National Cancer Registry Migration:
A database case study

Brian Shand

ecric
Fighting cancer with information

Eastern Cancer Registration
and Information Centre
http://www.ecric.nhs.uk/

Brian.Shand@cbcu.nhs.uk

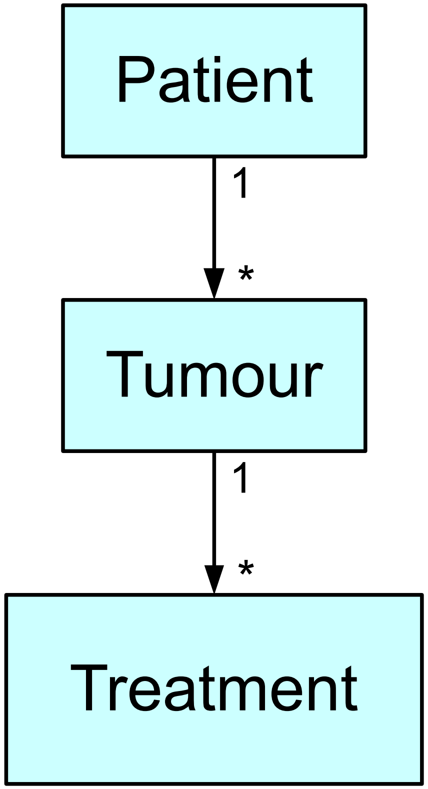
NHS

Overview

- Cancer registration overview
- Registry database structures
- Web-based access (Ruby on Rails) and data security
- Automation and electronic data processing
- Registry migration and scalability

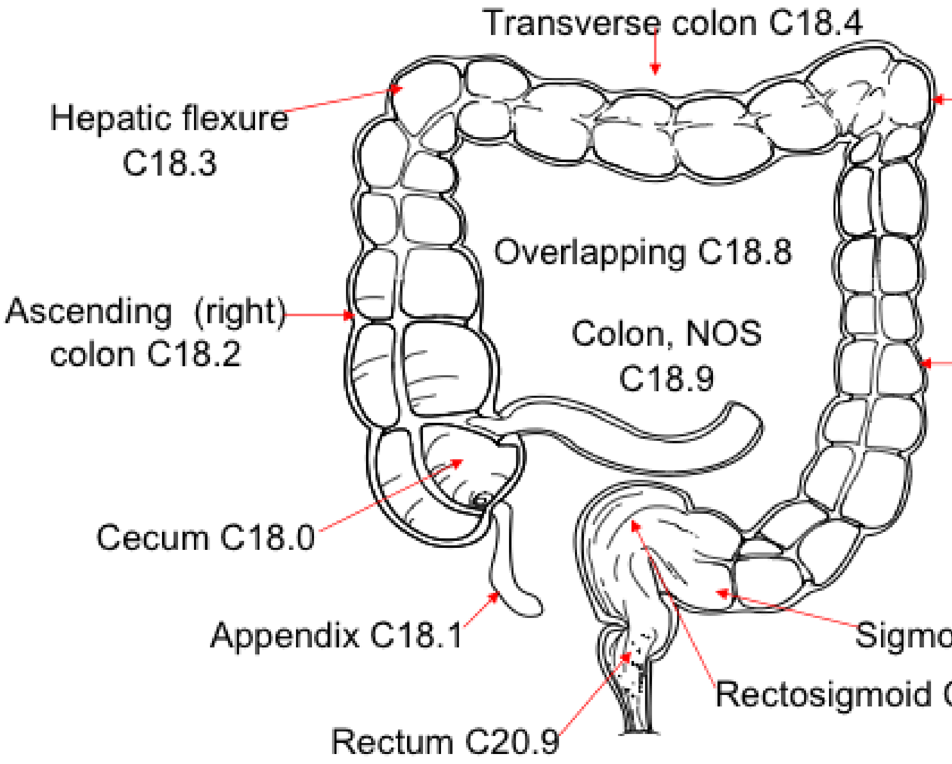
Cancer registration

- UK cancer registries collect population-based data about cancer incidence and mortality
  - Long-term statistical analysis
  - More immediate uses: clinical audit, planning for service delivery
  - Special authorisation to collect cancer data (NHS Act 2006 Section 251)
  - Expert knowledge provides a synthesis of available information, not simply an amalgamation of data
- Historically a loose federation of independent databases
  - Shared minimum dataset sent to Office for National Statistics (ONS)



Cancer registration (2)

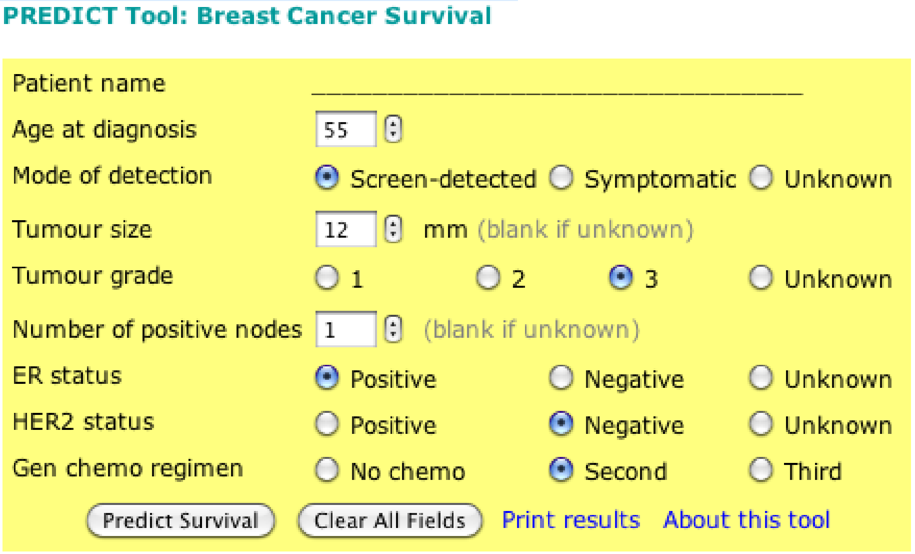
- Tumour registrations are the primary output
  - Tumour site
    C18.4 = transverse colon [ICD-O-3]
  - Morphology/behaviour
    8140/3 = adenocarcinoma
  - Stage at diagnosis
    T1 N0 M0 = stage I
  - Patient demographics,
    e.g. birth date, name,
    postcode at diagnosis
  - Treatment received
  - Hospitals where treated
- High data quality is needed to analyse rare tumours



