three categories; chemical-target, chemical-response, and genetic-response. However, some databases may be overlapping classification.

2.3.1 Chemical-Target

Most of pharmacological databases were classified into this category. Examples and details of each database were described below.

2.3.1.1 TTD (Therapeutic Target Database)

Numbers of proteins, nucleic acids, and other molecular entities have been discovered as therapeutic targets that effect by binding to and modulating the activity of a particular target. TTD (Chen, et al., 2002; Zhu, et al., 2010) was developed to provide information about the known therapeutic targets described in the literatures, the targeted disease

conditions, the pathway information, and corresponding drugs/ligands directed at each of these targets. The intention of this database is to provide comprehensive information about the primary targets and other drug data for the approved, clinical trial and experimental drugs. TTD is case insensitive search by target name, drug/ligand name or function, disease name, or drug therapeutic classification. Significantly, the updated TTD by Zhu, et al. (2010) increased data to 1,894 targets, 560 diseases, and 5,028 drugs, compared with 433 targets, 125 diseases, and 809 drugs in original released data by Chen, et al. (2002). The elements of targets were composed of 348 successful, 292 clinical trial, and 1,254 research targets while drugs were composed of 1,514 approved, 1,212 clinical trial and 2,302 experimental drugs. TTD can be accessed at http://bidd.nus.edu.sg/group/ cjttd/ttd.asp.

2.3.1.2 STITCH (Search Tool for Interactions of Chemicals)

STITCH (Kuhn, et al., 2008; Kuhn, et al., 2010b) is the database of accumulated information about interactions between protein and small molecules (drugs or drug-like molecules) which can be traced back to the original data sources. It intends to integrate dispersed data over the literature and various databases of biological and metabolic pathways, crystal structures, binding experiments, and drug-target relationships. STICH displays the interactions in network which can be explored interactively or used as the basis for large-scale analyses. Kuhn, et al. (2010b) had developed STICH 2.0 that connects protein from 630 organisms to over 74,000 different chemicals, including 2,200 drugs. STITCH can be accessed at http://stitch.embl.de/.

2.3.1.3 PDTD (Potential Drug Target Database)

PDTD (Gao, et al., 2008) is a dual function database that associates an informatics database to a structural database of known and potential drug targets. It queries drug target information and identifies the potential binding proteins of an active compound or an existing drug by using reverse docking approach. The comprehensive web-accessible database focuses on those drug targets with known 3D-structures. As of September 2010, PDTD had collected 1,207 entries covering 841 known and potential drug targets from Protein Data Bank (PDB) and literatures extractions. Drug targets of PDTD were categorized into 15 and 13 types according to two criteria: therapeutic areas and biochemical criteria. PDTD is available at http://www.dddc.ac.cn/pdtd/.

2.3.1.4 KEGG DRUG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is an integrated database resource that consists of 16 main databases. The KEGG DRUG database (Kanehisa, et al., 2010) is one of these projects. It is a chemical structure-based and component information resource for all prescription in Japan including crude drug and TCM (Traditional Chinese Medicine) formulas. Most prescription drugs in the USA and many prescription drugs from Europe are also accumulated in this database. As of September 2009, KEGG DRUG contained about 9,000 chemical structures and therapeutic efficacy of drugs. In addition, it contained two types of molecular networks. The first was a molecular interaction network denoting interaction and/or relations with target molecules, drug

metabolizing enzymes, drug transporters, and other drugs. The second was the network of chemical structure changes in small molecules which covered series of chemical modifications introduced by medicinal chemists in the history of drug development, secondary metabolic pathways for biosynthesis of druggable natural products and drug metabolism. KEGG DRUG is available at http://www.genome.jp/kegg/drug/.

