RX-tut

Acquisition of Knowledge from a Database

Gio Wiederhold, Ph.D. [1]
Robert L. Blum, M.D., Ph.D. [2]
Michael Walker [3]

Departments of Medicine[1,3] and Computer Science[1,2]
Stanford University

March 1988

Presentation

1. Review of general concepts as used by us
2. Overview of RX
3. Data and knowledge processing
4. The Architecture to support RX
5. General Conclusions
6. Future Work

Objectives

1. Gain an understanding for
   interactions in a large knowledge-data system
2. Get a feeling for some of the detailed
   implementation issues
3. Learn from a working system, not phantasy

This is not an introduction to AI . . .

1. Basic Concepts

Computing for DECISION-MAKING

a global objective

Combine

Data — the state of the world —
Knowledge — our abstractions —

$\boxed{\text{Computational Engine}}$

Predictions of the Future

	Paradigm	
	Traditional	Artificial Intelligence
Knowledge:	Program	Rules, ...
Data:	Files	Ground rules, instance frames
Engine:	CPU	CPU and interpreter

Data, Knowledge, Information

Data:

1. Factual observations on
   specific objects or events
2. Measured in the past
3. Objectively verifiable

Knowledge:

1. General descriptions or abstractions on
   classes of objects or events
2. Predicting the future
3. Obtained from experts
4. Uncertain and not verifiable

Information:

1. Data or knowledge
   previously unknown to the receiver
2. Used for decision-making

Litmus test:
If an automatic process or clerk can collect the material then we are talking about *data*.

If an expert has to provide the material then we are talking about *knowledge*.

AI System Modes

Early Expert Systems

1. Data poor
2. Goal driven (user request)
3. Backward chaining
4. Often focused
5. Minimize data requests from user

Knowledge Based Systems / EDS

1. Data rich
2. Can be data-driven (triggers)
3. Forward and backward chaining
4. Easily explosive
5. Minimize repetetive data requests

A Model of Information Flow

Experience | State Recording

Education

$\boxed{\text{Knowledge}}$ $\boxed{\text{Data}}$

selection

$\boxed{\text{Knowledge}}$ $\boxed{\text{Information}}$

↑ feedback | feedback ↑

analysis

$\boxed{\text{Decision-making}}$

Scientific Results | Actions

2. Overview of RX

Objective of RX

Knowledge Extraction from Databases

Hypothesis

Databases contain much experience,
(more than any single physician
can accumulate)
This knowledge can be extracted to serve
(eventually)
knowledge-based advice-giving systems

Knowledge is used to drive the system

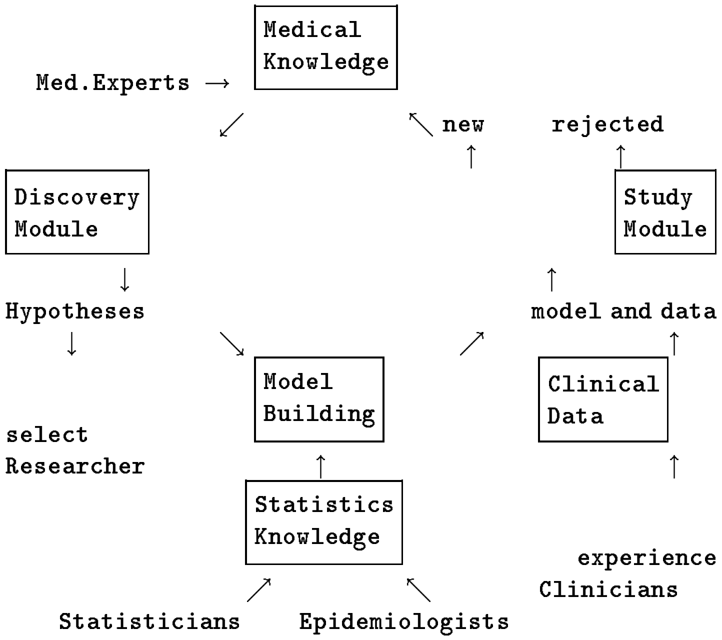
The Knowledge representation for
initial
and
derived
is identical

Components

1. Medical database:
   relational model, transposed
   extracted from clinical use
2. Medical Knowledge base: frames
   2.1 for multiple interpreters
   multi-objective
   2.2 interlinked structures
3. Statistical Knowledge: rules
4. Statistical Validation: programs

Processing Flow

Cycle: Discovery - Study - Modeling - Verification -
- Augmentation of Scientific Knowledge



Roles

Medical Knowledgebase

initial knowledge
directs inference →
→ accepts new knowledge

Medical Database

contains past experience
basis for inference →

Statistical Knowledge

processing rules

Interpreters

hand-coded engines
interpret the knowledge

Database

Aramis—American Rheumatism Association

uses TOD (began 1969)

Time-Oriented Database System

Features:

- Domain oriented data types
  1. Date
  2. Severity codes (0, +, ++, ...
  3. User defined codes (female, male, ...