Graphic from CS Steering Co

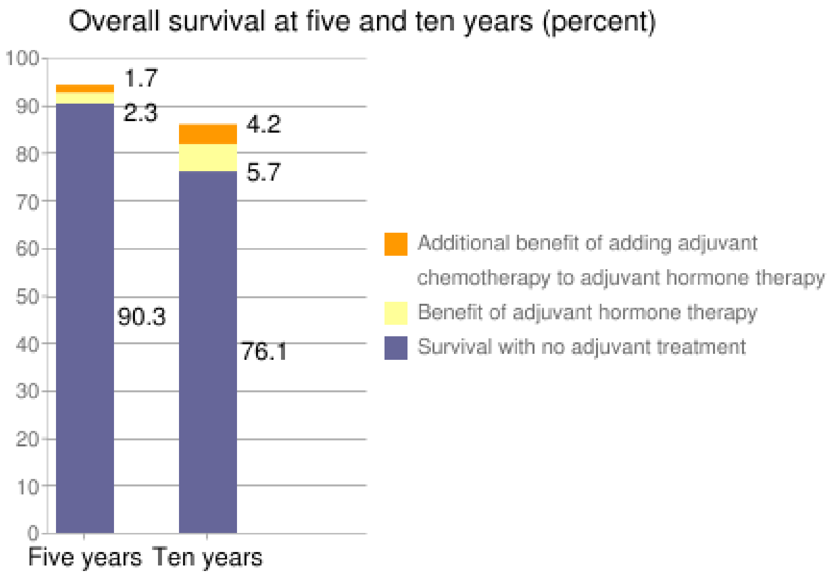
Examples of data use

- Historically: asbestosis, smoking causes cancer
- Melanoma study
  - Identified that patients are presenting earlier due to public awareness.
  - Early treatment has increased survival statistics.
- Predict tool
  - Helps patients and doctors choose the best course of treatment after breast surgery.



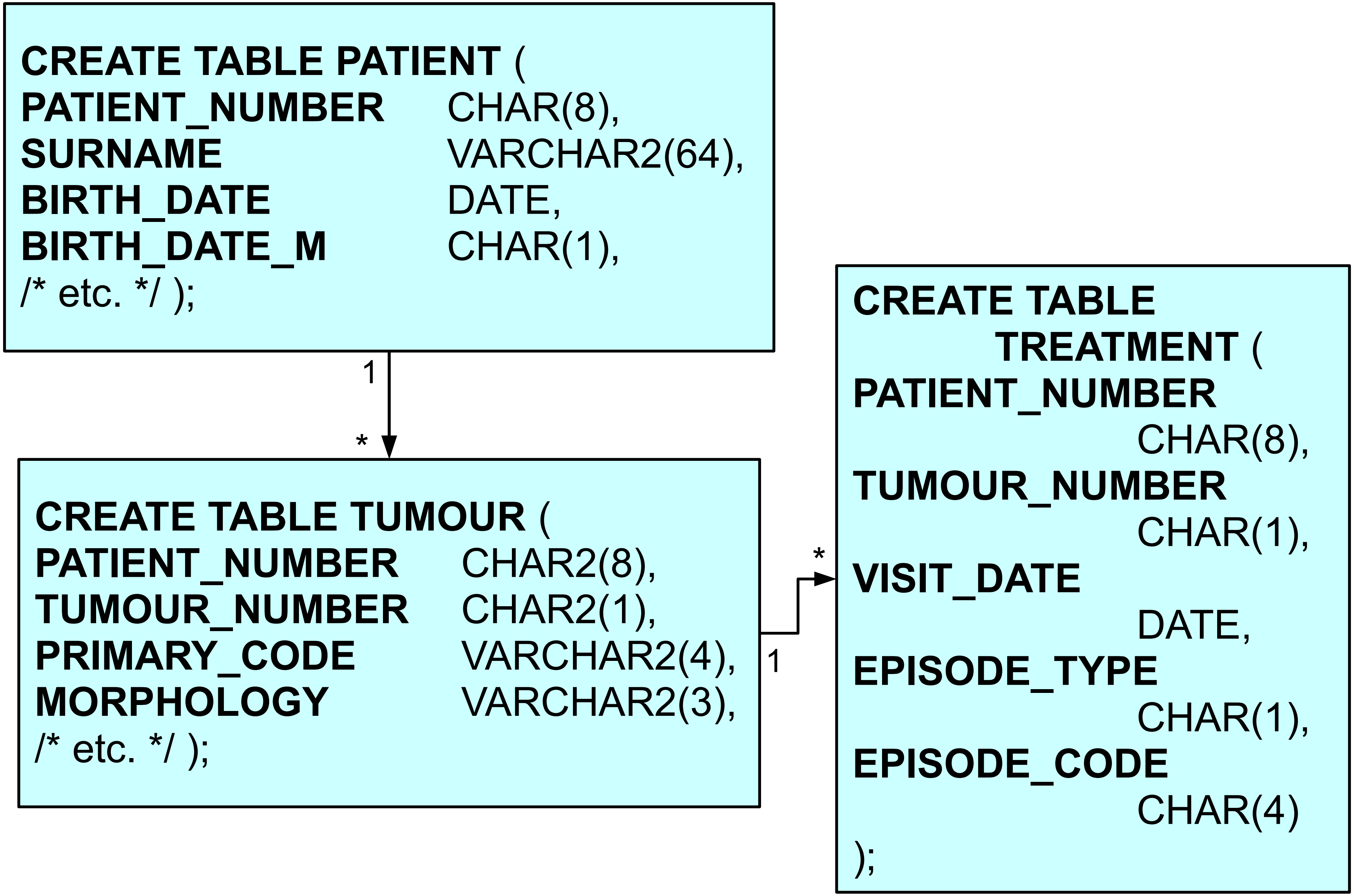
PREDICT Tool: Breast Cancer Survival

Patient name	______________________________			
Age at diagnosis	55			
Mode of detection	● Screen-detected	○ Symptomatic	○ Unknown	
Tumour size	12	mm (blank if unknown)		
Tumour grade	○ 1	○ 2	● 3	○ Unknown
Number of positive nodes	1	(blank if unknown)		
ER status	● Positive	○ Negative	○ Unknown	
HER2 status	○ Positive	● Negative	○ Unknown	
Gen chemo regimen	○ No chemo	● Second	○ Third	



Overall survival at five and ten years (percent)

