Health Care Applications of Data Mining

Mollie R. Poynton PhD, APRN
Assistant Professor, University of Utah College of Nursing
Adjunct Assistant Professor, University of Utah School of Medicine,
Department of Biomedical Informatics
February 8, 2007
Where you could be if you weren’t here…

Standing in line at the Department of Motor Vehicles
Or maybe...

Getting a “well check” from your pediatric nurse practitioner
Even better...

Giving birth
Today’s Agenda

- Coping with Health Care Data
- Data Warehousing & Data Mining
- Applications in Health Care
- Example: pharmacokinetic and pharmacodynamic modeling of remifentanil
Large amounts of data

1st problem: storage

- Solution: storage now inexpensive
Large amounts of data

1\textsuperscript{st} problem: storage
- Solution: storage now inexpensive

2\textsuperscript{nd} problem: optimize data for analysis
- Solution: data warehousing
Accessing data from applications

- **1970s** – mainframe, not easily accessed
- **1980s** – distributed database management systems, still some difficulty pulling data
- **1990s** – advent of enterprise data warehouses (EDW), stores of data optimized for analytics
What is a Data Warehouse?

- A collection of integrated subject-oriented data extracted from the enterprise source systems
- Each unit of data is relevant to some moment in time
- A data warehouse contains atomic data and lightly summarized data
UHC Data Warehouse
Architecture

- Hybrid Inmon Architecture
- Oracle Backend
- Data stored at atomic level
- Long term data store
- Non-volatile - All updates are tracked
- Real-time clinical data
- Batch financial data
- Over 200 Contributor Systems
• Data marts are organized by subject area.
• Data marts are optimized to support reporting and analysis.
What data does the EDW contain?

- Patient and Visit Data
- Operational Data
- Clinical Data
- Financial Data
- Research Data

Which represents…
- One Terabyte of data
- Over 500+ million records
Circumstance of Data Explosion

- Inexpensive storage
- Growth of data warehousing technology
- “Data dumping”
- Data volumes = terabytes
Large amounts of data

1st problem: storage
- Solution: storage now inexpensive

2nd problem: optimize data for analysis
- Solution: data warehousing

3rd problem: extract useful information and knowledge from data stores
- Solution: ???
Why is such large, complex data problematic?

- Not readily used by consumers of data/information/knowledge
- Data quality issues
- In-memory processing often not feasible
  - Will this problem be solved soon?
- Combinatorial explosion encountered
- Traditional statistical techniques may not be sufficiently flexible to model complex patterns
Data/ Knowledge Disparity

- Molecular biology: millions of genotypes, but only thousands of associated phenotypes
- Pharmaceuticals: Enormous increase in “leads”, but stunted productivity.
- Nursing: Warehousing of data from clinical information systems – improved nursing care?
- Human medical data in North America and Europe = thousands of terabytes – how much has our clinical knowledge increased?
Large amounts of data

- **1st problem: storage**
  - Solution: storage now inexpensive

- **2nd problem: optimize data for analysis**
  - Solution: data warehousing

- **3rd problem: extract useful knowledge from data stores**
  - Solution: knowledge discovery in databases/data mining
Gaining Knowledge from Big Data

- Queries/ summaries
- OLAP (online analytic processing)
- Visualization
- Statistical methods
- Machine learning methods
The Knowledge Discovery in Databases (KDD) Process

From: Fayyad, Piatetsky-Shapiro, & Smyth, 1996
Why Data Mining/ KDD?