     . . .
- Subsetting operations
- Transposition

Size of Stanford subset: about 30 Mbytes

Knowledge Base

More complex than the database

Interlinked structures:

- Categorical Knowledge is Hierarchically Organized
- Definitional Knowledge has distinct Hierarchies
- Causal Knowledge links across the hierarchies

Aggregate knowledge is a network structure

- represented by frames with references to each other

```
                 ALL-UNITS

   STATES         ACTIONS         STAT'M'DS
 DIAGN'C-CAT'S      DRUGS        REGRESSION
     ...             ...             ...
 CARDIAC-DIS'S    ANTIBIOTIC     MULT-REG'N
         etc              etc              etc
```

Displaying the RX Knowledge Base

Menu of Display Options

MACRO	function(args)	EXAMPLE Display
DS	schema(node)	DS Nephrotic-syndrome
DP	paths(c↔e)	DP SLE Cholesterol
DC	causes(e-node)	DC WBC
DE	effects(c-node)	DE Prednisone
DD	distribut'n(c e)	DD Prednisone Cholesterol
DM	model(c e)	DM Prednisone Cholesterol
DEV	evidence(c e)	DEV Prednisone Cholesterol
DF	frequencies	DF
D	desc'dnts-tree	D Diagnostic-Categories
CLASS	classificat'n	CLASS Azathioprine
SPEC	children	SPEC Diagnostic-categories
SIBS	siblings	SIBS Azathioprine
TR	traverse right	TR Glomerulonephritis
TL	traverse left	TL Glomerulonephritis
PL	print property list	PL Validity
PPL	print verbose pr.list	PPL Frequency

(These functions provided many of the slides below)

Hierarchical Classification of Diseases

each frame has

- a generalization slot and
- a specialization slot:

```
              respiratory diseases:
          genl: all categories of disease
        spec: pneumonia, asthma, emphysema

pneumonia            asthma               emphysema
genl: resp'ry dis.   genl: resp'ry dis.   genl: resp'ry dis.

spec:                spec:                spec:
 pneumococcal pn.     allergic asthma      pco2 retention
 klebsiella pn.       intrinsic asthma
```

- Assumptions: Completeness across
  Inheritance

Display Hierarchical Frames

Display the Descendants in the Hierarchy

```
D Autoimmune-Disorders

Autoimmune-disorders
    SLE
        Lupus-nephritis
        Cardiac-lupus
        CNS-lupus
        lupus-serositis
    Ra
    Arteritis
```

Hierarchical Classification

```
CLASS Glomerulonephritis

(Glomerulonephritis
    Renal-disorders
        Diagnostic-categories
            States)
```

Definitions

Definitions may be in Terms of other Attributes
of other Objects

It is important that medical knowledge is
available at a high level of abstraction,

but the definition may use other (lower) frames,
in another hierarchical subtree

`Pneumonia`

`definition: Temperature` $>$ `102 degrees F.`
`and WBC` $>$ `10,000 cells per` mm^3
`and Chest X-RAY = Lobar Infiltrate`

At the lowest level the frames correspond
to attributes found in the DATABASE

Causal Knowledge Linkes Nodes

i.e.: `Temperature is Affected by Pneumonia`

```
Pneumonia              Temperature
affected-by:           affected-by:
          Alcoholism                Pneumonia
            Diabetes                Influenza
effects:               effects:
         Temperature                Perspiration
                 WBC
          Chest-XRAY
```

Each causal relationship is represented as
a set of features:

`intensity, frequency, direction,` *setting, functional form, validity, evidence*

The relationship "`Pneumonia increases temperature`":

```
intensity: to 104 degrees F.
frequency: common
direction: +
setting:  studied in middle-aged patients
    with pneumococcal pneumonia
functional form: .5log (severity pneumonia) + 98
validity: widely confirmed
evidence: citations to medical literature
```

SUMMARY of Round 1

Most 'KNOWLEDGE' is in the relationships

Frames (*and people*) define their meaning
through relationships to others

In a small knowledgebase
linkages can be arbitrary

Semantic Nets

As the knowledge grows
we imposed structure

1. Categorical,
2. Definitional,
3. Causal.

To relate knowledge to the data this structure
must be applicable to data instances

in class frames
schema frames at DB level

3. Data and Knowledge Processing

Scientific process is a cycle:

Instances → Experience

Education + Experience → Knowledge

Unexpected Instances → Questions

Questions + Scientific training → Hypothesis

Hypothesis + Knowledge → Model

Model + Data → Validation

Validation + Dissemination → New Knowledge

How and Who?