ECRIC's old database schema

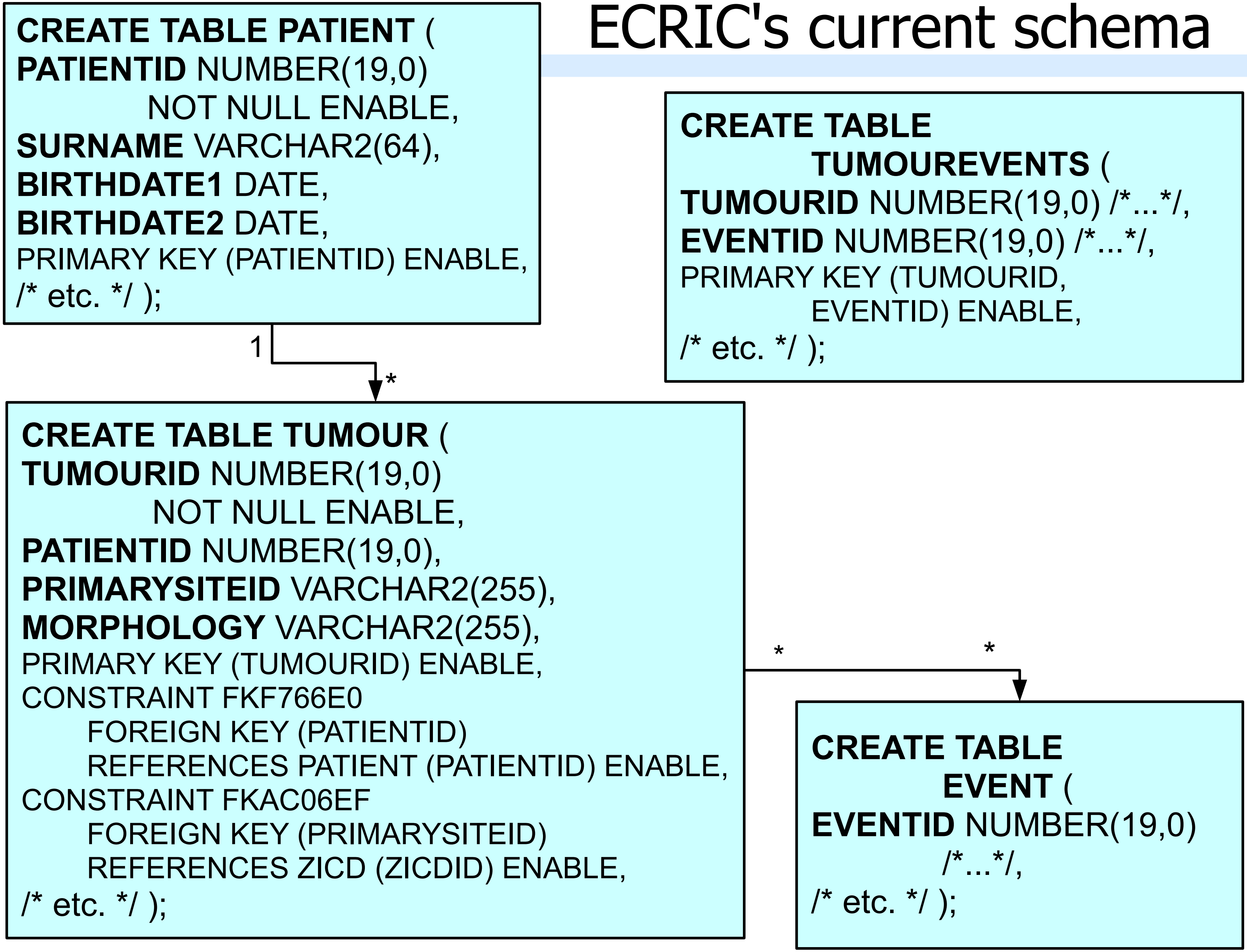


```
CREATE TABLE PATIENT (
PATIENT_NUMBER      CHAR(8),
SURNAME             VARCHAR2(64),
BIRTH_DATE          DATE,
BIRTH_DATE_M        CHAR(1),
/* etc. */ );
```

```
CREATE TABLE TUMOUR (
PATIENT_NUMBER      CHAR2(8),
TUMOUR_NUMBER       CHAR2(1),
PRIMARY_CODE        VARCHAR2(4),
MORPHOLOGY          VARCHAR2(3),
/* etc. */ );
```

```
CREATE TABLE
          TREATMENT (
PATIENT_NUMBER
                    CHAR(8),
TUMOUR_NUMBER
                    CHAR(1),
VISIT_DATE
                    DATE,
EPISODE_TYPE
                    CHAR(1),
EPISODE_CODE
                    CHAR(4)
);
```

ECRIC's current schema



```
CREATE TABLE PATIENT (
PATIENTID NUMBER(19,0)
          NOT NULL ENABLE,
SURNAME VARCHAR2(64),
BIRTHDATE1 DATE,
BIRTHDATE2 DATE,
PRIMARY KEY (PATIENTID) ENABLE,
/* etc. */ );
```

```
CREATE TABLE
          TUMOUREVENTS (
TUMOURID NUMBER(19,0) /*...*/,
EVENTID NUMBER(19,0) /*...*/,
PRIMARY KEY (TUMOURID,
          EVENTID) ENABLE,
/* etc. */ );
```

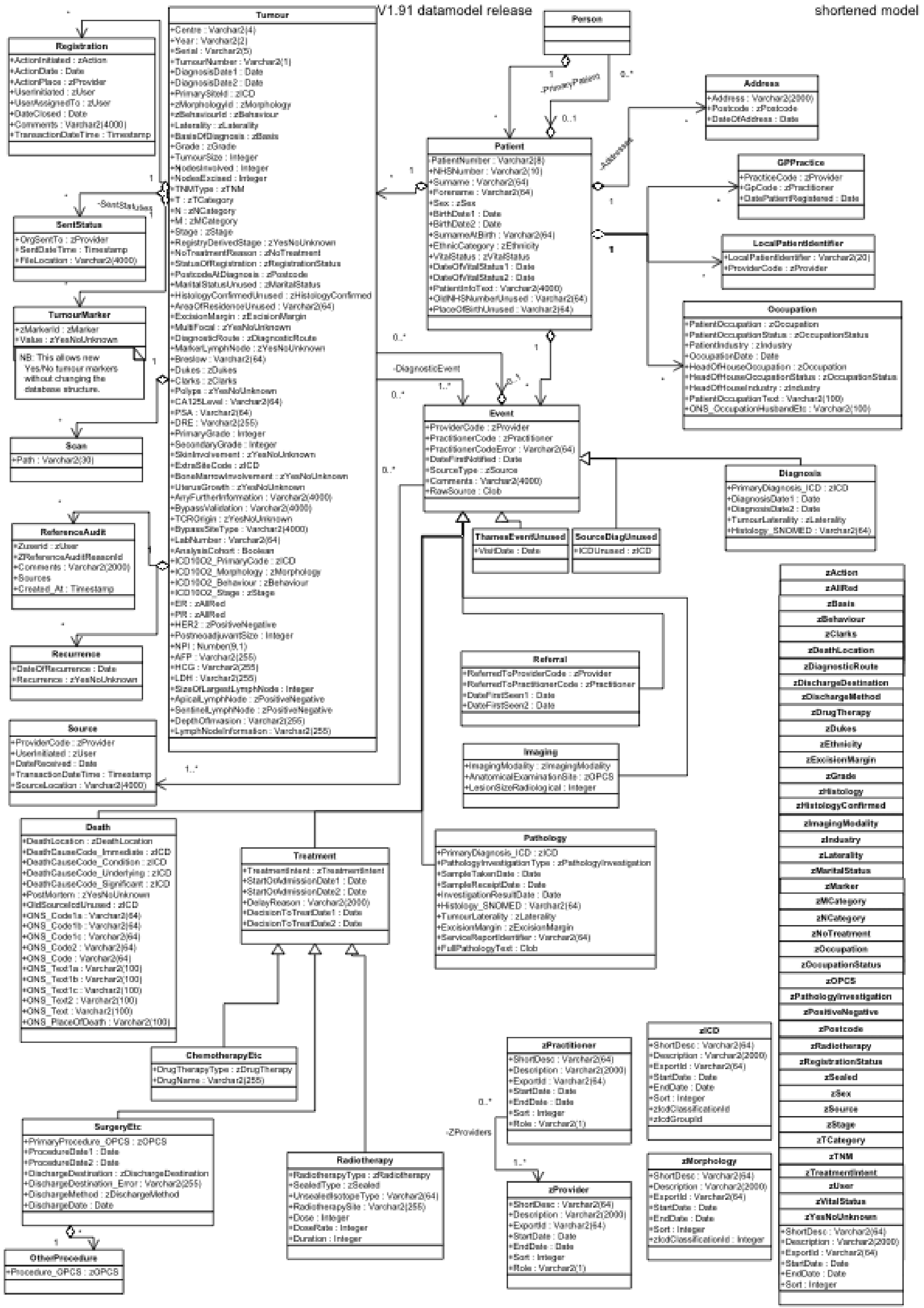
```
CREATE TABLE TUMOUR (
TUMOURID NUMBER(19,0)
          NOT NULL ENABLE,
PATIENTID NUMBER(19,0),
PRIMARYSITEID VARCHAR2(255),
MORPHOLOGY VARCHAR2(255),
PRIMARY KEY (TUMOURID) ENABLE,
CONSTRAINT FKF766E0
     FOREIGN KEY (PATIENTID)
     REFERENCES PATIENT (PATIENTID) ENABLE,
CONSTRAINT FKAC06EF
     FOREIGN KEY (PRIMARYSITEID)
     REFERENCES ZICD (ZICDID) ENABLE,
/* etc. */ );
```

```
CREATE TABLE
          EVENT (
EVENTID NUMBER(19,0)
          /*...*/,
/* etc. */ );
```

