CEME

A PURPOSE-BUILT TOOL FOR VALUE-DIFFERENTIATION

August 2015

KEITH McCORMACK PhD

Email: keithmccormack@mccormack-group.com

www.mccormackpharma.com

CONTENTS

DATA MINING, TEXT MINING AND ASSOCIATION RULE LEARNING

Data mining or knowledge discovery in databases is the detection of interesting patterns from large data collections. Association rule learning is one of the most well-studied data mining tasks that discovers interesting relationships among variables, producing “if-then” statements. An association rule X ⇒ Y proposes that within the database where X occurs, there is a certain probability (as defined by the values of concepts known as *support* (S) and *confidence* (C)) that Y will appear). Here, X and Y are defined respectively as the antecedent and the consequent parts of the rule. Since this rule is learned from data then the association between X and Y may constitute a previously-unknown relationship that likely was unsuspected and is interesting; consequently, this association may represent an important contribution.

There is a consensus that the concept of association rules was popularised largely by the 1993 article of Agrawal and coworkers (*Proceedings of the 1993 ACM SIGMOD international conference on management of data, SIGMOD ’93 p207*) which has acquired more than 6000 citations according to Google Scholar, and is thus one of the most cited papers in the data mining field. However, it is possible that what is now called "association rules" is similar to what appears in the 1966 publication on General Unary Hypotheses Automaton (GUHA) (*Computing 1966 1 293-308*), a general data mining method developed by Petr Hájek and coworkers (*In: Database Support for Data Mining Aplications, 2004 ISBN 978-3-540-22479-2, Springer*).

A simple example of association rule learning is the discovery of buying patterns in large-scale transaction data recorded by point-of-sale checkout systems in high-street supermarkets. The rule {onions, potatoes} ⇒ {white

wine} (S = 0.01; C=0.6) derived from the sales data of a supermarket gathered during July and August indicates that 60% of all customers that bought onions and potatoes also bought white wine, and 1% of all customers bought all three together. Given the extensive selection of vegetables within a large store then such an association is interesting and will assist the supermarket sales team in positioning goods on the shelves during the barbeque season!

The challenges of mining text

Searching text using keywords and phrases is “information extraction” and not mining. Moreover, by comparison with the kind of data stored in structured databases, natural language text is unstructured, amorphous, and difficult to mine using traditional algorithms. Nevertheless, by a process of logic and deduction information extraction can lead to the generation of new hypotheses.

For example, Swanson suggested in a series of creative papers that novel information might be unearthed by systematically studying seemingly unrelated and non-interactive research literatures, which he called "complementary but disjoint" *(For review see Bekhuis T Biomed Digit Libr 2006 Apr 3; 3:2).* Swanson and Smalheiser *(Artif Intell 1997 91 183-203)* recognized that two literatures might be spuriously linked because of shared language in the larger discipline. For example, while investigating causes of migraine headaches, Swanson *(J Am Soc Inf Sci 1987 38(4) 228-233; Artiff Intell 1997 91 183-203)* extracted information from titles of articles in the biomedical literature, leading to clues like these:

> *Stress is associated with migraines*
> *Stress can lead to loss of magnesium*
> *Calcium channel blockers prevent some migraines*
> *Magnesium is a natural calcium channel blocker…*

…Spreading cortical depression is implicated in some migraines
High levels of magnesium inhibit spreading cortical depression
Migraine patients have high platelet aggregatability
Magnesium can suppress platelet aggregatability

These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache, a hypothesis that did not exist in the literature at the time Swanson found these links. However, while the "hypothesis discovery" of Swanson is undoubtedly an invaluable contribution, generally the process is limited to keywords of diseases and mediators. In this way, while connections become apparent and result in the generation of a hypothesis, value-differentiation of a selected drug is a remote prospect unless individual drugs are cited or become included within the deductive process. That is, the hypothesis discovery method is limited by the selection of literature that defines the deductive process and is not powered to explore differences in clinical outcomes that exist between drugs sharing the same pharmacophore or are indicated for the same approved indication within the context of the new hypothesis. Consequently, Swanson's methods, whilst pioneering and wholly innovative, lack sensitivity for the routine application of discovering value-differentiation of a selected brand. But in any case, that was not what Swanson and coworkers intended.