- Data explosion/ highly dimensional health care databases
- Growth of data warehouses
- Need for technologies that accelerate modeling of clinical knowledge.
- Exciting potential for discovering useful information and knowledge from existing data stores
Data warehousing v. Data mining

- Data warehousing creates a single point of query, a single resource for data from heterogeneous sources.
- Data mining (aka knowledge discovery in databases) extracts useful knowledge/patterns from the data (predictive models, trends, etc...)
- Data warehousing facilitates data mining.
Applications of Data Mining in Health Care

- Modeling Health Outcomes
- Modeling clinical knowledge for decision support systems
- Bioinformatics
- Pharmaceutical research
- Business Intelligence
“data mining” OR “knowledge discovery”
Application: Hospital QI/QA

- Infection Control
  - Surveillance systems developed using these techniques shown more sensitive and specific.¹
- Care Pathways/ Guidelines
  - Prompt use of clinical pathways for appropriate patients²
Application: Predicting Patient Outcomes

- Predicting smoking cessation status$^{3,4}$
- Predicting pre-term labor$^5$
- Predicting pneumonia for diagnostic decision support$^{6,7}$
- Predicting cancer prognosis$^8$
Application: Genomic Medicine

- Data-intensive area of research
- Data mining methods are hypothesis-generating
- Generates “leads” or new directions for pharmaceutical development
- Discover relationships from published literature\(^9\)
Other Applications

- Pharmacokinetic and pharmacodynamic modeling
  - Remifentanil\textsuperscript{10}

- Decision Support for Pathology/ Radiology
  - Classifying tumors/ lesions, identifying abnormalities\textsuperscript{11,12}
Special characteristics of health care data

1. Heterogeneity
2. Lack of canonical form
3. Poor mathematical characterization
4. Sensitive/ private nature of health care data
5. High dimensionality
6. Volume
Heterogeneity of Data in Health Care
Health care concepts often lack canonical form

- Concepts lack a single, preferred notation that encompasses all equivalent concepts
- Ex: chest pain, radiates to left arm
  
  CP -> L arm

  Chest pressure with radiation to left arm
  Chest pressure radiating to arm

  *(We are making great progress here with standardized vocabs, snomed, NIC/NOC/NANDA, etc.…)*
Poor mathematical characterization

- Concepts in healthcare lack formal structure (like that found in basic sciences).
- No standard representations that can be entered into formulas/models (does mild depression = 0.5 moderate depression?).
- Significant need for canonical form – standardized vocabularies, so that we can identify equivalent concepts.
Privacy Issues

- Both patient and provider!!
- Follow HIPAA guidelines...
- Theoretical possibility of re-identification
- Sensitive “predictions”
- Important to weight risks v. benefits
Population pharmacokinetic and pharmacodynamic models of remifentanil in healthy volunteers using machine learning methods
Our team

S. Kang, University of Inje, School of Health Administration
M. Poynton, Informatics Program, University of Utah College of Nursing and University of Utah School of Medicine, Department of Biomedical Informatics
G. Noh, Asan Medical Center, University of Ulsan College of Medicine, Clinical Pharmacology and Therapeutics, Anesthesiology and Pain Medicine
K. Kim, Sanggye Paik Hospital, Inje University College of Medicine, Anesthesiology and Pain Medicine
H. Lee, Mokdong Hospital, Ewha Womans University College of Medicine, Anesthesiology and Pain Medicine
D. Kim, Dankook University College of Medicine, Department of Anesthesiology and Pain Medicine
K. Bae, Asan Medical Center, University of Ulsan College of Medicine, Clinical Pharmacology and Therapeutics
O. Linares, University of Utah, Bioengineering
S. Kern, University of Utah College of Pharmacy, Pharmaceutics/Pharmaceutical Chemistry, Anesthesiology
Terms:

Pharmacokinetics (PK) v. Pharmacodynamics (PD)

- The process by which a drug is absorbed, distributed, metabolized and eliminated.
- The action or effect(s) of a drug.

PharmacokinetiCS  Pharmacodynamics
Other terms

- **EEG (electroencephalogram)**
  - measures the electrical activity of the brain/ CNS

- **CNS (central nervous system)**
  - brain and spinal cord

- **ApEN (approximate entropy)**
  - a parameter measured by electroencephalography
  - can be used as a surrogate measure of CNS activity

- **ANN (artificial neural network)**

- **NONMEM**
  - Dominant software package used for pharmacokinetic and pharmacodynamic modeling – builds mixed nonlinear effects models
remifentanil

- Purpose: anesthesia during surgical procedures (along with a muscle relaxant)
- Opioid mu-receptor agonist (similar to Fentanyl)
- CNS depressant
- Very rapid achievement of peak effect
- Very short duration of action
Background