ECRIC's current schema (2)

- So it's a better schema
  - Primary keys on every table
  - Foreign key constraints wherever possible
    - PRIMARYSITEID includes the classification system
  - Well normalised (can link 1 pathology report to 2 tumours)
  - Date ranges instead of approximate dates
- But it's changing in nature
  - We can now store anything / everything we receive
  - Now, instead of simply an expert summary, it also encapsulates the backing data
    - and audits changes to the core data
    - and structural changes to coding spaces over time e.g. ICD10-O-2 vs ICD-O-3 vs ICD-O-3 (2011 update)

- Before you know it, the core of your data model looks like this ---------->
- Broadly consistent with HL7 version 3 structures
- It's changed from a summary of the data to as much original data as possible, plus summary information.



V1.91 datamodel release

shortened model

Tumour
+Centre : Varchar2(4)
+Year : Varchar2(2)
+Serial : Varchar2(5)
+TumourNumber : Varchar2(1)
+DiagnosisDate1 : Date
+DiagnosisDate2 : Date
+PrimarySiteId : zICD
+zMorphologyId : zMorphology
+zBehaviourId : zBehaviour
+Laterality : zLaterality
+BasisOfDiagnosis : zBasis
+Grade : zGrade
+TumourSize : Integer
+NodesInvolved : Integer
+NodesExcised : Integer
+TNMType : zTNM
+T : zTCategory
+N : zNCategory
+M : zMCategory
+Stage : zStage
+RegistryDerivedStage : zYesNoUnknown
+NoTreatmentReason : zNoTreatment
+StatusOfRegistration : zRegistrationStatus
+PostcodeAtDiagnosis : zPostcode
+MaritalStatusUnused : zMaritalStatus
+HistologyConfirmedUnused : zHistologyConfirmed
+AreaOfResidenceUnused : Varchar2(64)
+ExcisionMargin : zExcisionMargin
+MultiFocal : zYesNoUnknown
+DiagnosticRoute : zDiagnosticRoute
+MarkerLymphNode : zYesNoUnknown
+Breslow : Varchar2(64)
+Dukes : zDukes
+Clarks : zClarks
+Polyps : zYesNoUnknown
+CA125Level : Varchar2(64)
+PSA : Varchar2(64)
+DRE : Varchar2(255)
+PrimaryGrade : Integer
+SecondaryGrade : Integer
+SkinInvolvement : zYesNoUnknown
+ExtraSiteCode : zICD
+BoneMarrowInvolvement : zYesNoUnknown
+UterusGrowth : zYesNoUnknown
+AnyFurtherInformation : Varchar2(4000)
+BypassValidation : Varchar2(4000)
+TCROrigin : zYesNoUnknown
+BypassSiteType : Varchar2(4000)
+LabNumber : Varchar2(64)
+AnalysisCohort : Boolean
+ICD10O2_PrimaryCode : zICD
+ICD10O2_Morphology : zMorphology
+ICD10O2_Behaviour : zBehaviour
+ICD10O2_Stage : zStage
+ER : zAllRed
+PR : zAllRed
+HER2 : zPositiveNegative
+PostneoadjuvantSize : Integer
+NPI : Number(9,1)
+AFP : Varchar2(255)
+HCG : Varchar2(255)
+LDH : Varchar2(255)
+SizeOfLargestLymphNode : Integer
+ApicalLymphNode : zPositiveNegative
+SentinelLymphNode : zPositiveNegative
+DepthOfInvasion : Varchar2(255)
+LymphNodeInformation : Varchar2(255)

Registration
+ActionInitiated : zAction
+ActionDate : Date
+ActionPlace : zProvider
+UserInitiated : zUser
+UserAssignedTo : zUser
+DateClosed : Date
+Comments : Varchar2(4000)
+TransactionDateTime : Timestamp

SentStatus
+OrgSentTo : zProvider
+SentDateTime : Timestamp
+FileLocation : Varchar2(4000)

TumourMarker
+zMarkerId : zMarker
+Value : zYesNoUnknown

NB: This allows new Yes/No tumour markers without changing the database structure.