Our Example: MEDICINE

Student learns 8 y

cycle starts

Clinician	treats	5 y
Clinician	observes exceptions	+ 1 y
Clinician	studies cases	+ 2 y
Clinician	formulates Hx	= 2 y
Archivist	collects data	= 2 y
Epidemiologist	formulates model	+ 3 m
Statistician	applies methods	+ 3 m
Data Analyst	selects and processes data	+ 6 m
All	write	+ 1 y
Editors	review	+ 1 y
Journal	publishes	+ 1 y
Clinicians	adapt practice	+ 3 y/cr

Many participants

Much time in cycle: 16 y net

Operational Cycle of RX

Models this scientific process

1. Collect data – done outside of RX
2. Collect and represent Medical Knowledge
   - 2.1 Define categorical frames from 'textbook' knowledge
     - densely in area of interest (≡ DB)
     - use inheritance in categorical hierarchy outside
   - 2.2 Make Definitions to link Concepts to Database
   - 2.3 Initialize known cause/effect linkages
     - in area of interest
3. Collect rules for statistical processing
   tied to the data description
4. Program control mechanisms for 5. – 10.
5. Discover unusual events
   RX: Brute force correlation
   RADIX: scan for time-variations
6. Generate hypotheses
7. Build model for hypothesis testing
8. extract data for statistical Hx test
9. run test
10. append validated HX to knowledge base
11. iterate to 5.

4. The Architecture of RX

to support the cycle

DATABASE

and

KNOWLEDGE BASE

DISCOVERY module → generate hypotheses

validate ← STUDY Module

STATISTICAL PROGRAMS

all controlled by several

KNOWLEDGE INTERPRETERS

AI paradigm

Similar to DENDRAL

GENERATE and TEST

Discovery Module → Study Module

All kinds of correlations →

independendent, significant correlations

Clinical Database

Data is a byproduct of medical practice
Cases are representative
Many uses of data:
    Health care
    Billing
    Medical Audit
    Research

ARAMIS

Relational model

2 relations
    Patients: (pat-no, DoB, ... (50 values))
    Visits: (pat-no, date-of-visit, reason ... (500 values))

Internally transposed

The ARAMIS Database

Transposed:
    by ATTRIBUTE
        by PATIENT and VISIT

Data attribute column (values p1.v1 p1.v2 . . . p2.v1 . . .)

```
Patient-Id (1 1 1 1 1 1   3 3 3 3 3 3 3 3   . . .
        6 6 6 6 6 6 6 . . .    78 78 78 . . .
Visit-date (10Mar78 11Apr78 23Jun78 1Jul78 10Jul78
        4Dec78   15May78 . . .
Cholesterol (31 29 24 30 31 29   23 - 27 25 = 23 - =
          . . .   20 22 = = = 21 . . .     32 34 . . .
Prednisone . . .
```

Columns stored as a variable length
    compressed records
    controlled by a prefix table
        0(!), 1(0), 2(-), 3(=)

Collected using Forms

The Database in RX

Transposed
    by PATIENT
        by FRAME (States, Actions)
            by VISIT → VALUE

Data attribute strings:

```
(Patient1
 (Aspirin ((1 30)(2 20)(3 20)(4 20)(5 20) ...))
 (Cholesterol ((1 215)(2 229)(4 230)(...))
 . . .
 (Prednisone ((1 50)(2 27)(4 25))
 . . .
 (Visit-date (1Jun80 15Jun80 12Jul80 ...)
 . . .
(Patient78
 (Aspirin ((6 10)( ... ))
 (Cholesterol ((1 ... ))
 . . .
. . .
```

Schema Frames

Other slots define computational parameters:
    Example: Schema for Hemoglobin

```
Hemoglobin                       explanation
----------
attribute-type:                  represented as a
              point-event                  time:value pair
value-type: real                 i.e. a real-valued number
range: 0 < value < 25            the legal range of values
units: grams per deciliter       units of measurement
significance: .1                 used for rounding off values
```

and Real World Knowledge

```
----------
function: oxygen transport
molecular-weight: 67,000 daltons
structure: Fe + heme + 4 polypeptide chains
part-of: red blood cell
affected-by: high altitudes, genetic make-up
clinical-effects: deficiency causes fatigue
    severe deficiency may cause cardiac failure
```

Discovery Module

RX uses database directly — (no knowledge used)

Generate Hypothesis of relationships

Search for binary correlations of events/concepts, time lagged

Imprecise, often false, or useless:

- HX may be known but overlooked in knowledgebase acquisition (discard Hx, update KB)
- HX may be trivial (discard)
- HX is worthy of study (to see if it seems valid)

Costly

— use subsets of data

— run on weekends

Select

— strong → rank by correlation (R-value)

— interesting

— non–obvious

Final selection of hypothesis

by manual inspection

Data for Discovery

Initially

only use a subset

50 patients

50 attributes → 50! interactions

6—50 visits

12 timelags

Future— use AI - model guidance ?

but avoid excessive restrictions

RADIX trigger from changes in the data

at a high level of abstraction (see later)

— the validation is done on the full set of data —

Important: first a patient's course is characterized

then correlations are computed over

the characterizations.