- Some basic investigation of machine learning methods for PK and PD modeling, but dominant method = mixed nonlinear effects model (NONMEM).\textsuperscript{14,15}
- Noh data set excellent test bed for comparison
An ideal data set…

- 30 healthy volunteers infused with 1-8 $\mu$g/kg/min for 20 minutes or continuous infusion 3 $\mu$g/kg/min.
- *Unusually broad* spectrum of pharmacokinetic and pharmacodynamic parameters
- Ideal data set for examining relative performance of machine learning methods and dominant method, mixed non-linear effects model (NONMEM), for predicting individual blood concentrations
Objectives

1. Compare accuracy of ANN and NONMEM in predicting remifentanil blood concentration.\textsuperscript{10}

2. Determine whether an artificial neural network can predict overshoot of ApEn (approximate entropy) during recovery from profound remifentanil effect.\textsuperscript{10}

3. Compare performance of multiple methods in predicting remifentanil blood concentration: NONMEM, ANN, SVM, ensemble methods.\textsuperscript{16}
Background Obj. 2

- EEG effect/ CNS suppression of remifentanil thought to be the same on administration and recovery
- Noh and colleagues observed an “overshoot” of CNS depression not predicted by NONMEM.
Overshoot of EEG ApEn
EEG (electroencephalogram) Data

- Abundant! (Measurements q 10-20 seconds)
- 24,509 measurements of ApEN (approximate entropy)
- Does NONMEM require such a reduction of measurements that important information is lost?
Objectives

1. Compare accuracy of ANN and NONMEM in predicting remifentanil blood concentration.\textsuperscript{10}

2. Determine whether an artificial neural network can predict overshoot of ApEn (approximate entropy) during recovery from profound remifentanil effect.\textsuperscript{10}

3. Compare performance of multiple methods in predicting remifentanil blood concentration: NONMEM, ANN, SVM, ensemble methods.\textsuperscript{16}
We hypothesized that...

1. The ANN PK model would predict blood concentrations of remifentanil with equivalent or greater accuracy than a NONMEM PK model.

2. An ANN PD model including all instances of ApEN (n = 24,509) with corresponding predicted blood concentrations of remifentanil would predict ApEn overshoot.
Our strategy

1. Build and compare PK models using NONMEM & ANN.
2. If ANN as accurate as NONMEM, build a PD model with ANN predictions...
Table 1. Number of Instances, Input Variables and Number of Hidden Layer Neurons of Pharmacokinetic (PK) and Pharmacodynamic (PD) Models of Artificial Neural Network

<table>
<thead>
<tr>
<th></th>
<th>PK</th>
<th>PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of instances</td>
<td>1340(^\dagger)</td>
<td>1077(^\dagger)</td>
</tr>
<tr>
<td>Input variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Model A</td>
<td>Model B</td>
</tr>
<tr>
<td>Amount</td>
<td>Age</td>
<td>Age</td>
</tr>
<tr>
<td>Infusion</td>
<td>Sex</td>
<td>Sex</td>
</tr>
<tr>
<td>Rate</td>
<td>Weight</td>
<td>Weight</td>
</tr>
<tr>
<td>Age</td>
<td>Height</td>
<td>Height</td>
</tr>
<tr>
<td>Sex</td>
<td>Measured concentrations of remifentanil</td>
<td>Predicted concentrations of remifentanil</td>
</tr>
<tr>
<td>Weight</td>
<td>Height</td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*: measured concentrations of remifentanil, \(^\dagger\): measured concentrations of remifentanil versus corresponding ApEn values, \(^\ddagger\): predicted concentrations of remifentanil calculated from the pharmacokinetic model of ANN versus complete ApEn data.