Scan
+Path : Varchar2(30)

ReferenceAudit
+Zuserid : zUser
+ZReferenceAuditReasonId
+Comments : Varchar2(2000)
+Sources
+Created_At : Timestamp

Recurrence
+DateOfRecurrence : Date
+Recurrence : zYesNoUnknown

Source
+ProviderCode : zProvider
+UserInitiated : zUser
+DateReceived : Date
+TransactionDateTime : Timestamp
+SourceLocation : Varchar2(4000)

Person

Patient
-PatientNumber : Varchar2(8)
+NHSNumber : Varchar2(10)
+Surname : Varchar2(64)
+Forename : Varchar2(64)
+Sex : zSex
+BirthDate1 : Date
+BirthDate2 : Date
+SurnameAtBirth : Varchar2(64)
+EthnicCategory : zEthnicity
+VitalStatus : zVitalStatus
+DateOfVitalStatus1 : Date
+DateOfVitalStatus2 : Date
+PatientInfoText : Varchar2(4000)
+OldNHSNumberUnused : Varchar2(64)
+PlaceOfBirthUnused : Varchar2(64)

Address
+Address : Varchar2(2000)
+Postcode : zPostcode
+DateOfAddress : Date

GPPractice
+PracticeCode : zProvider
+GpCode : zPractitioner
+DatePatientRegistered : Date

LocalPatientIdentifier
+LocalPatientIdentifier : Varchar2(20)
+ProviderCode : zProvider

Occupation
+PatientOccupation : zOccupation
+PatientOccupationStatus : zOccupationStatus
+PatientIndustry : zIndustry
+OccupationDate : Date
+HeadOfHouseOccupation : zOccupation
+HeadOfHouseOccupationStatus : zOccupationStatus
+HeadOfHouseIndustry : zIndustry
+PatientOccupationText : Varchar2(100)
+ONS_OccupationHusbandEtc : Varchar2(100)

Event
+ProviderCode : zProvider
+PractitionerCode : zPractitioner
+PractitionerCodeError : Varchar2(64)
+DateFirstNotified : Date
+SourceType : zSource
+Comments : Varchar2(4000)
+RawSource : Clob

ThamesEventUnused
+VisitDate : Date

SourceDiagUnused
+ICDUnused : zICD

Diagnosis
+PrimaryDiagnosis_ICD : zICD
+DiagnosisDate1 : Date
+DiagnosisDate2 : Date
+TumourLaterality : zLaterality
+Histology_SNOMED : Varchar2(64)

Referral
+ReferredToProviderCode : zProvider
+ReferredToPractitionerCode : zPractitioner
+DateFirstSeen1 : Date
+DateFirstSeen2 : Date

Imaging
+ImagingModality : zImagingModality
+AnatomicalExaminationSite : zOPCS
+LesionSizeRadiological : Integer

Pathology
+PrimaryDiagnosis_ICD : zICD
+PathologyInvestigationType : zPathologyInvestigation
+SampleTakenDate : Date
+SampleReceiptDate : Date
+InvestigationResultDate : Date
+Histology_SNOMED : Varchar2(64)
+TumourLaterality : zLaterality
+ExcisionMargin : zExcisionMargin
+ServiceReportIdentifier : Varchar2(64)
+FullPathologyText : Clob

Death
+DeathLocation : zDeathLocation
+DeathCauseCode_Immediate : zICD
+DeathCauseCode_Condition : zICD
+DeathCauseCode_Underlying : zICD
+DeathCauseCode_Significant : zICD
+PostMortem : zYesNoUnknown
+OldSourceIcdUnused : zICD
+ONS_Code1a : Varchar2(64)
+ONS_Code1b : Varchar2(64)
+ONS_Code1c : Varchar2(64)
+ONS_Code2 : Varchar2(64)
+ONS_Code : Varchar2(64)
+ONS_Text1a : Varchar2(100)
+ONS_Text1b : Varchar2(100)
+ONS_Text1c : Varchar2(100)
+ONS_Text2 : Varchar2(100)
+ONS_Text : Varchar2(100)
+ONS_PlaceOfDeath : Varchar2(100)

Treatment
+TreatmentIntent : zTreatmentIntent
+StartOfAdmissionDate1 : Date
+StartOfAdmissionDate2 : Date
+DelayReason : Varchar2(2000)
+DecisionToTreatDate1 : Date
+DecisionToTreatDate2 : Date

ChemotherapyEtc
+DrugTherapyType : zDrugTherapy
+DrugName : Varchar2(255)

SurgeryEtc
+PrimaryProcedure_OPCS : zOPCS
+ProcedureDate1 : Date
+ProcedureDate2 : Date
+DischargeDestination : zDischargeDestination
+DischargeDestination_Error : Varchar2(255)
+DischargeMethod : zDischargeMethod
+DischargeDate : Date

OtherProcedure
+Procedure_OPCS : zOPCS

Radiotherapy
+RadiotherapyType : zRadiotherapy
+SealedType : zSealed
+UnsealedIsotopeType : Varchar2(64)
+RadiotherapySite : Varchar2(255)
+Dose : Integer
+DoseRate : Integer
+Duration : Integer

zPractitioner
+ShortDesc : Varchar2(64)
+Description : Varchar2(2000)
+ExportId : Varchar2(64)
+StartDate : Date
+EndDate : Date
+Sort : Integer
+Role : Varchar2(1)

zProvider
+ShortDesc : Varchar2(64)
+Description : Varchar2(2000)
+ExportId : Varchar2(64)
+StartDate : Date
+EndDate : Date
+Sort : Integer
+Role : Varchar2(1)

zICD
+ShortDesc : Varchar2(64)
+Description : Varchar2(2000)
+ExportId : Varchar2(64)
+StartDate : Date
+EndDate : Date
+Sort : Integer
+zIcdClassificationId
+zIcdGroupId

zMorphology
+ShortDesc : Varchar2(64)
+Description : Varchar2(2000)
+ExportId : Varchar2(64)
+StartDate : Date
+EndDate : Date
+Sort : Integer
+zIcdClassificationId : Integer

