An introduction to radiology infrastructure in clinical practice

Mark Thurston

Consultant radiologist
University Plymouth NHS Trust

Mark.Thurston@nhs.net

Objectives

- Overview of systems in place for medical imaging studies
  - Appointments, reports, and other metadata
  - Pixel data
    - Acquisition, storage, display
- Background for computer scientists and neurologists
  - Inspire ideas for research
- Image download pipeline

Radiology information system (RIS)

- Management information system, specific to radiology
- A database client used by all staff
  - Administrative staff
  - Radiographers
  - Reporters: integrated voice recognition
- Stores all metadata about the pixel data
  - Separate to the PACS system
  - Unique identifiers common to both systems
  - Link between PACS and RIS: XML-RPC/COM
- Some hospitals don’t use a separate RIS
  - In-PACS reporting

DATABASE RESEARCH GROUP
University of California at Berkeley
POSTGRES * Tioga browser * Mariposa DBMS * S2K storage system research

RIS

- Main brands in UK
    - Wellbeing CRIS (inc. UHP)
    - Soliton

- Architecture:
    - CRIS
        - Postgres 9 on RHEL Linux Relational Database Management System backend
        - Java 8 frontend, perhaps a web frontend at other sites

    - Soliton
        - (Probably) MS SQL server on Windows Server
        - .NET frontend

- Peninsula region currently use one Postgres database for the region (5 hospitals)
    - Production write database instance
    - Overnight replication to a read only reporting DB

CRIS database schema

- Complicated database schema
    - not really possible to illustrate full extent with a single entity relation diagram

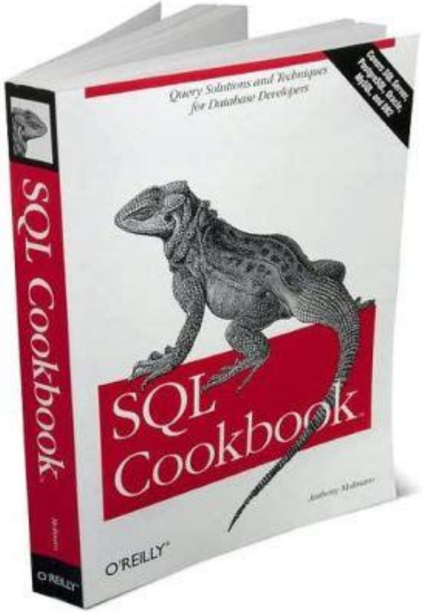
- Large amount of organisational and diagnostic data. E.g:
    - Appointment times
    - Scan vetting information
    - Free text reports
        - Some structure/standardised phrases
    - 320GB for the whole region

RIS research skills

- Database administration for initial extracts
  - SQL (Postgres dialect)
  - Resources
    - SQLZoo: https://sqlzoo.net/
    - Postgres official documentation: https://www.postgresql.org/docs/current/
    - SQL Cookbook, Molinaro, O'Reilly

- Tabular data analysis
  - Python/Pandas, R
    - Multiple online tutorials (including Kaggle and Hackerrank)
    - Python for Data Analysis, McKinney, O'Reilly

