EFデータ抽出ツール

東京大学
堀口　裕正

DPCデータを分析に使うには

- HISによるシステム化
- DPC分析ソフト(サービス)の利用
  - ARROWS/EVE/ヒラソル・・・・・

高
有償
低い

- 自分でDB化して分析

無料

研究班がセミナーで支援

医療機関自らDBを作って分析する

- メリット
  - コストがかからない
  - 分析に必要な考え方が身につく
  - 1度作ってしまえば、同じデータを毎回出すのには手間がかからない
- とはいっても
  - 分析用データセットを作るのには手間と時間がかかる

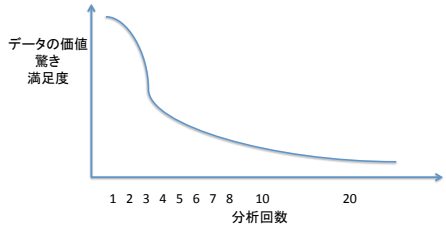
例えば・・・

- ACCESSでのデータ分析
- 毎月のDataのインポートにかかる手間
- 分析したいものに応じて作る中間テーブルの設計
- データを修正したときの差し替えの手間

- 毎月決まったものを出すOR自分で高度な分析をしたい(しできるスキルがある)のでないと継続しない？？

分析の結果ほしいもの

- 毎月継続してほしいもの
- 1回か2回データを出せば十分なもの



結果として

- ちょっと1回データの分析結果を見てみたい
- あまり手間をかけたくない

といった希望が出てくる。そうすると

- DB化は面倒、DPCのデータファイルをそのまま使って分析できないかなあ・・・。

ここからは研究班側の話

- そもそも研究遂行のために必要なので、情報システム＋分析システムを構築
- 自分たちが使いやすいよう、あるいはより多くの成果が出るように改良やノウハウの蓄積を進めてきた。

研究班での分析システムの歴史

- 初期　統計ソフト（SPSS/STATA)＋ACCESS

これでは処理しきれなくなって

- 中期　SQLサーバー＋統計ソフト＋BIソフト
  - BIソフト（QlikView/Spotfire・・・）

ここで得たノウハウやData Management手法をセミナーで医療機関の皆様に還元

近年の研究班におけるIT的課題1

- SQLサーバーでは量をこなせない？？
  - 2009年　Fファイル9億行
  - 2010年　E/F統合ファイル　16億行
  - 7年分で約1000万退院分のデータ
  - Fファイル換算約50億行分

処理に時間がかかる、たいへん・・・。今年もまた16億行分のデータ・・・。非常にうれしいけど・・・。

近年の研究班におけるIT的課題2

- 論文になったり、公表したりした成果を作るために利用した加工データの加工アイディアやロジックがそのあとお蔵入り・・・。
- （ありがたいことですが）年間数十本の論文成果、このデータの作成ロジック自体も世の中に貢献できないだろうか

課題1を解決するために

- 今「はやり」の分散処理システムを使うことにしました。
- Hadoop/Pig
- たくさんのコンピュータを並列で使ってデータが増えても機材の台数を増やすだけで解決。
- 処理をするためのスクリプト言語は統計ソフトのスクリプト並み

これはいいということで導入
(詳しい説明は後述)

分散処理システム

- これからはBigDataの時代・・
- スケールアップよりスケールアウト
- SQLなんて古い、これからはNOSQLの時代

- なんてもてはやされています
- なんといってもOSS（オープンソースソフトウエア）なので無償で使える

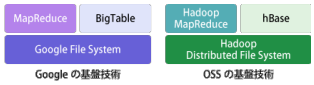
すばらしい、やはり時代はこちらに・・・
（By　東京のIT屋さん）

もちろん

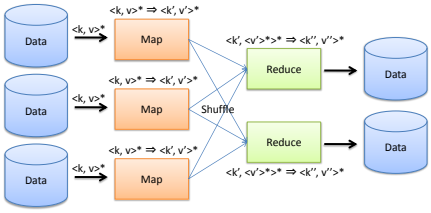
- SQLサーバーを使用し続けながらデータ量をこなす問題を解決する手法もあります。
  - SSD＋高速RAIDを使う
  - データベースを分割する
  - 機械の性能を上げる
- こちらの手法も研究班内では利用していままでのSQLで培った資産を生かす試みもやっています。
- （ただ、この手法は大量データを処理するためのノウハウなので、そこまでデータ量のない医療機関の皆様に還元できる成果が少ないです）

hadoop とは？

- Googleの基盤ソフトウェアのクローン
  - Google File System, MapReduce
- Yahoo Research の Doug Cutting氏が開発
  - 元々はLuceneのサブプロジェクト
  - Apache Project
- Javaで記述!