Time: time elapsed from the beginning of remifentanil infusion (minutes), Amount: total dose infused (mg), Infusion: 1 = during remifentanil infusion, 0 = after termination of remifentanil infusion, Rate (mg/min), Age (yr), Sex (1 = male, 0 = female), Weight (kg), Height (cm).
## PK models

### Table 3. Predictive Performance of the Pharmacokinetic Models of NONMEM and Artificial Neural Network (ANN)

<table>
<thead>
<tr>
<th></th>
<th>NONMEM (95% CI)</th>
<th>ANN* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>95.8 (79.2-112.3)</td>
<td>57.1† (3.22-71.03)</td>
</tr>
<tr>
<td>MAE</td>
<td>5.4 (4.99-5.87)</td>
<td>4.1† (3.76-4.44)</td>
</tr>
<tr>
<td>MDAWR</td>
<td>0.27</td>
<td>0.31</td>
</tr>
<tr>
<td>$\overline{MAWR}$</td>
<td>0.37 (0.31-0.42)</td>
<td>2.22† (1.66-2.78)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.83</td>
<td>0.90</td>
</tr>
</tbody>
</table>

*: Five testing sets of five-fold cross-validation were combined to obtain predictive performance estimates for ANN.
†: $P < 0.001$ vs. NONMEM.

MSE: mean squared error, MAE: mean absolute error, MDAWR: median absolute weighted residual, $\overline{MAWR}$: mean of the individual mean absolute weighted residuals.
### Table 4. Predictive Performance of the Pharmacodynamic Models of NONMEM and Artificial Neural Network (ANN)

<table>
<thead>
<tr>
<th></th>
<th>NONMEM</th>
<th>ANN</th>
<th>Model B (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ordinary sigmoid E&lt;sub&gt;max&lt;/sub&gt; model (95% CI)</strong></td>
<td></td>
<td>Model A (95% CI)</td>
<td>Model B (95% CI)</td>
</tr>
<tr>
<td>MSE</td>
<td>0.00725 (0.00594 - 0.00731)</td>
<td>0.148 (0.004 - 0.007)</td>
<td>0.0018* (0.0017 - 0.0019)</td>
</tr>
<tr>
<td>MAE</td>
<td>0.058 (0.056 - 0.061)</td>
<td>0.052 (0.047 - 0.058)</td>
<td>0.0300* (0.0293 - 0.0307)</td>
</tr>
<tr>
<td>MAR</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.52</td>
<td>0.58</td>
<td>0.74</td>
</tr>
</tbody>
</table>

*: P < 0.01 vs. ordinary sigmoid E<sub>max</sub> model.

N = 1581 for the pharmacodynamic model of NONMEM [1].

ANN - Model A: the data set was measured concentrations of remifentanil versus corresponding ApEn values (n = 1077).

ANN - Model B: the data set was predicted concentrations of remifentanil calculated from the pharmacokinetic model using ANN versus complete ApEn data (n = 24509).

MSE: mean squared error, MAE: mean absolute error, MAR: median absolute residual.
Predicting ApEn overshoot with an ANN
Conclusions from remifentanil study

1. Reduction of pharmacodynamic data in order to make PD modeling with NONMEM computationally feasible resulted in inaccurate representation of REAL clinical response to remifentanil.

2. An ANN PD model was able to predict ApEN overshoot during recovery from profound remifentanil effect.

3. Overall, the predictive accuracy of the ANN PK model was better than that of the NONMEM PK model, but the ANN PK model tended to underpredict the lower range of measured concentrations of remifentanil.
Objectives

1. Compare performance of ANN and NONMEM in predicting remifentanil concentration, using basic pharmacokinetic parameters.$^{10}$

2. Determine whether an artificial neural network can predict overshoot of ApEn (approximate entropy).$^{10}$

3. Compare performance of multiple methods in predicting remifentanil concentration: NONMEM, ANN, SVM, ensemble methods.$^{16}$
References


4. Poynton M. *Classification of smoking cessation status with machine learning methods.* Bloomington, IN: Graduate School, Indiana University; 2005.


Informatics at University of Utah

- University of Utah College of Nursing, Nursing Informatics Program
  http://www.nurs.utah.edu/programs/informatics/index.htm

- University of Utah School of Medicine, Department of Biomedical Informatics
  http://uuhsc.utah.edu/medinfo/