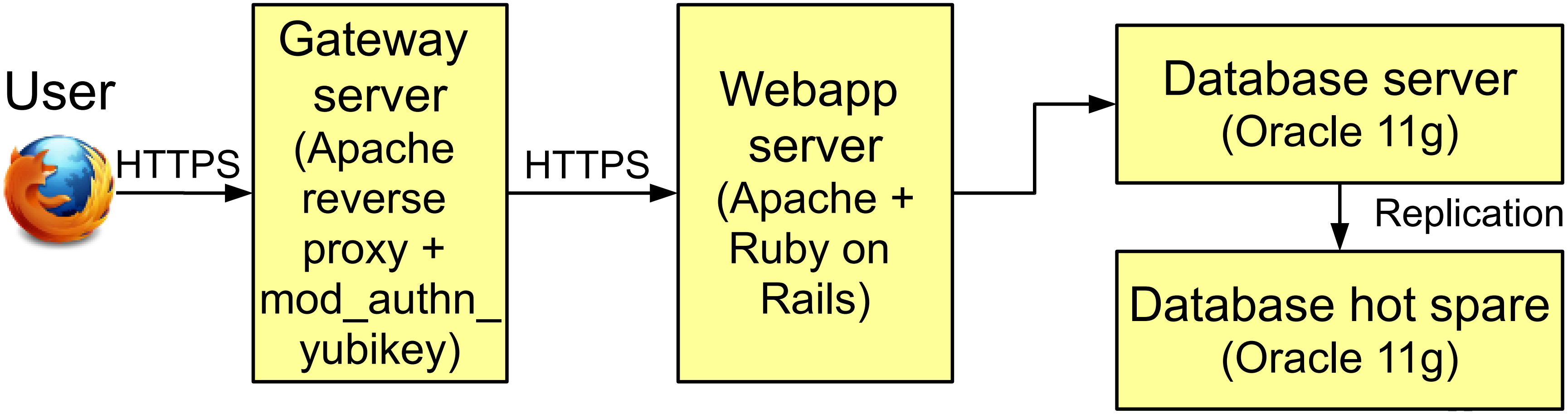
zAction
zAllRed
zBasis
zBehaviour
zClarks
zDeathLocation
zDiagnosticRoute
zDischargeDestination
zDischargeMethod
zDrugTherapy
zDukes
zEthnicity
zExcisionMargin
zGrade
zHistology
zHistologyConfirmed
zImagingModality
zIndustry
zLaterality
zMaritalStatus
zMarker
zMCategory
zNCategory
zNoTreatment
zOccupation
zOccupationStatus
zOPCS
zPathologyInvestigation
zPositiveNegative
zPostcode
zRadiotherapy
zRegistrationStatus
zSealed
zSex
zSource
zStage
zTCategory
zTNM
zTreatmentIntent
zUser
zVitalStatus
zYesNoUnknown
+ShortDesc : Varchar2(64)
+Description : Varchar2(2000)
+ExportId : Varchar2(64)
+StartDate : Date
+EndDate : Date
+Sort : Integer

ECRIC database schema overview

- The schema represents the core of the database
  - It aims to follow the Cancer Registration Dataset in the NHS Data Dictionary wherever practical http://www.datadictionary.nhs.uk/data_dictionary/messages/clinical_data_sets/cancer_registration_data_set_fr.asp?shownav=1
- The database is not the complete workflow
  - People and physical workflows
  - It's taken years to stop turning every tumour registration into paper
  - Secondary databases are also hard to avoid
  - Auxilliary tables (not shown) drive the import of electronic data sources

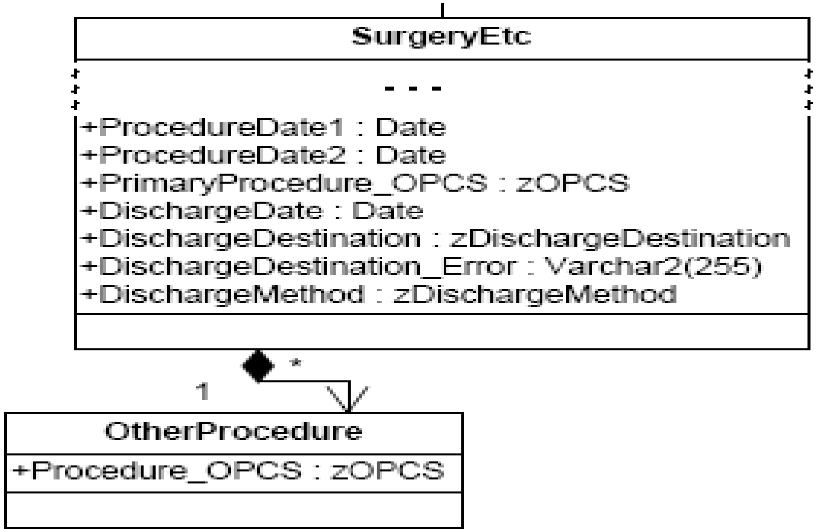
Overview (2)

- Cancer registration overview
- Registry database structures
- **Web-based access (Ruby on Rails) and data security**
- Automation and electronic data processing
- Registry migration and scalability


User
HTTPS
Gateway server (Apache reverse proxy + mod_authn_yubikey)
HTTPS
Webapp server (Apache + Ruby on Rails)
Database server (Oracle 11g)
Replication
Database hot spare (Oracle 11g)

Database

- Oracle database back-end provides high availability and scalable performance
  - Multiple redundant backups (including off-site) without downtime; continuous redo logs
  - Triggers record a forensic timeline of all data changes
- Table structures follow the National Cancer Data Set



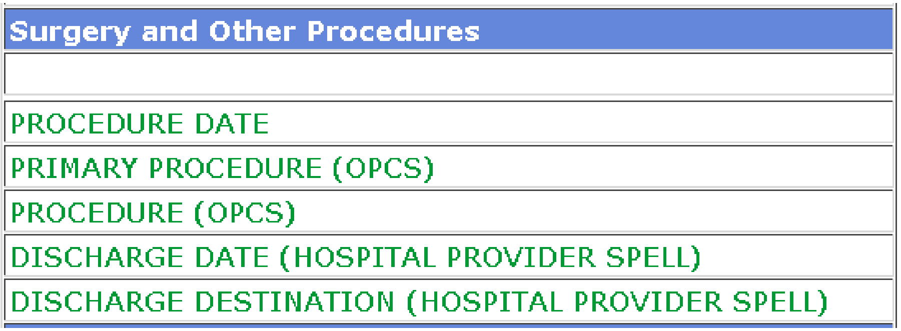
SurgeryEtc

- - -

+ProcedureDate1 : Date
+ProcedureDate2 : Date
+PrimaryProcedure_OPCS : zOPCS
+DischargeDate : Date
+DischargeDestination : zDischargeDestination
+DischargeDestination_Error : Varchar2(255)
+DischargeMethod : zDischargeMethod

OtherProcedure

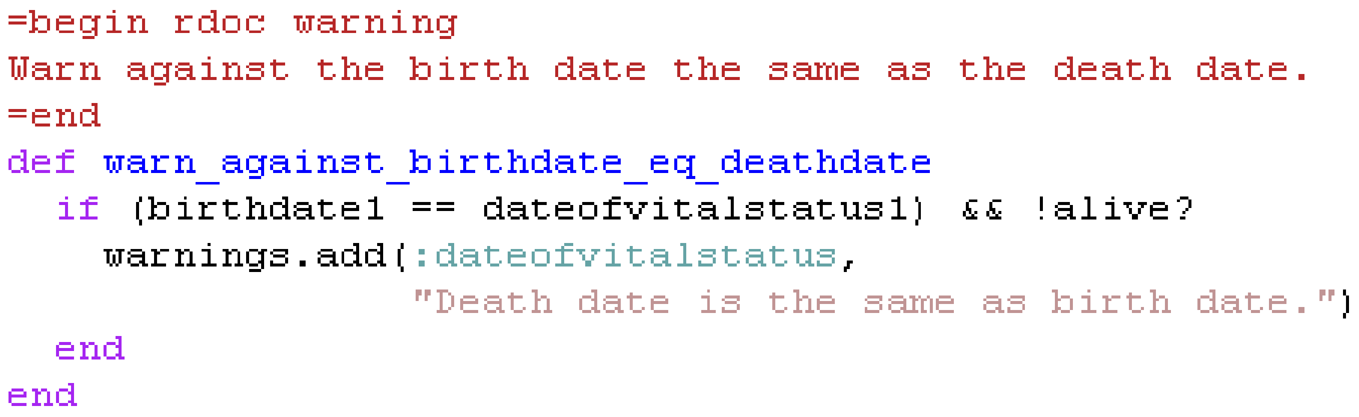
+Procedure_OPCS : zOPCS



Surgery and Other Procedures
PROCEDURE DATE
PRIMARY PROCEDURE (OPCS)
PROCEDURE (OPCS)
DISCHARGE DATE (HOSPITAL PROVIDER SPELL)
DISCHARGE DESTINATION (HOSPITAL PROVIDER SPELL)