MapReduceの実行フロー



Pig

- Hadoopを扱いやすくするミドルウェアの1つ
- Pig Latinという専用言語で簡単にDWH的な処理を書けるようにしたミドルウェア
  - Javaを使用せずに、SQL的な言語でMapReduce処理を記述する事が出来る
  - Googleでは、同様にSawzallというスクリプト言語で、MapReduce処理が簡単に書けるようになっている
- Yahoo!が開発
- データのロード・結合やフィルタ処理を楽に書ける

システム開発の前提条件

- 利用者は研究者
  - 分析は自分の手慣れたソフトで試行錯誤を繰り返しながら使いたい
  - SAS/SPSS/STATA/R等の統計ソフトを使う＋シンタックスぐらいはかける（理解できる）
- 結果データは1患者1レコードのCSVデータとして作成、それ以上の加工はしない
- Pigというスクリプト言語を利用し、ほしいデータを抽出するスクリプトは利用者が自分で作成する（UIを作らない）

開発の概要

- 今回の抽出・加工を行うためのUDF（スクリプト言語で使う関数）を開発し、それを利用した作業が行えるようにする
- UDFで実現する内容
  - 抽出・加工内容を定義した定義ファイルがわかりやすくかける
  - 定義ファイルを使って必要なレコードを抽出する
  - 定義ファイルを使ってクロス表を作成する
  - （おまけ）日付データの取り扱いを楽にする

GroupFilterFormat（抽出・集計用定義）

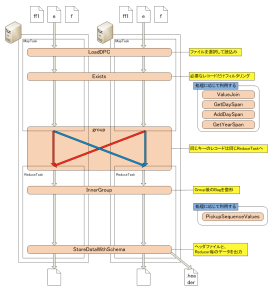
- フォーマット
  - <グループ名>(<項目名>:[<付加値>, …], …), …
- サンプル

  ボスミン(642450005[1,注],
  642450164[2,注],
  620517902[1 ,注]),
  ワルファリン(613330003[1,錠],
  621938101[0.2,細粒])

作成した関数

関数	定義
Exists	指定した文字が含まれるか判定する。
ValueJoin	指定した付加値を Bag として追加する。
InnerGroup	Bag を Group してタプルにする。
GetDaySpan	2つの日付から経過日数を算出する。
GetYearSpan	2つの日付から経過年数を算出する。
AddDaySpan	日付に値を加算して日付を得る。
PickupSequenceValues	Bag 中で値が連続しているタプルだけ取り出す。

処理イメージ



処理速度

- 本研究室内環境で2009年DPCデータから20種類程度の薬剤・処置の実施の有無を1患者1レコード型に書き出す加工
- 様式1 約300万・ Fファイル 約9．5億行

台数	処理時間
2	6900sec
4	3650sec
8	2080sec
16	1010sec
48	380sec

本システムの今後

- このシステムは処理がDPCデータに特化していないため、多様な応用が可能
  - オーダリングのデータ
  - 電子レセプトデータ
  - Webのアクセスログ・・・
- スクリプトの記載が簡単であることを確保しながら、関数を増やしていく。
- 同じスクリプトを医療機関で簡便に（パソコン1台で）利用できるプログラムも開発中

新しい分析ソフトのご紹介

- 研究班で分散処理技術の開発を行ってきた副産物としてできたソフト
- パソコン1台で実行可能
- DPCデータと処理用のスクリプトを入れればその場で結果ファイルをはき出す。

DPCデータ抽出・整形プログラム（通称 Durok）

DPCデータ抽出・整形プログラム

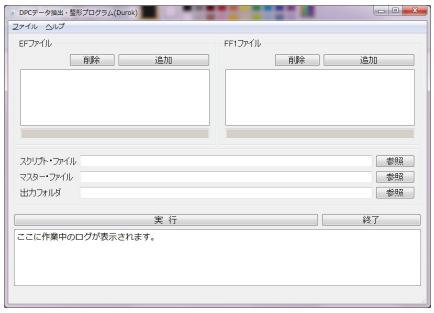
- 医療機関の「DB化は面倒、DPCのデータファイルをそのまま使って簡単に分析できないかなあ・・・」を実現/もちろんOSSなので無料（ただし無保証）
  - （注）なお簡単かどうかについては感じ方に個人差があります。
- 研究班側の「このデータの作成ロジック自体も世の中に貢献できないだろうか」についても実現可能
  - 但し以前のデータは複雑なSQL分を駆使して作成されていますので、今後の分析分からとなります