2.3.1.5 DrugBank

DrugBank is an unrivalled bioinformatics/cheminformatics resource that combines detailed drug data with comprehensive drug target information. Some databases, for example, KEGG DRUG, ChEBI, and PubChem are not specifically designed to be drug database that provide specific pharmaceutical information or link to specific drug targets. Conversely, others, such as online pharmaceutical encyclopedias, tend to offer much more detailed clinical information about many drugs but they are not designed to contain structural, chemical or physicochemical information. DrugBank solves these problems by combining the strengths of each data source to create a single drug resource that links sequence, structure, and mechanistic data about drug molecules with sequence, structure and mechanistic data about their drug targets (Wishart, et al., 2006). In addition, DrugCard is a summary table describing the drug of interest in much greater detail. It contains more than 100 data fields with half of the information being devoted to drug or chemical data and the other half being devoted to pharmacological, pharmacogenomics and molecular biological data. The latest version of DrugBank (release 2.0) had been expanded enormously to cover about 4,800 drug entries. Moreover, Wishart, et al. (2008) added two new drug categories: Withdrawn drugs and Illicit drugs. Table 2.3 present the content data comparison between DrugBank (release 1.0) versus DrugBank (release 2.0). DrugBank is available at http://www.drugbank.ca.

Table 2.3 Comparison between the data content in DrugBank (release 1.0) versus DrugBank (release 2.0), modified from Wishart, et al. (2008)

Category	**Release 1.0**	**Release 2.0**
No. of FDA-approved small molecule drugs	841	1,344
No. of biotech drugs	113	123
No. of nutraceutical drugs	61	69
No. of withdrawn drugs	0	57
No. of illicit drugs	0	188
No. of experimental drugs	2,894	3,116
No. of total Small molecule drugs	3,796	4,774
No. of names/brand names/synonyms	18,304	28,447
No. of data fields	88	108
No. of search types	8	12

2.3.2 Chemical-Response

The examples of chemical-response pharmacological database were detailed below.

2.3.2.1 TRMP (A Database of Therapeutically Relevant Multiple Pathways)

The cross-talks between proteins of different pathways are common phenomena and often implicate therapeutic efficacies. TRMP (Zheng, et al., 2004) stores the therapeutically relevant multiple pathways. Therefore, it gives information for facilitating the analysis of the potential implications on multiple target-based therapies and understanding of how therapeutic targets interact with other molecules in disease and physiological processes. This database contained 11 entries of multiple pathways, 97 entries of individual pathways, 120 targets covering 72 disease conditions along with 120 sets of drugs directed at each of these targets. TRMP can be accessed at http://bidd.nus.edu.sg/group/trmp/trmp_ns.asp.

2.3.2.2 SIDER (Side Effect Resource)

Predicting the possible side effects of drug candidates based on the binding pattern, chemical structure, and other properties is attracted to pharmaceutical industry. Side effects can also be used to predict novel drug-target interactions and might be utilizable for drug re-purposing. Kuhn, et al. (2010a) had compiled package inserts from several public resources associated with FDA. The standardized Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) were used as the basic lexicon of side effects and drug names were mapped to PubChem identifiers. SIDER contained 62,269 drug–side effect pairs that covered a total of 888 drugs and 1,450 side effects terms. About 70% of drugs have between 10 and 100 different side effects (Figure 2.7A), whereas about 55% of all side effects occur for <10 drugs (Figure 2.7B). SIDER is freely available for academic research at http://sideeffects.embl.de.

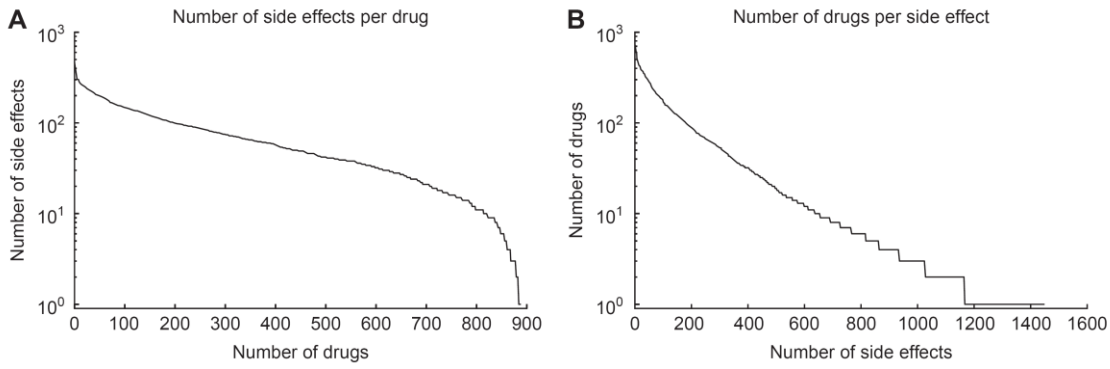


Figure 2.7 Statics of SIDER (Kuhn, et al., 2010a)

(A) A plotted graph of the number of side effects per drug is shown about 200 drugs with at least 100 side effects. (B) A plotted graph of drugs per side effects

2.3.3 Genetic-Response

Two databases were evaluated that they contained the information about genetic and drug response. The lists and their features were explained as follow.