Combining Correlations Across Patients

The patient is the entity –

not the events,

the #(events/observations) differ greatly

Patient-based score

Patient 1: $r_1 = cor(x, y)$ $log[pval(r_1)]$

Patient 2: $r_2 = cor(x, y)$ $log[pval(r_2)]$

Etc.

$$score(x, y) = -2 \sum_{i \in allpatients} log[pval(r_i)]$$

$$score(x, y) \approx \chi^2 2p$$

Output from Discovery Module

Possible Causal Effects of Prednisone

variable	lag strength
Hemoglobin	(B + 518)
Anti-DNA-Hemagglut	(B - 514)
Disease-Activity	(R + 469)
C3	(B + 389)
Fatigue	(R + 370)
Urine-WBCS	(R + 350)
Albumin	(R - 346)
BP-Diastolic	(C + 322)
WBC	(C + 306)
Urine-RBCS	(B -293)
Temperature	(B - 275)
Weight	(C + 269)
LDH	(C + 268)
Glucose	(C + 256)
Log-Fana	(C - 238)
Lymphs	(C - 194)
BP-Systolic	(C + 167)
...	...

Study Module

1. Use knowledge to build
   model for statistical analysis
   1.1 look at confounders and
   1.2 their temporal relationships
2. Use data estimates to select
   statistical procedures
   2.1 use rules
   2.2 use meta data
   cardinality
   type information

(1. and 2. are interdependent, but
iteration is not now automated)

3. Extract required data
   from database
4. Perform analysis
5. Inspect result
   if significant - insert into knowledge base

> Another study will now take
> the new knowledge into account

Next: build statistical models

We have to locate all (known) confounders

```
GLOMERULONEPHRITIS as a confounding
  variable for Prednisone and Cholesterol:
GLOMERULONEPHRITIS (30 pct activity) increases
 NEPHROTIC-SYNDROME (3 gms proteinuria/24 hrs)
  is treated by PREDNISONE (604 % of baseline)
GLOMERULONEPHRITIS (30 pct act'y) is treated by
 PREDNISONE (182 % of baseline)
GLOMERULONEPHRITIS (30 pct act'y) increases
 NEPHROTIC-SYNDROME (3 gms ...) increases
  CHOLESTEROL (120 mgms/dl)
GLOMERULONEPHRITIS (30 pct act'y) is treated by
 PREDNISONE (182 % of baseline) attenuates
 NEPHROTIC-SYNDROME (-1 gms ...) decreases
 CHOLESTEROL (-22 mgms/dl)
GLOMERULONEPHRITIS (30 pct act'y) is treated by
 PREDNISONE (182 % of baseline) increases
  CHOLESTEROL (11 mgms/dl)          new
GLOMERULONEPHRITIS (30 pct act'y) is treated by
 PREDNISONE (182 % of baseline) attenuates
  SLE (-6 pct activity) attenuates
   NEPHROTIC-SYNDROME (0 gms ... ) decreases
    CHOLESTEROL (-5 mgms/dl)
```

Displaying the Schema

The critical information for a disease

DS NEPHROTIC-SYNDROME

NEPHROTIC-SYNDROME

Classification:

(NEPHROTIC-SYNDROME RENAL-DISORDERS
DIAGNOSTIC-CATEGORIES STATES ALL-UNITS)

```
Nephrotic-syndromeis defined as follows:
 or
      during HEAVY-PROTEINURIA
   and
           during HYPOALBUMINEMIA
           during PROTEINURIA
Units of measurement: gms proteinuria/24 hrs
Default values for nephrotic-syndrome's
      time parameters:
           ||minimum-interval||
           ||---- = 30 days --||
||onset-delay||             ||--carry-over--||
||- = 7 days -||              ||-- = 30 days --||
```

A patient index based on nephrotic-syndrome is available.

The Category Hierarchy is Complete

Necessary for the CLOSED-WORLD assumption
made by its interpreter

Example: The Specialization at a top level

```
SPEC Diagnostic-categories
(Arthritic-disorders Autoimmune-disorders
Cardiac-dis'rs Dermatologic-dis'rs
Electrolytic-dis'rs Endocrine-dis'rs
Gi-dis'rs Gynecologic-dis'rs Hematologic-
dis'rs Hepatic-dis'rs Hypertensive-dis'rs
Immunologic-dis'rs Infectious-dis'rs
Metabolic-dis'rs Neurologic-dis'rs
Non-specific-dis'rs Nutritional-dis'rs
Oncologic-dis'rs Ophthalmologic-dis'rs
Psychiatric-dis'rs Pulmonary-dis'rs
Renal-dis'rs Urologic-dis'rs Vascular-dis'rs)
```

SIBLINGS

```
SIBS AZATHIOPRINE
(CHLORAMBUCIL CYCLOPHOSPHAMIDE)
```

At low levels made feasible through inheritance

Complete Property List

PL NEPHROTIC-SYNDROME

```
GENL: RENAL-DISORDERS
SPEC: (PROTEINURIA HEAVY-PROTEINURIA)
DEFINITION: (OR (DURING & --) (AND & --))
TYPE: INTERVAL
EFFECTS: (URINE-PROTEIN-RANGE ALBUMIN
          24-HR-URINE-PROTEIN --)
MINIMUM-DURATION: 30
MINIMUM-POINTS: 2
INTERVALFN: MEAN-DURING-INTERVAL
VALUE-TYPE: BINARY
INTRA-EPISODE-GAP: 100
INTER-EPISODE-GAP: 180
RECORDS: INVERTED
AFFECTED-BY: ((PREDNISONE &) (GLOMERULO-
          NEPHRITIS &) (SLE &))
PARTITION: (0 .5 1 --)
UNITS: "gms proteinuria/24 hrs"
PROXIES: (ALBUMIN 24-HR-URINE-PROTEIN
          URINE-PROTEIN-RANGE)
ONSET-DELAY: 7
MINIMUM-INTERVAL: 30
CARRY-OVER: 30
```