The use of antecedent surrogate variables

Regardless of the origins of association rules learning, in the 1980s McCormack Pharma was pioneering the use of antecedent surrogate variables in association rules learning before such methods were widely used or even understood. Moreover, at that time the field of text mining was barely nascent. McCormack Pharma's use of surrogate variables offers a novel and innovative route for probing natural language text for the discovery of new and

interesting associations that otherwise would remain lost. The use of antecedent surrogate variables as a means to augment missing data is illustrated later in this document.

CEME EXPLORES QUASI-CHAOTIC SYSTEMS

Historically, the CEME methodology evolved from the fundamental tenet whereby a quasi-chaotic relationship exists between experimental data and real world clinical outcomes. That is, small differences between physical properties of drugs result in significant differences in real world clinical outcomes.

The chemical structure of a drug determines the nature (physical properties), and formulation and route of administration determine the character of effects *in vivo.* More generally, in a therapeutic context this axiom may be restated whereby for an approved drug, experimental data uniquely code for the drug's real world clinical profile and patterns.

For a deterministic system that is chaotic, final outcomes are not predictable and are sensitive to starting values. However, in a quasi-chaotic context-dependent system, by comparison with behaviour in a chaotic system small differences in starting values likewise result in large differences in final outcomes, but unlike a true chaotic system, the relationship between the initial conditions and the final outcome is fixed. In a quasi-chaotic system CEME displays exquisite sensitivity in differentiating drugs that share the same approved indication.

EXPERIMENTAL DATA CODE FOR REAL WORLD OBSERVATIONS

For a specific drug, given that experimental data uniquely code for the drug's real world clinical profile and patterns then within an approved indication value-differentiation is determined by some previously-unknown relationship between the drug and data that characterizes the approved indication. It is the clinical correlate of this previously-unknown relationship that shows with a high probability of success where value-differentiation (as real world evidence) resides within big data. As I reasoned almost three decades ago the probability of successfully discovering some previously-unknown relationship (decoding the experimental data) within a specific indication covaries with the number of surrogate variables. This postulate can be illustrated using a real case study.

Case Study

In the early 90s, the French manufacturer Rhone-Poulenc Rorer (RPR) asked McCormack Pharma to differentiate its nonsteroidal (NSAID) analgesic Orudis (ketoprofen) when used to treat and manage pain. Since ketoprofen had previously been shown to enter the cerebrospinal fluid (CSF) and achieve therapeutic levels then value-differentiation was sought within the management of postoperative pain by searching for some action of ketoprofen upon attenuating the phenomenon of central sensitization that had been reported by Woolf *(Br Med Bull 1991 47(3) 523-533)* as a critical mediator in the development of pain following surgery.

Does the experimental data code for some previously-unknown centrally-mediated effect of ketoprofen in managing postoperative pain? In order to answer the question, and as a prerequisite to probing the experimental data,

the first task was to build a list of surrogate variables for the broader picture of the postoperative stress response that subsumes perioperative/postoperative pain. The list of antecedent surrogate variables is extensive and includes…..

Human Growth Hormone (HGH)/Growth Hormone/Somatotropin/Somatropin
Somatostatin/Growth Hormone-Inhibiting Hormone (GHIH)/Somatotropin Release-Inhibiting Factor (SRIF)/Somatotropin Release-Inhibiting Hormone (SRIH)
Corticotrophin/Corticotropin/Adrenocorticotrophic Hormone (ACTH)
Cortisol/Hydrocortisone
Vasopressin/Arginine Vasopressin/Argipressin/Antidiuretic Hormone (ADH)
Glucagon
Insulin
Catecholamines/Adrenaline/Epinephrine/Norepinephrine
Anterior pituitary
Posterior pituitary
Hormone
Pro-opiomelanocortin
Thyroid-Stimulating Hormone (TSH)
Follicle-Stimulating Hormone (FSH)
Luteinizing Hormone (LH)
Hypothyroid
Euthyroid
Thyroid
Thyroxine (T_4)
Tri-iodothyronine (T_3)
Reverse T_3 (rT_3)…….and so on…

Mining text for associations between each NSAID and each member of this extensive catalogue of antecedent surrogate variables (including synonyms), CEME discovered the association rule:

(Ketoprofen, Hypothyroid) ⇒ ↑Tryptophan 2,3-dioxygenase

Translating the association rule

When in an *in vivo* hypothyroid state ketoprofen increases the activity of the hepatic enzyme tryptophan 2,3-dioxygenase (TDO). In both adult rats and humans, an increase in activity of hepatic TDO is associated with an increase in

the level in brain and CSF of the tryptophan metabolite, kynurenic acid *(Life Sci 1977 21 755-768; Drug Chem Toxicol 1985 8 145-154; Biochem J 1986 234 635-647; Kynurenine and Serotonin Pathways, Plenum Press, New York, 1991 pp 281-288; Toxicol Appl Pharmacol 1990 102 251-258; Kynurenine and Serotonin pathways, Plenum Press, New York, 1991 pp 299-308; J Neurochem 1991 57 1630-1635; Life Sci 1991 49 527-534; Kynurenine and Serotonin Pathways, Plenum Press, New York, 1991 pp 185-199)*, the only known endogenous excitatory amino acid receptor antagonist in mammalian brain which selectively antagonizes activity of the N-methyl D-aspartate (NMDA) in the spinal cord, especially at concentrations below those normally required for antagonism of non-NMDA receptor activity *(see Pain 1994 59(1) 9-43 for full discussion and supporting bibliography).*

The evolution of new teachings

Kynurenic acid is an endogenous selective antagonist of the NMDA receptor within the spinal cord that acts at the strychnine-insensitive glycine site within the receptor. Importantly, thyroid function is depressed/inhibited (=hypothyroid) in post-surgical patients, especially after general anaesthesia.

Is surgical stress a covariate for the analgesic efficacy of ketoprofen?

Previously no one had considered the possibility that an NSAID could block NMDA receptor activity indirectly. Likewise, no one had given much thought to the prospect of additional drug-specific mechanisms of action of an NSAID within subgroups of patients. These findings represent a new contribution in the management of the surgical patient, both pre-emptively and post-operatively.

Acceptance and validation of CEME outcomes

New teachings about Orudis were followed with an acceptance by healthcare professionals of a new classification of ketoprofen as a dual-acting agent; this classification challenged traditional dogma on peripherally- versus centrally-acting analgesics. The International Association for the Study of Pain (IASP) subsequently included key publications by myself and coworkers on the effects of NSAIDs upon spinal nociceptive processing within their Core Curriculum for Postgraduate Education. *Keith McCormack Pain 1994 59 9-43* remains a landmark publication in pain medicine.

At that time, CEME enabled ketoprofen (Orudis), but not any other NSAID to be aligned with the emerging view that postinjury hypersensitivity is mediated by both peripheral and central phenomena. Subsequently, several groups independently validated the CEME outcomes and demonstrated up-regulation of kynurenic acid by ketoprofen *in vivo.* For example…

"...ketoprofen also acts through kynurenic acid as a central antagonist on the NMDA receptor. Due to this central analgesic mechanism of ketoprofen, we expected an analgesic preemptive effect. This study was carried out following the Breivik/Stubhaug preemptive effect study design..."

"...we showed a preemptive effect with ketoprofen..." (Anaesthesist 1998 47(4):303-10).

Dissemination of the CEME outcomes resulted in immediate new business for Orudis in excess of $100M (1993).

Summary

The case study of Orudis illustrates the power of CEME in discovering a previously-unknown relationship. Validation of the novel role of Orudis in both

pre-emptive and postoperative analgesia through a novel cyclooxygenase-independent mechanism was forthcoming. In those days the era of real world evidence had yet to emerge. But, if such observational data had been available, the clinical correlate of CEME, pre-emptive and postoperative analgesia would point to precisely where value-differentiation resides.

CEME: CONSTRUCTION OF A NEW SQL DATABASE

In 2002 McCormack Pharma undertook the construction of the first structure queried language (SQL) real-time database that catalogued all known clinical applications for the existing UK and US pharmacopoeia*. The database incorporated metadata layers for searching by chemical structure/fragment** and all known human genome-derived and non-human genome-derived targets. These databases increased by orders of magnitude the capacity of CEME to locate value-differentiation of any drug.

CEME IN 2015

Today, CEME is recognized as a formidable association rule learning algorithm. Additionally, CEME employs SQL relational databases that incorporate various metadata layers. Located at a single site in the UK, CEME remains the only purpose-built tool for value-differentiation within the Pharma industry; with over three decades of outstanding achievements.

*
Note: Public access to this database by subscription via www.off-label.com was discontinued in 2011 and all rights to use are retained exclusively by McCormack Pharma

**
In collaboration with Cambridgesoft