2.3.3.1 PharmGKB (The Pharmacogenetics Knowledge Base)

PharmGKB (Hewett, et al., 2002; Klein and Altman, 2004) is a central repository of genotype and phenotype information that relate to pharmacogenetics and is collected by laboratories in research network. The objective of the PharmGKB is to boost research in the field of pharmacogenetics. PharmGKB models information about cellular phenotypes and includes clinical content. The database has a large collection of genotypes for genes of pharmacogenetics interest, for instance, 150 genes under study and a large ontology of pharmacogenetics concepts, which shows the patterns of polymorphisms identified in different populations. Furthermore, it also provides tools for submitting, editing viewing, and processing the information. Researchers can submit genomic information, drugs, diseases, populations, and so on through web forms or in uploaded files containing PharmGKB-defined XML elements. In 2002, PharmGKB had incorporated over 600 different relationships that will always be available for interpretation and correlation as new hypotheses emerge. This database is available at http://www.pharmgkb.org/.

2.3.3.2 PharmGED (Pharmacogenetic Effect Database)

Knowledge about protein polymorphisms of drug-related proteins and individual drug responses significantly promote pharmacogenetics study and individual prediction of drug responses. Zheng, et al. (2007), therefore, had developed PharmGED to provide the information about effects of a particular protein polymorphism, non-coding region mutation, splicing alteration, and expression variation on the response of a particular drug. Entries of this database can be searched by protein name, drug/ligand name, disease name, and drug class. The search is case insensitive and wildcards are supported. The wild characters are '?' and '*' that represent any single character and a string of character of any length respectively. In 2007, PharmGED contained 1,825 entries covering 108 disease conditions, 266 distinct proteins, 693 polymorphisms, and 414 drugs/ligands cited from 856 references. This database can be accessed at http://bidd.cz3.nus.edu.sg/phg/ that free of charge for academic use.

2.4 Biological Network

High-throughput (HT) technologies, such as DNA microarray, potentiate us to enumerate and study the dynamics and mechanisms of biological components in cell as systems view (Liao, et al., 2003). This concept shares structural principles with engineered networks (Alon, 2003). Specific components and their interacting partners or substrates can be employed to assemble high-confidence pathways. The topological structures yield valuable information about the functions of individual components and unexpected relationships between components and cellular processes. The cell comprises various types of interaction webs, or networks. None of these networks

function independently. Instead, they form a ‘network of networks’ that is responsible for the behavior of the cell (Joyce and Palsson, 2006). The integration of all interactions may reveal the ultimate description of how complex biological processes occur and can be controlled.

2.4.1 Concepts and Principles

Cellular interactions are generally assembled into network maps. The network description allows application of tools and concepts developed in fields, for examples, graph theory, physics, and sociology that have dealt with network problems before. They comprise of biological molecules (proteins or genes) entitled vertices or nodes and interactions between them defined as edges (in undirected networks) or arcs (in directed networks) (Alon, 2003; Zhu, et al., 2007). The directionality of a network depends on the characteristics of the biological data. Protein-protein and genetic interactions are usually represented with an undirected network. On the contrary, transcription factor binding, phosphorylation, and metabolic networks have directions on their interactions. Other feature is the strength of interactions. However, this information is rarely used in most network analyses (Zhu, et al., 2007). The understanding of network architecture and performance are represented by network topology. The most important and commonly used topological feature is degree (Figure 2.8). *Degree* is the number of connections linked to one node. A node with high degree (hub) may play an important role in maintaining the network structure.

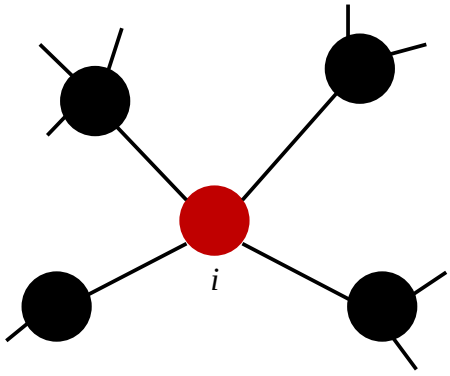


Figure 2.8 Degree: the number of links connected to node *i*.