Definitions over Time

Our observations are actually over time
This has important effects:

1. The DEFINITIONS combine
   EVENT observations
   as recorded in the database into
   INTERVAL information

1.1 INTERVALS have parameters as
   `MAX, MIN, AVE, RATE, . . .`

1.2 Patients differ in the number of
   EVENTS observed for a disease course
   but a Course should be one interval
   a treatment should be one interval
   (*same for other time-based data – most data in planning extrapolate from past series to future*)

Definitions and Missing Data

2. We elevate detailed observations to
   higher level concepts and
   do $\boxed{\text{Statistics on Concepts, not on Facts}}$?

2.1 Our experts, and the knowledgebase
   deals better with higher level concepts

2.2 We can combine multiple event-types
   to substantiate an interval concept
   (more credibility
   in the face of missing data)
   `NEPHROTIC SYNDROME during`
   `HEAVY-PROTENURIA or`
   `PROTENURIA and ...`

2.3 We can acount for masked symptoms
   `... during SYMPTOM or`
   `DRUG` *given for that symptom*.

Why ignore the Facts?

Whats wrong with data:

1. variable number of observations
2. taken at unpredictable intervals
3. often incomplete

Use higher level concepts
defined by frames
to aggregate incomplete facts
into meaningful concepts: *Labeling*

1. Intial finding + continuing treatment
   = continuing disease state
   (treatment can mask findings,
   comtinued test for findings are costly)

2. Findings of events over time $\rightarrow$ interval
   = worsening/steady/improving
   disease state
   (matters more than level of state)

Use of This Information

MODEL BUILDING

There can be many Paths between
two nodes in our network
even at the higher, CONCEPT level

New knowledge $\rightarrow$ New direct causal path
with parameters

But, any alternative path can
also explain a hypotheses

If $\sum$ of alternate paths explains
all of the relationship
no new knowledge!

Hypothesis is invalidated

So:

1. look for all paths –
   intermediate nodes are covariates
2. prune subsumed paths
3. omit infrequent covariates
   to simplify model
   (omitting frequent covariates –
   too much loss of data)

Cycles

Note there can be cycles,
but the time delay
imposes an ordering:

Example

	intensity –delay	
Sedentary Life	—+2 —months →	Diet
Diet	—+2 —months →	Cholesterol
Cholesterol	—+2 —years →	Coronary Art.sc.
Coronary Art.sc.	—+4 —months →	Heart Attack
Heart Attack	—-2 —days →	A-type behavior
Heart Attack	—-1 —hours →	Smoking
Heart Attack	—+3 —minutes →	Sedentary Life
Heart Attack	—+5 —minutes →	Death
Heart Attack	—+2 —days →	Death
Coronary Spasms	—+3 —months →	Heart Attack
Smoking	—+1 —months →	Coronary Spasms
Hypertension	—+4 —years →	Coronary Art.sc.
Hypertension	—+3 —months →	Coronary Spasms
A-type behavior	—+1 —years →	Hypertension
A-type behavior	—+1 —varied →	Coronary Spasms
Age	—+1 —years →	Cholesterol
Age	—+2 —years →	Hypertension

There are positive and negative paths and loops
Cannot be captured by a simple logical model

Modal Effects of Prednisone

Frequency and strength of causal relationship

```
DE PREDNISONE MODE    /* one link away */
PREDNISONE, at a level of 30 mgms/day,
usually increases CHOLESTEROL
          by 50 to 130 mgms/dl,
regularly attenuates NEPHROTIC-SYNDROME
          by 1.0 to 2.0 gms prot/24 hrs,
regularly attenuates GLOMERULONEPHRITIS
          by 10.0 to 30.0 percent,
commonly attenuates SLE
          by 10.0 to 30.0 percent activity,
regularly decreases ANTI-DNA-HEMAGGLUT
          by 50 to 90 percent,
regularly increases IMMUNOSUPPRESSION
          by 16 to 32 percent activity,
regularly decreases EOS by 2 to 3 % of WBC,
occasionally increases KETOACIDOSIS
          by 20 to 100 mgms/dl of glucose,
```

* Note that all the terms are represented by
numerically encoded values

Display all Paths above threshold

to collect significant covariates:

DP SLE CHOLESTEROL (default > 0.1)

```
SLE {30 percent activity} increases
   NEPHROTIC-SYNDROME
          {1 gms proteinuria/24 hrs } increases
      CHOLESTEROL {24 mgms/dl}

SLE {30 percent activity} is treated by
   PREDNISONE {182 % of baseline} increases
      CHOLESTEROL {14 mgms/dl}

SLE {30 percent activity} increases
   NEPHROTIC-SYNDROME
      {1 gms proteinuria/24 hrs } is treated by
      PREDNISONE {143 % of baseline} increases
          CHOLESTEROL {8 mgms/dl}

SLE {30 percent activity} increases
   IMMUNOSUPPRESSION
          {18 percent activity} increases
      HEPATITIS {5 Iu/ml of SGOT} increases
          CHOLESTEROL {6 mgms/dl}
```

Displaying the Causes of Cholesterol

i.e., the Set of Nodes that Affect it

`DC CHOLESTEROL` */* other direction */*

```
CHOLESTEROL

always is increased by PREDNISONE
regularly is increased by HEPATITIS
regularly is increased by KETOACIDOSIS
usually is increased by NEPHROTIC-SYNDROME
```

Interpretation of the Frequencies

Is not linear over the range of terms:

DF	Cell	Adverb	Probability
	1	never*	.001
	2	very-rarely	.005
	3	rarely	.01
	4	infrequently	.04
	5	occasionally	.16
	6	commonly	.32
	7	regularly	.64
	8	usually	.95
	9	almost-always	.99
	10	always	1.00

* well hardly ever

Causal Inference

Generating New Knowledge in RX

Means

Establishing and Quantifying

New Causal Linkages

Correlations discovered do not establish

1. causality
2. directness

ad 1. causality: A causes B ?

heuristic if B consistently follows in time A,

then B does not cause A

(there may be an unknown covariate C,

causing both with different delays)

ad 2. directness: the correlation may be due to

known covariates – check the model

as shown previously

Use of the Covariate Model

Rule driven, but

uses medical knowledge in frames

uses metadata in frames

is controlled by a frame hierarchy

deterministic execution

1. Select proper statistical method:
   Use info about data from Schema Frames
2. Check if enough data is available:
   Ask DBMS portion for cardinality
   of subsets needed
   which distinguish
   the remaining covariates

Running the Study Module

Statistical knowledge is encoded as RULES.

The statistical knowledge in RX is not deep:

(no derivation from Probability theory)

```
Selecting instance of the class: STUDY-DESIGNS
The candidate selected is: LONGITUDANAL-DESIGN
Would you like to see rules that determined
    selection of study design?
**YES
LONGITUDINAL-DESIGN
PREREQUISITES: Can the EFFECT occur
    more than once in a patient's record?
Do we have patient records in which values for
    the EFFECT have occurred more than once
CROSS-SECTIONAL-DESIGN
PREREQUISITES: If the dependent variable is
    not a function of time, then use
    the CROSS-SECTIONAL-DESIGN
CROSS-SECTIONAL-DESIGN will also be used when
    most patient records have only a few values
```

Now select statistical procedure

The rule categorization is also hierarchical

```
Selecting instance of class: STATISTICAL-METHODS
 Considering instance: CONTINGENCY-TABLES
 Considering instance: T-TEST
 Considering instance: ANOVA
 Considering instance: REGRESSION
 Selecting instance of class: REGRESSION
  Considering instance: MULTIPLE-REGRESSION
  Considering instance: SPEARMAN-RHO
  Considering instance: KENDALL-TAU
  Considering instance: PEARSON-R
 Candidates whose prerequisites are satisfied:
  (MULTIPLE-REGRESSION SPEARMAN-
   RHO KENDALL-TAU PEARSON-R)
```

Conflict resolution rules are used to decide among these

```
The candidate selected is: MULTIPLE-REGRESSION
 Considering the instance: DISCRIMINANT-ANALYSIS
 Considering the instance: FACTOR-ANALYSIS
 Considering the instance: LIFE-TABLES
Candidates whose prerequisites are satisfied:
(MULTIPLE-REGRESSION)
The candidate selected is: MULTIPLE-REGRESSION
```

Explanation

```
The candidate selected is: MULTIPLE-REGRESSION
Wou ld you like to see decision criteria for
    selecting statistical methods?
**YES
MULTIPLE-REGRESSION
RULES:
  If the independent variables are causally ordered,
    then do a hierarchical regression.
    otherwise, do a standard regression.
PREREQUISITES:
 Multiple regression is appropriate when the number
    of independent variables is greater than 1
    All variables must be at least of
      measurement level = binary.
    All variables must be normally distributed.
Statistical method: MULTIPLE-REGRESSION
```

More explanation

```
The # of values recorded for the dependent var.
    for each patient must be > 1 +
        the # of independent variables
Next, there is the same minimum required
   # of values for the independent variable
    of primary interest
To estimate the effect of the independent variable
    for a single patient, the coefficient of
    variation must be > threshold = 10 percent
Finally, to do individual estimation, the
    total number of events must be > 1+
        # of indep. vars:
    the costliest criterion computationally
```