ソフトウェアの使用方法

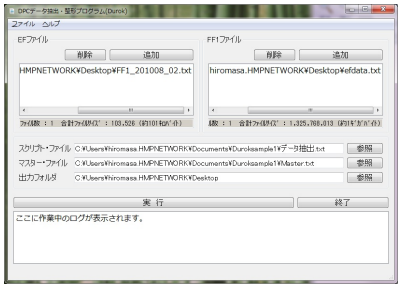
- もしかするとここからが本題
- ソフトの入手方法
  - 東京大学医療経営政策学のホームページから
  - 昨年度研究班の報告書DVDの中から
  - 東大のページには今から説明するサンプルスクリプトも一緒に置いてあります。
  - このスクリプトは随時追加予定

http://plaza.umin.ac.jp/hmp/

ソフトを立ち上げると



分析をするには



こんな風に5箇所にファイルを設定して実行ボタンを押します。

サンプルスクリプト1

- 統合EFファイル抽出スクリプト
  - 医療資源をもっとも投入した傷病名のICD10がooである患者の統合EFファイルのデータだけを抜き出したファイルを作るスクリプト
- 「このスクリプトは少数の患者さんのデータだけになっていれば自分でも（EXCELか何かで）E/Fファイルの分析ができるのに」と思っている方に、特定の患者のEFデータのみを抜き出してファイルを作成するものです。

Master1ファイルの中身

このマスターに抽出対象ICDコードを1行1こづつ記載します。

E100
E101
E102
E103
E104
E105
E106
E107
E108
E109
E110
E111
E112
E113
E114
E115
E116
E117
E118
E119

- ファイル内のICD10リストを抽出してほしいICD10のリストに書き換えます。

データ抽出スクリプトの中身

```
MASTER = load 'master.txt';
A = JOIN FF1 by $32,MASTER by $0;
A1 = Foreach A generate $0,$3,$9,$2;
A2 = FILTER A1 by $3==0;
B = JOIN A2 by ($0,$1,$2),EF by
($0,$1,$3);
FINALDATA =FOREACH B generate
$4,$5,$6,$7,$8,$9,$10,
$11,$12,$13,$14,$15,$16,$17,$18,$19,
$20,$21,$22,$23,$24,$25,$26,$27,$28,
$29,$30,$31,$32,$33,$34;
```

- こちらは特段いじる必要はありません。

結果ファイル

- 出力フォルダーにoutput.txtというファイルができます。そこに抽出された患者の統合EFファイルが書き出されます。（フォーマットはそのままです）
- あとはお好みでEXCELでもACCESSでもでファイルを開いて分析を行ってください。

サンプルスクリプト2

- 標準化死亡比計算スクリプト
  - 国立病院機構が平成22年度医療の質評価・公表推進事業における臨床評価指標において公表(http://www.hosp.go.jp/7,9502.html)した標準化死亡比の計算方法に基づいて、自分のデータで標準化死亡比を計算するスクリプトです。

使い方

- Durokにて統合EFファイル・様式1ファイル・本スクリプト及びマスターを選択し、実行します。
- 出力フォルダーのoutput.txtというファイルにデータが書き出されます。

出力データ

左から順に
- 医療機関番号
- 退院患者数
- 死亡退院患者数
- 予測死亡数
- （実）死亡退院率
- 予測死亡退院率
- 標準化死亡比
- 標準化死亡比の95%信頼区間上限
- 標準化死亡比の95%信頼区間下限

の順に記載されています。

このほかにこのソフトでできること

患者番号	レセプトコード	名称	投与量	点数	実施日
1000036	610443053	バイアスピリン錠100mg	1	5.8	20101220
1000036	610443053	バイアスピリン錠100mg	1	5.8	20101224
1000036	610443053	バイアスピリン錠100mg	1	5.8	20101229
1000036	610443053	バイアスピリン錠100mg	1	5.8	20101229
1002404	610443053	バイアスピリン錠100mg	1	5.8	20100805
1002404	610443053	バイアスピリン錠100mg	1	5.8	20100812
1002404	620002191	グリセオール注200ml	3	909	20100729
1002404	620002191	グリセオール注200ml	3	909	20100730

ログ型データから

	バイアスピリン			グリセオール		
患者番号	投与量計	初回投与日	最終投与日	投与量計	初回投与日	最終投与日
1000036	4	20101220	20101229			
1002404	2	20100805	20100812	6	20100729	20100730

1患者1レコード型へのデータ変換

これで統計ソフトで、分析ができる

1患者1レコード型のデータ変換スクリプト

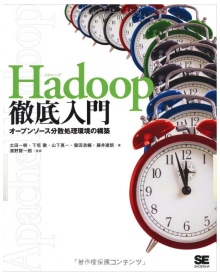
- 現在鋭意プログラム開発中
- シンプルに表現できるスクリプトでこの処理ができるようになる予定
- 目標今年夏が終わるまで。

- 今のソフトのままでも実現可能ですが、スクリプトが複雑/速度が遅いので現在公開を見合わせています。

自分でスクリプトが書きたい

- もちろん自分で書くこともできます。
- 日本語で紹介されている本で使えるものはこの本
- 3800円　450ページ
- Pigのスクリプトについて書かれている部分は20ページあまりです。

- 分散処理そのものに興味がある人には非常におすすめです。



まとめ DPCデータを分析に使うには

- HISによるシステム化
- DPC分析ソフト(サービス)の利用
  - ARROWS/EVE/ヒラソル・・・・・

高
有償
低い

- 自分でDB化して分析

無料

- DPCデータから直接分析をする

研究班新規のやり方の提案