Surprisingly, biological networks share three structural principles with engineered networks (Alon, 2003). The first principle is *modularity*, a set of nodes that have strong interactions and common function. For example, proteins work in slightly overlapping and regulate within groups as pathways. These complexes are optimized by evolutionary process. The second principle is *robustness to component tolerances*. In both engineering and biology, the network design has to operate under all interferences that come from intrinsic properties of the components or the environment. The third principle is *the use of recurring circuit elements* to operate the thousands of occurrences. As operational amplifiers and memory registers in electronic device, metabolic networks apply regulatory circuits, for instance, feedback inhibition and

network motifs in transcriptional network can perform a specific information processing task. These concepts ultimately allow characterizing and understanding the laws of nature that evolved and designed systems.

2.4.2 Biological Data Integration

The amount of data in biology has indeed grown exponentially over the past decade. This data is available in a wide variety of formats, annotated, and stored in flat files and relational or object-oriented databases. Each data type alone has a limited utility because technologies that investigate biological systems have inherently high false-positive and false-negative rates. To generate an effective knowledge from data, scientists must integrate these large and diverse data sets (Hwang, et al., 2005; Louie, et al., 2007). The integration of multiple data types provides the greatest information about a particular cellular mechanism. Increment of integration experiments help to demonstrate the power of combining and correlating several data domains. This covers not only biological notions of sequence, expression, interaction, localization, and variation but also compounds in an approach called chemogenomics (Table 2.4).

There are four problems that make data integration important (Hernandez and Kambhampati, 2004; Hwang, et al., 2005). First, the variety of data covers several biological and genomic research fields. They store different types of data, different degree of reliability, and different amounts of error. Second, similar data can be contained in several sources but represent heterogeneity depending on the source. This representational heterogeneity includes structural, naming, semantic, and content differences. Third, web-based sources operate autonomously. They are instability and unpredictability that are free to modify and/or remove data. Finally, Individual sources serve their own user-access interface, which means different querying capabilities.

Table 2.4 Examples of large-scale scientific data integration across domains, modified from Searls (2005)

Domain	Description
Sequence x Expression	Novel transcription elements are discovered by aligning the upstream regions of genes that are co-expressed in microarray experiments and detecting especially conserved sites.
Sequence x Interaction	Phage display finds consensus sequences of ligands for peptide recognition modules, from which protein-interaction maps are inferred by scanning whole genomes and then tested by yeast two-hybrid methods. Interaction data can also be used to discover *cis*-regulatory motifs in upstream sequences.
Interaction x Expression	By integrating protein-interaction data with microarray results it is possible to reconstruct complex signaling pathways accurately, without prior knowledge of pathway intermediates. Interaction, expression and other functional data sources can be integrated to predict participation in protein complexes.
Sequence x Localization	Subcellular localizations of gene products are assessed by extensive motif analysis of cDNA sequences and by high-throughput tracking of expressed fusion proteins in living cells, demonstrating good mutual confirmation.
Variation x Expression	Natural variation in microarray-based gene expression levels segregating as quantitative traits in human families is traced to specific chromosomal regions, representing *cis*- and *trans*-acting loci as well as putative 'master regulators', by linkage analysis using a database of single-nucleotide polymorphisms (SNPs).
Variation x Interaction	Evidence for evolutionary co-variation of interacting protein families can be seen in correlated mutations and similarities in phylogenetic tree topologies, according well with databases of known structural domain interactions.
Compounds x Sequence	Kinases clustered on both compound selectivity data and sequence similarity produce comparable dendrograms in all but higher-level groupings. Yeast deletion screens find proteins that functionally interact with compounds to inhibit cellular proliferation, finding both known and novel on- and off-target effects.
Compounds x Expression	Microarray platforms are used to profile compounds for *in vivo* effects over whole transcriptomes, where, for example, expression profiles have been used to accurately classify a variety of psychiatric drugs. Expression patterns can be related not only to drugs but to substructures and other chemical features.
Compounds x Interaction	New pharmacological approaches to protein–protein interaction, such as inhibitors of dimerization or allosteric modulators, can benefit from interaction maps that help characterize interfaces, mutation studies that pinpoint crucial amino-acid residues within large contact surfaces and database-driven design.