The Rules are in LISP:

```
Would you like to see the machine readable
            eligibility criteria?
**YES
Eligibility criteria:
[AND (IGEQ (#VALUES (QUOTE CHOLESTEROL)
               PAT)
            (ADD1 (FLENGTH VARS)))
    (IGEQ (#VALUES (QUOTE PREDNISONE)
               PAT)
            (ADD1 (FLENGTH VARS)))
        (GREATERP (COEF-VAR (QUOTE PREDNISONE)
               PAT)
                      .1)
        (IGEQ (FLENGTH (ENTRIES (QUOTE PRED-CHOL)
                      NIL PAT))
        (ADD1 (FLENGTH VARS]
```

Running the Analysis

1. Obtain the needed from the database
   attached already to the schema class frames
2a. Run through IDS (XEROX LISP package)
2b. Set up a run for a batch package
   BMDP, SPSS
   and extract results from output (*a pain*)
3. inpect the resulting t-value for
   statistical significance

If high,

Place into Knowledge base a new causal link !

Validity 4 or 5 /10

10 indisputable mechanism
8 wide experimental confirmation
6 confirmed in multiple
retrospective studies
4 confirmed by single retrospective study
2 case citations
1 based on astrology

Display a Causal Relationship

initialized or

produced by an RX analysis

```
DM PREDNISONE CHOLESTEROL

Statistical Model for Prednisone/Cholesterol effect:

cholesterol =
        - 59.17575 albumin
        + 23.13479 log(prednisone)
        + 188.1984
        + error term

Setting of Effect:
and
    not during KETOACIDOSIS      omitted
    not during HEPATITIS         covariates
```

More from the study managed by RX :

Display Evidence

```
DEV PREDNISONE CHOLESTEROL
Validity of and Evidence in Support of the Causal
 Relationship: PREDNISONE influences CHOLESTEROL.
Study Design: Longitudinal Design
Performed on:  5-May-81
Database: ARAMIS/Stanford-Immunology
Validity: 6 on a scale from 1 to 10
Interpretation of Validity:
     strong correlation and time precedence:
     known covariates controlled
Name of Study: PC
Total Number of Patients: 21
Median Number of Visits with Complete Data: 7.0
Range (Visits with Complete Data): 5.0 to 30.0
p-values of variables in the model:
  PREVIOUS-CHOLESTEROL 0.4603615
  ALBUMIN              7.450581E-9
  PREDNISONE           1.564622E-7
```

A detailed synopsis of this study is available
in the EVIDENCE file.

When knowledge is entered from other sources
this command will list citations to the literature

Distribution Across Patients

to visually check normality - (improve!) .
prerequisite for many Statistical Methods

```
DD PREDNISONE CHOLESTEROL
```

tt

Distribution across patients of CHOLESTEROL
in units of mgms/dl,
given a baseline value of 230 mgms/dl and,
given a change in PREDNISONE from 0 to 30 mgms/day,
using deciles

Range of CHOLESTEROL	Percentage of Patients	Magnitude of Change
100 150	0	extreme -
150 195	0	strong -
195 210	0	moderate -
210 225	0	weak -
225 230	0	equivocal -
230 235	0	equivocal +
235 250	0	weak +
250 280	10	moderate +
280 360	82	strong +
360 700	8	extreme +

This presents a rough representation of the function.
It provides hints on normality versus
bi-modal versus
uniform distributions.

RX Summary

The RX structure presents a
significant attempt to encode DEEP knowledge
Such knowledge is distinguished from
the SHALLOW operational knowledge in
rule-based Expert systems.

If rules are adequate they are easier to collect and use.
Frames allow a degree of structuring which is
helpful in organizing larger bodies of knowledge

- Categorical Knowledge satisfies Closed World assumption
- Definitional Knowledge understand Time
- Causal Knowledge represents the critical relationships

Aggregate knowledge is a complex network

Information Flow

Data is used in three phases

1. Initial experience for Medical Experts
   browsing
   DISCOVERY $\rightarrow$ tenative knowledge
   hypotheses = proposed goals
2. Estimating if the hypotheses can be tested
   Model building
3. Statistical processing
   validation
   new KNOWLEDGE

Future

$\boxed{\text{RX / RADIX}}$

Expand model to cover
all knowledge related to facts available from the database

Apply the RX concepts to other medical specialities

Integrate multiple models

Long Range

Use machine-processable knowledge in medicine?