SQLAlchemy

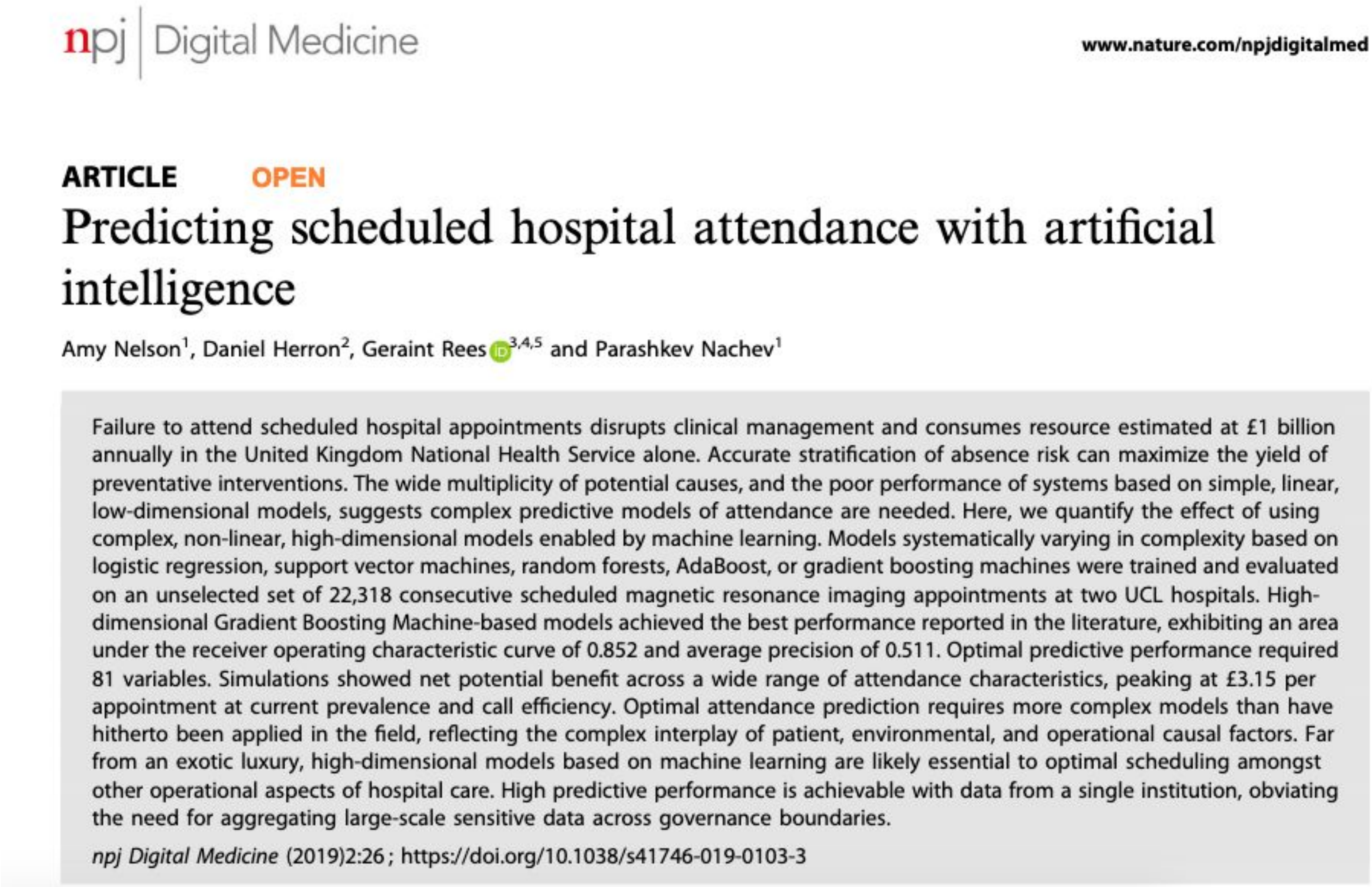


Learning the RIS

- Unfortunately not possible to get a fake patient schema from the vendor for training staff
    - “due to commercial sensitivity”

- Postgres
    - Open source, easy to learn
    - Example database schemas: https://github.com/xivSolutions/ChinookDb_Pg_Modified

- Might be worth considering running PD registry databases on RDBMS
    - Postgres, Maria, or MS SQL



npj Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Predicting scheduled hospital attendance with artificial intelligence

Amy Nelson[1], Daniel Herron[2], Geraint Rees[3,4,5] and Parashkev Nachev[1]

Failure to attend scheduled hospital appointments disrupts clinical management and consumes resource estimated at £1 billion annually in the United Kingdom National Health Service alone. Accurate stratification of absence risk can maximize the yield of preventative interventions. The wide multiplicity of potential causes, and the poor performance of systems based on simple, linear, low-dimensional models, suggests complex predictive models of attendance are needed. Here, we quantify the effect of using complex, non-linear, high-dimensional models enabled by machine learning. Models systematically varying in complexity based on logistic regression, support vector machines, random forests, AdaBoost, or gradient boosting machines were trained and evaluated on an unselected set of 22,318 consecutive scheduled magnetic resonance imaging appointments at two UCL hospitals. High-dimensional Gradient Boosting Machine-based models achieved the best performance reported in the literature, exhibiting an area under the receiver operating characteristic curve of 0.852 and average precision of 0.511. Optimal predictive performance required 81 variables. Simulations showed net potential benefit across a wide range of attendance characteristics, peaking at £3.15 per appointment at current prevalence and call efficiency. Optimal attendance prediction requires more complex models than have hitherto been applied in the field, reflecting the complex interplay of patient, environmental, and operational causal factors. Far from an exotic luxury, high-dimensional models based on machine learning are likely essential to optimal scheduling amongst other operational aspects of hospital care. High predictive performance is achievable with data from a single institution, obviating the need for aggregating large-scale sensitive data across governance boundaries.