2.5 Knowledge Discovery (KD)

The current trends of technology inexorably lead to data flood. Huge amounts of data, especially as in text, have been accumulated at a very fast pace. Although text expresses the vast and wealthy knowledge, pure text is not useful and meaningful. Information is hidden and can be seen as the patterns or characteristics of the data. The new discipline, called *knowledge discovery* (KD), has thus emerged to make sense and use of data (Chen, et al., 1996). It is a highly complex and demanding process that requires careful analysis, specification, implementation and testing. Here two methods of KD, data mining and literature mining, were reviewed.

2.5.1 Data Mining

Data amass like earths and rocks in a mountain that most of them are not useful. The valuable materials are needed to be dug out. Uncovering valuable knowledge may as well be excavated from a large amount of data. There is a huge amount of veiled information that is potentially important but has not yet been discovery or expressed. Data mining is therefore becoming more popular and is needed for efficient data analysis. It is the process of extracting valid, previously unknown, comprehensible, and actionable information from large data and then utilizes it to make crucial decisions (Hsu, 2006).

2.5.1.1 Data Mining Process

Data mining consists of six processes (Wirth, 2000), named Cross Industry Standard Process for Data Mining (CRISP-DM) and illustrated in Figure 2.9. The sequence of procedure is not strict. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle represents the cyclic nature of data mining. A data mining process continues after a solution has been set out. In the following, each process of CRISP-DM is briefly delineated.

Business understanding is the primary phase that focuses on understanding the project objectives and requirements. It then converts the knowledge into a data mining problem explanation and designs the method to achieve the purposes.

Data understanding starts with a data collection. Only proper data is useful to be mined. Domain experts are needed for the selection of data for certain problems.

Data preparation covers all actions to construct the final data that will be fed into the modeling. The data preparation includes table, record and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools. This task is likely to be performed multiple times and seems to be the longest period of data mining.

Modeling usually involves building a model for the data. Typically, different algorithms and techniques are conducted for the same data mining problem type. Some techniques have specific requirements on the form of data that stepping back to the data

preparation phase is therefore often needed. However, for a certain task, suitable techniques should be chosen.

Evaluation of the model requires to be interpreted by human experts from its correctness, comprehensibility, and usefulness. The model should have the high quality from a data analysis perspective and properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment will be needed to increase knowledge of the data. Creation of the model is not the end of the project. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

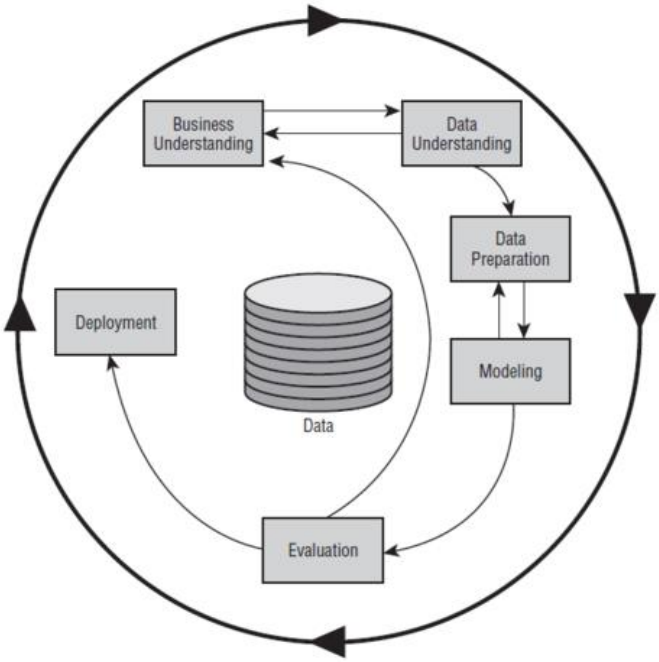


Figure 2.9 Process diagram of Cross Industry Standard Process for Data Mining (CRISP-DM), modified from Wirth (2000)

2.5.1.2 Classification

Any form of interesting information that can be discovered from the data can formulate a specific data mining task. There have many different types of data mining tasks, for example, classification, clustering, and association. Some of them might overlap with others. Only the most popular task called classification was reviewed.