Use models to share scientific knowledge

to replace papers?

to aid the review process

Transfer to other areas

Prerequisites

Need to learn from data

Reliable and deep data bases

Other Application areas

Economic models
(conditions for up/down swings)

Personnel Management
(conditions for good/poor performance,
re-enlistment)

Monitoring of Production lines
(conditions for product failures)

5. General Conclusions

drawn from RX

- Hypothesis Generation
  traditionally a scientists' function
  Inspiration derived from experience
  (Database contains more experience
  than any physician or expert!)
- Hypothesis Validatation
  (equivalent to a very smart query)
  AI provides a powerful control tool
  processing driven by knowledge
- Inserting Validated Hypothesis
  back into knowledge base
  $\boxed{\text{Learning}}$ from Data!
- Knowledge needs Structure
  for interpretation
  for sharing
  for maintenance

AI as the Control Mechanism

AI deals with knowledge
about general object types

Databases deal with data,
facts about specific objects

Use AI as the control mechanism
for large programs

A large program has

1. Computational sections
2. Data organizing sections
3. Controls to decide
   what sections to invoke
   in sequence
   iteratively
   and set parameters

Control using AI technology

Control is based on general concepts

1 Iteration has converged
2 Goal has been reached
3 Computation got stuck
4 . . .

Numeric programs use FORTRAN
IF statements for control

Statistical packages use
control 'cards' for control

Data-processing programs use
COBOL conditions for control

Real-time programs use
interrupts for control

We believe that
CONTROL BY AI IS BETTER

Advantages of AI

Control statements are based
on a variety of conditions

if these conditions are only
testable sequentially
(a fixed decision tree)
then all conditions have to be known
or the decision tree gets very complex

Rules in AI systems do not depend on
evaluation order for correctness.

HERE:
AI to serve applications,
not for its own sake.

Knowledge Base

$\boxed{\text{Literature/Education}} + \boxed{\text{Data}} + \boxed{\text{Experience}}$

$\downarrow$

$\boxed{\text{Knowledge}}$

$\downarrow$

$\boxed{\text{Consultation}} + \boxed{\text{Surveillance}} +$
$\boxed{\text{Teaching}} + \boxed{\text{Research}}$

---

6. Future Research Problems

Some
addressed by Current Research in KBMS

Long Range Research Focus

• Develop and Validate Concepts
for Information systems that store

1. factual data as well as
2. knowledge about the data

and support inferencing over the totality

Applications in
Decision support
Planning
Design support

Current State

• Little Internal Structure
meta knowledge, control rules, facts ...

• Lack of Generality / Sharability
tight linkage to interpreter

OK while sizes are modest

Problems foreseen

Performance $\downarrow$ === Size $\uparrow$

Maintenance cost $\uparrow$ === Coverage $\uparrow$

Consistency $\downarrow$ === Breadth $\uparrow$

Sharability of Knowledge 0 === Amortization of Cost 0

Specifically
impossible to maintain by multiple experts

single objective –
lack of knowledge reuse drives up cost

Questions:

How large are large KB's really?
and how will they grow?

several dozen rules – interesting
several hundred rules – large
several thousand rules – rare, major

What Is Proportion of

- Ground Rules or Facts . . . ↑
- Application knowledge . . . . ↗
- Control knowledge . . . . →

what can we expect in the future

Can KB's serve
multiple objectives?

System	size
MYCIN	100's of rules, one interpreter
XCON	large, >10,000 small OPS rules
XSEL	also large, and large overlap but distinct
R1ME	towards algorithmic approach
RX	few hundred frames, 30MB data, several interpreters
RADIX	generate and test interpreters, new KB

Knowledge Sharing

- Difficult

RADIX did NOT use
RX knowledge as represented in RX

XCON and XSEL (although similar) do not share
the knowledge representation

The Interpreter and the Knowledge

- are interlinked
- must be interlinked

so that sharing is infeasible?

Databases are a means for
SHARING data for diverse tasks

Why?
Consistency
Cost reduction

Adopt Similar Paradigm for Knowledge

Partitioning

For maintenance and control
not to limit inference

Horizontal: Knowledge / Database

Vertical: Domain 1, Domain 2, Domain n

Objects (Frames) can be in several domains

Faculty → Smith ← Consultant

Domain Characteristics

Structure		interpretation
is-a		inheritance
part-of		synchrony, ownership
derivation	boolean	dependency
	rule	pickup locally
	proc.	exec. embedded proc
completeness		negation
		univ.quantification
uniqueness		first result
disjointness		sum = total

Restrict an interpreter to
operate on knowledge
with identical characteristcs
during one [sub-] task

Planned Organization

KSYS

Defined Frames	→	object types
Instance Frames	→	selected data objects
Slots	→	Attributes
associated with		Subset: SoD
SoD	→	Features
Feature	→	Interpretable Description
Description	→	defines type of SoD
SoD	→	Subset of discourse
Active SoD	→	Scope of Interpreter
	best:	Hierarchy
instantiated via		DBMS View

Work in Progress

Modeling of SoD Interaction by theories

Definition of features of SoD

Definition of operations within SoD

Instantiation of values within SoD from DBMS

Expectations

Conceptual Linkage of KNOWLEDGE and DATA
will be more effective than
physical access linkage of
EXPERT SYSTEMS and DBMS

Fundamental CS Issues

- Representation of Knowledge
- Management of Knowledge
- Exploitation of Knowledge
- Representation of Heuristics
- Use of AI as a Control Mechanism
- Multi-step Planning
- Use of Database for Planning

also

- Dealing with Uncertain data
- Data that are distributed on
  Autonomous systems
- Dealing with Data that do
  not match syntactically
  do match semantically
- Non-monotonic updates
  that imply (eventual) model changes