Application server

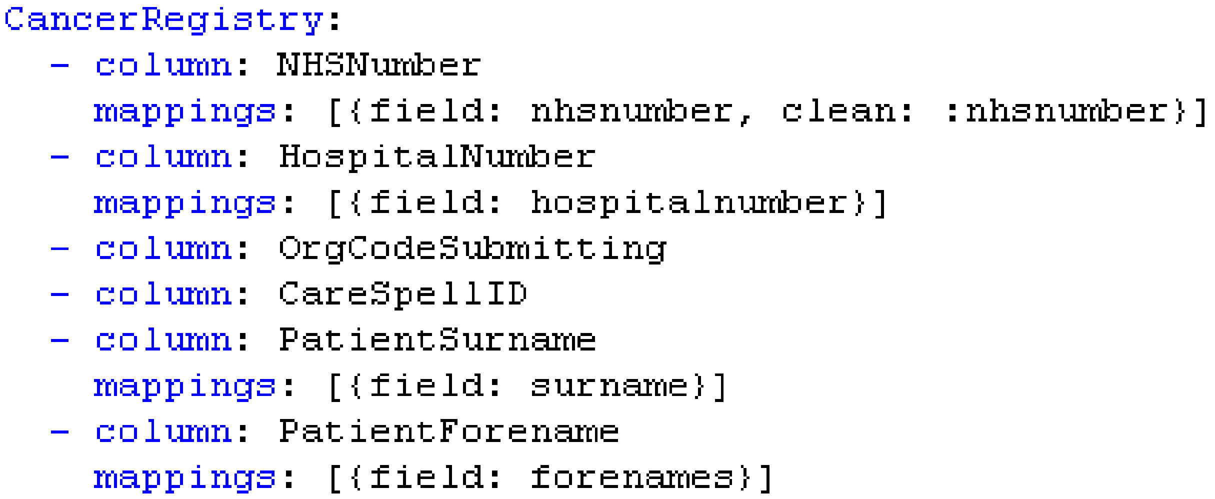
- Ruby on Rails provides a responsive system, allowing continuous, incremental evolution
- Consistent data validations apply to all new data
  - This includes automated processing, interactive import of electronic data, and manual data entry
  - As validations evolve, historical data can be reassessed, improving the quality of the whole dataset
  - Warnings also protect against common potential errors



```
=begin rdoc warning
Warn against the birth date the same as the death date.
=end
def warn_against_birthdate_eq_deathdate
  if (birthdate1 == dateofvitalstatus1) && !alive?
    warnings.add(:dateofvitalstatus,
                 "Death date is the same as birth date.")
  end
end
```

Electronic data processing

- Electronic data sources are processed as soon as possible after receipt
  - Automated scripts scan for new data from regular feeds
  - Quick processing enables rapid QA of data source quality
- Source mapping files simplify adding new data sources, e.g. Somerset mapping snippet:



```
CancerRegistry:
  - column: NHSNumber
    mappings: [{field: nhsnumber, clean: :nhsnumber}]
  - column: HospitalNumber
    mappings: [{field: hospitalnumber}]
  - column: OrgCodeSubmitting
  - column: CareSpellID
  - column: PatientSurname
    mappings: [{field: surname}]
  - column: PatientForename
    mappings: [{field: forenames}]
```

Workflow

- Electronic data sources pass through a customisable, multi-stage workflow
  - Preprocessing (using Monarch or Ruby)
  - Validation of format, postcodes, and internal data consistency
  - Tracing
  - Automatic patient matching
  - Manual patient matching (of ambiguous matches)
  - Record deduplication
    - E.g. automatically handles overlap between NBTR vs ECRIC cancer waits records
  - Assignment of batches of work to users / automatic scripts
  - Transfer of records to patients / tumours, coded by registration officers where appropriate

Workflow (2)

- When transferring electronic sources, records for a single patient are batched together, allowing a more complete view of the circumstances of diagnosis
  - Optimise for human context switches, and minimise page round trips
- To support the information gathering and QA, follow-up actions can be assigned to tumours
- At the end of a registration period, registerable tumours are staged, and flagged as “Final”

Data security

- A formal security audit identifies key security requirements
- The security of the code and the system is continuously monitored, and tested frequently
  - Separate code review for security (e.g. SQL injection attacks)
- Defense in depth: multiple overlaid security protections
  - Independent audits of database logins and data changes
- We provide extra security training before granting users external (web) access
  - Two-factor authentication