Classification is a learning method for predicting the class or group of unseen instance from pre-labeled (classified) instances (Witten and Frank, 2005; Hsu, 2006). There should be at least two classes. The input of a classification model is the attributes of a data sample and the output is the class that data sample belongs to. Classification takes supervised learning to build a model. A set of data with known classes (training data) is needed to estimate the parameters of the classification model. In numeric determination, the outcome to be predicted is not a discrete class but a numeric quantity. After the

parameters are set, the model can be applied to automatically classify the new data samples. There are many approaches of classification, for example, decision tree, nearest neighbor, neural networks, and support vector machine.

Decision Tree

A decision tree is a judgment support tool that applies a tree-like flowchart for the possible consequence. From its structure, a node symbolizes the test on an attribute, each branch stands for outcome of test, and leaf node represents class label. A divide-and-conquer approach to the problem of learning is a technique of decision tree. An attribute is placed and splits branches for every value of the attribute. The process is repeated recursively for each branch. Consequently, the tree stops to develop at the leaf node when all instances are the same class or almost homogenous. Even though decision tree is simple to understand, illustrate, and interpret, its calculations can get very complex particularly if many values are uncertain.

Nearest Neighbor

One of the classic forms of learning is plain memorization. When a set of training instances has been learned, a new instance is searched for the most resembles the training data. So instance-based learning is lazy that the closest existing instance is used to assign the class to the new one. Sometimes more than one nearest neighbor is employed and the majority class of the closest k neighbors is given to the new instance. This is termed the k-nearest-neighbor method. Although nearest neighbor is simple and effective, it is generally slow. The way to find which member of the training set is closest to an unknown test sample is to calculate the distance from every member of the training set and select the smallest. The time it takes to make a prediction is thus proportional to the number of training instances.

Neural Networks

Unlike decision tree which divides the space of examples using straight lines, neural networks can be thought as more complex regions. The inspiration of neural networks comes from biological networks of central nervous systems. They are presented as systems of interconnected neurons that compute values from inputs by feeding information through the network. The message is weighted and passed on. When the weight associated with an input corresponds to the threshold of the perceptron, an output neuron is activated. Although neural networks are successfully applied to a wide range of supervised and unsupervised learning problems, the comprehensibility of learned models and the time required to induce models from large data sets are two fundamental considerations.

Support Vector Machine (SVM)

There is resurgence of interest in linear models with the introduction of SVM, the combination of linear modeling and instance-based learning. SVM selects a small number of critical boundary instances, called support vectors, from each class and builds a linear discriminant function, named maximum margin hyperplane, which

separates classes as widely as possible. A linear model constructs the new space that can represent a nonlinear decision boundary in the original space. These systems transcend the limitations of linear boundaries by making it practical to include extra nonlinear terms in the function. This is possible to form quadratic, cubic, and higher-order decision boundaries.

2.5.2 Literature Mining

The amount of biomedical literature available online continues to grow rapidly today. PubMed/Medline, the largest published literature repositories in the biomedical world, contains more than 19 million citations and abstracts from about 5,000 journals (Yu, et al., 2007; Lok, 2010). These supply nearly 830,000 articles published in 2009, up from some 814,000 in 2008 and around 772,000 in 2007. The advance of genome sequencing, the increasing number of genes addressed in single studies, and HT experimentation generate the enlargement of publications. Moreover, the growth rate exhibits no signs of decreasing, especially as becoming appearance countries such as China and Brazil that speed up their research. These occurrences cause the readers would stop either when they found the first instance of the necessary fact or after reaching their attention/frustration limit. At best, the manual approach does not explore the complete range of values available from different sources. In contrast, it leaves many of the values blank at worst. Additionally, it is no longer possible for a researcher to keep up-to-date with all the relevant literature manually, even on specialized topics (Hale, 2005). There are so many techniques that attempt to identify, extract, manage, integrate, and discover novel or hidden or unsuspected knowledge. Even though computer can rapidly process and integrate this wealth of information, an overwhelming amount of biomedical knowledge is recorded in electronic texts and is written down in natural language and pictures. Computer, like a human, has difficulty in making sense of these and need specialized knowledge in order to understand (Rodriguez-Esteban, 2009). To overcome this difficulty, literature mining has been developed to systematically compare large data sets with all the knowledge that is derived from the published data, which allows the biological relevance of the data to be interpreted (Krallinger, et al., 2005).