npj Digital Medicine (2019)2:26 ; https://doi.org/10.1038/s41746-019-0103-3

PACS

- Picture Archiving and Communication System
    - Refers to both the **storage server** and the viewing client
    - Image data resides on PACS

- PACS uses the DICOM standard
    - File format
    - Network protocol

- Derriford uses Insignia PACS
    - Orthanc as test bed

DICOM

- Allows connect and download image data
    - Dicomserver.co.uk
    - Orthanc
    - https://tutorial.mdvthu.com/

- Open source client software available
    - Pynetdicom
    - DCMTK

Modalities

- “Modality” refers to the image acquisition device:
  - MRI, Plain radiograph (== x-ray), CT, Nuclear medicine, Ultrasound etc.

- Modalities will store images in DICOM file format and transfer DICOM files to PACS using the DICOM network protocol (over TCP/IP)

- Many different brands and different devices - all with slightly different properties
  - End of life kit (e.g. Win XP embedded, Linux 2.x s.i.c.) in regular use
  - Very difficult to get vendors to upgrade (including critical security vulnerabilities)

- Different brand modalities, different DICOM values
  - Importance of thorough testing for anonymisation pipeline

Modality examples

Image download pipeline: progress to date

1. SQL search to identify unique IDs for studies of interest
    a. Python/SQLAlchemy
    b. Pandas

2. Bash helper script to connect and download the study from PACS
    a. Native Python pynetdicom is more involved than using DCMTK
    b. Linux/Bash

3. Post processing of pixel data
    a. Python/Matplotlib
    b. anonymisation/renaming
        i. use a sha256 hash to ensure unique studies but without reverse lookup of patient ID

Image download: SQL logic

- Aim:
  - **Obtain a unique identifier for each study of interest**
- **Connects to the radiology information system**
- Pacemaker CXRs training sets:
  - pre- pacemaker insertion
  - post- pacemaker insertion CXR
- Parkinson's patients:
  - A function that accepts an NHS number from the registry list and returns a list of unique IDs for all relevant head imaging
- Can't be done easily without direct database access
  - Stats tool frontend to CRIS is very limited

DICOM logic (and bugs)

- Bash helper script
  - Accepts unique identifier
  - Connects to locally hosted PACS
  - Dumps image data into a predefined directory
    - Training, test, validation split created afterwards

- Not very resilient
  - Error handling needs to be improved to maximise usable data

- Anonymisation logic: **risk**
  - review of images from various scanners
  - Compare with official anonymisation standard

Manual review (post download)

- Labelling of image data is very important
    - No Mechanical Turk for radiologists (like *ImageNet)*
        - outsourcing could perhaps be used for large projects but would be expensive

- 2 board certified radiologists, post download
    - Good quality SQL reduces workload significantly

- Arbitration process
    - Informal review of discrepancies

Data structures

- *“[Design] your code around the data, rather than the other way around”*
- Ideally, decide on how to store the downloaded data before starting the download
  - Consider the metadata carefully
- Evolution of data structures may require redownload

Pacemaker data structures

- 2 folders
  - RCHT/Paced; UHP/Paced
  - RCHT/ Unpaced; UHP/Unpaced
- Images named as:
  - Hash value of unique identifier
  - .jpg and .npy
- Bash script to create random symbolic links
  - Training/Paced; Training/Unpaced
  - Test/Paced; Test/Unpaced

Parkinson’s disease

- Progress
    - Draft SQL search has been created, for Plymouth patients
        - identification of scans for inclusion

- Priorities
    - Formulate data structures
        - How much metadata do we need to keep?
        - Minimum possible
        - Need to maintain compliance with approved research protocol and IG requirements

    - Robust testing of volumetric DICOM datasets to ensure appropriate anonymisation
        - Differing modalities

?

Questions