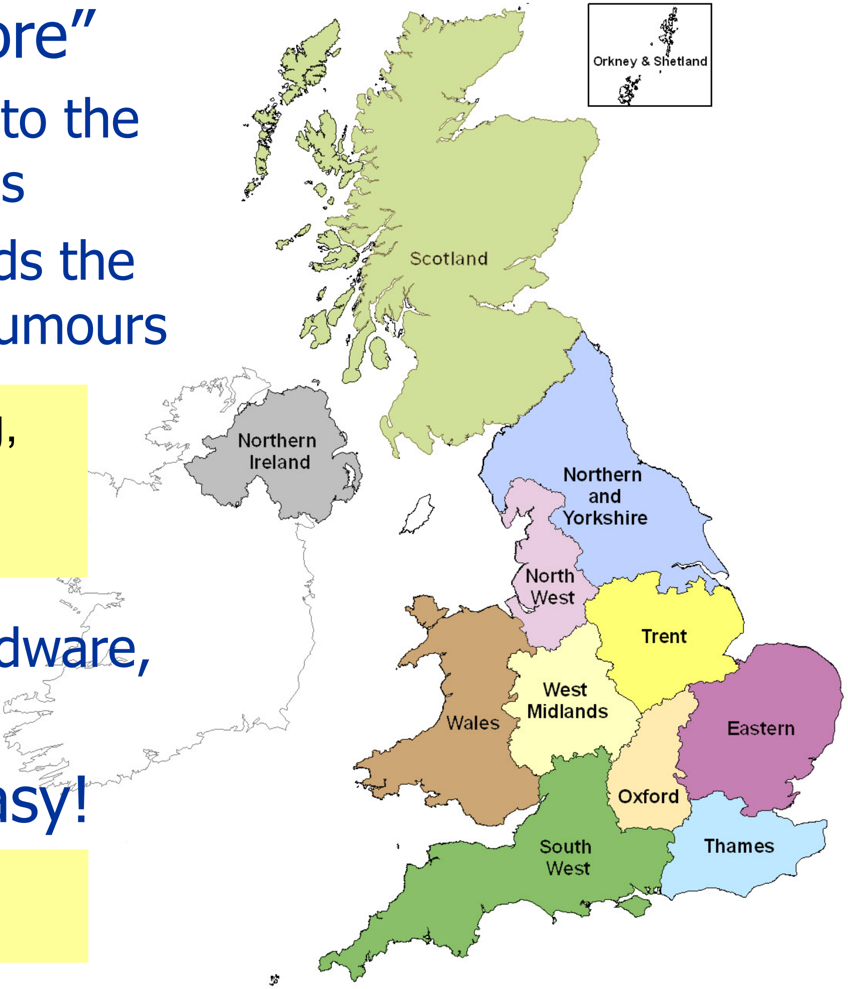
Cancer registry migration

- All 8 registries in England are migrating to a single shared system "encore"
  - Tumours are registered according to the patient's postcode at first diagnosis
  - A single centralised database avoids the need to exchange extra-regional tumours

Distributed systems are a Good Thing, but also hard – especially with an effective latency of months!

  - Other efficiencies, e.g. shared hardware, less duplicated development
- Migrating the data should be easy!

It's the easy bit!



Scale / scope of the task

- Scaling up a production system x 8, over two years
  - Without significant downtime [a few planned weekends]
- A web-based interactive cancer registration system
  - About 300 active users; about ½ use it full time
  - Ruby on Rails provides a rapid development environment
- Automated processing of electronic data sources
  - Pathology reports, PAS data, Death notifications, Multi-Disciplinary Team reports, Cancer Waiting Times, Hospital Episode Statistics, ...
  - Automate the routine, minimise human context switching
- Unified analysis platform
  - Simplify access to cancer data for researchers

It's never quite so easy

Scale / scope of the task

Continuous upgrade is hard

- Scaling up a production system x 8, over two years

It's a production system

  - Without significant downtime [a few planned weekends]

Can't defer all other changes

- A web-based interactive cancer registration system

More timely tumour registration

  - About 300 active users; about ½ use it full time
  - Ruby on Rails provides a rapid development environment

Rails 3.1, 3.2, …

ICD-O-3 (2011 update)

IE 7 (!)

- Automated processing of electronic data sources

IE 6 (?!)

  - Pathology reports, PAS data, Death notifications, Multi-Disciplinary Team reports, Cancer Waiting Times, Hospital Episode Statistics, …

Imaging

Radiotherapy

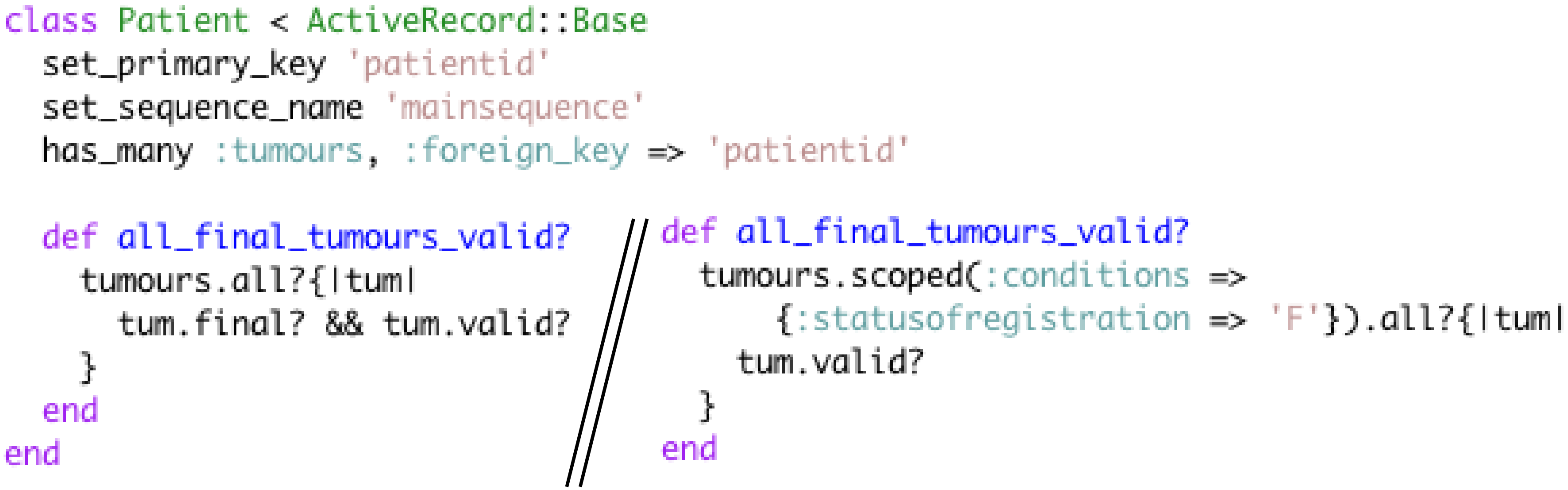
  - Automate the routine, minimise human context switching
- Unified analysis platform
  - Simplify access to cancer data for researchers

Different coding systems => different ways of counting

“Low grade endometrial stromal sarcomas are behaviour 3 in ICD-O-3, but behaviour 1 in ICD-O-2; should they be included in our count of all xnmsc?”

But will it scale?

- Essentially, it's an append-only dataset
  - Larger data blobs (e.g. pathology reports) are usually added and then never changed
  - Tumours are seldom updated (few more than 10 times)
  - 40GB for one registry (ECRIC) => 1TB should fit 8 registries
- Agile ≈ Lazy development *(i.e. just in time)*
  - Especially with the help of a nice ORM framework



```
class Patient < ActiveRecord::Base
  set_primary_key 'patientid'
  set_sequence_name 'mainsequence'
  has_many :tumours, :foreign_key => 'patientid'

  def all_final_tumours_valid?
    tumours.all?{|tum|
      tum.final? && tum.valid?
    }
  end
end
```

```
def all_final_tumours_valid?
  tumours.scoped(:conditions =>
      {:statusofregistration => 'F'}).all?{|tum|
    tum.valid?
  }
end
```

Conclusion

- Real data is full of exceptions
  - Most sane validations will have occasional, genuine exceptions
    - Tumour diagnosis date after date of death
    - Different patients with the same NHS number

Reality is like this, and recorded data more so.

- Data migration
  - Common core fields may have surprisingly different interpretations. New fields are actually easier.
  - The current owners / custodiens of the data can be responsible for the schema transformation.

Or you could bring in some management consultants, and blame every future glitch on them.

  - Avoiding the second system problem is essential.
- Future directions