Literature-mining tools are becoming essential to researchers. They empower researchers to identify relevant papers, recognize entity, and pull out specific facts. The advanced tools, called text mining, are based on these methods. More than literature mining, text mining can automatically discover the novel models and understand patterns from large amounts of data (Zaki, et al., 2007). Even literature mining and text mining are used interchangeably, literature mining is more general term. Here the most important three methodologies, information retrieval, entity recognition, and information extraction, that are used for both literature and text mining were briefly described (de Bruijn and Martin, 2002; Ananiadou and Mcnaught, 2005; Krallinger, et al., 2005; Jensen, et al., 2006; Rzhetsky, et al., 2009).

2.5.2.1 Information Retrieval (IR): finding the papers

IR systems aim to identify the text segments (full articles, abstracts, paragraphs, or sentences) in a collection which match a user's query. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. IR technologies are in wide-spread use. The most well known IR systems are search engines such as Google, which identify those documents on the website that are relevant to a set of given words. As part of biomedical world, PubMed is the best-known IR systems. Most experimental biologists take advantage of the PubMed information-retrieval system available at the NCBI. PubMed is an *ad hoc* system that uses two established IR methodologies: the "Boolean model" and the "Vector model". The Boolean model enables the user to retrieve all documents that contain certain combinations of terms by using a logical operation. The Vector model, by contrast, typifies each document by a term vector (a value according to a frequency-based weighting scheme) and compare to a query vector. *Ad hoc* IR systems generally give more difficulty than text-categorization systems in dealing with the many abbreviations, synonyms, and ambiguities in biomedical terminology. PubMed and many other good biomedical IR systems thus apply thesauri to automatically expand the query with other related terms.

2.5.2.2 Entity Recognition (ER): indentifying the substance(s)

The identification of entity types in textual data is known as 'name entity recognition' or 'semantic tagging'. In molecular biology, most of these entities are molecules, such as genes and proteins that have many of their aliases (objects, concepts, and symbols). The seemingly modest goal of ER is to find the biological entities that are mentioned within a text. This task is often separated into two sub-tasks. First task is the recognition of words. It refers to entities that take advantage of string regularity and write patterns to capture the known naming conventions. Second is the unique identification of the entities in question. The latter task is lexicon based that uses name lists to tag terms, or likely components of entity names. The main difficulty in ER occurs from the lack of standardization of names. Each gene or protein typically has several names and abbreviations. The recent improvement for resolving ambiguity in gene or protein names is therefore an urgent for ER. Although ER might at first seem neither challenging nor useful, it is possibly the most difficult task in biomedical literature mining and is an essential for both information extraction and information retrieval.

2.5.2.3 Information Extraction (IE): formalizing the facts

Readers do not understand text if they only know the entities. They must also realize the interactions or relationships between those entities. In contrast to IR systems, IE systems propose to extract pre-defined types of fact, in particular, relationships between biological entities. Information extraction attempts to identify biologically meaningful semantic structures within free text. An example of the using IE applications in molecular biology is the identification of protein interactions. There are two different fundamental approaches to extracting relationships from biological texts: co-occurrence and natural-language processing (NLP).

Co-occurrence

The simplest and straightforward approach to capture entity relationships is to search for sentences that frequently mention two entities within abstracts or sentences. If two entities are repeatedly mentioned together, it is likely that they are somehow related, although the type of relationship is not known. Co-occurrence methods tend to give better recall and sensitivity. But co-occurrence produces worse precision and specificity than natural-language processing (NLP). Besides its defect, co-occurrence method arise erroneous extracted relationships from complex sentences that contain multiple relationships.

Natural-language processing (NLP)

This method combines the analysis of syntax and semantics. A syntax tree is derived for each sentence to delineate noun phrases and represent their interrelationships. Subsequently IE is semantically tag the relevant biological entities and other keywords. Finally, a rule set is utilized to extract relationships on the basis of the syntax tree and the semantic labels. Unfortunately, most NLP systems are unable to extract relationships that extend across multiple sentences. However, this is not a complete mischance. Because of relationships are usually mentioned within a single sentence, it may overcome the limitation.