Pseudonymisation at source

"preserving patient confidentiality
& public trust in doctors"

Julia Hippisley-Cox
25 Oct 2013
PHSCG, Stratford

My roles & acknowlegements

1. NHS GP
2. Professor GP/clinical epidemiology
3. New Member PHCSG Committee
4. Member of EMIS National User Group
5. Member Confidentiality Advisory Group
6. Director ClinRisk Ltd (software company)
7. Co-Director QResearch (Notts/EMIS) who have funded this work for public benefit

Outline

- QResearch database
- Open Pseudonymiser & data linkage

QResearch database
www.qresearch.org

- Established 2002 joint venture EMIS & UoN
- Patient level pseudonymised data
- Only used for research
- No patient identifiers, no free text
- Strong IG framework with no breaches
- Approved by ethics, BMA/RCGP
- Advisory board with NUG & practice reps
- Currently 680 practices; 15 million patients

Information on QResearch – GP derived data

- Demographic data – age, sex, ethnicity, SHA, deprivation
- Diagnoses
- Clinical values –blood pressure, BMI etc
- Laboratory tests – FBC, U&E, LFTs etc
- Prescribed medication – drug, dose, duration, frequency, route
- Referrals
- Consultations

QPrediction Scores
A new family of Risk Prediction tools

- Individual assessment
  - ➢Who is most at risk of preventable disease?
  - ➢What is level of that risk and how does it compare?
  - ➢Who is likely to benefit from interventions?
  - ➢What is the balance of risks and benefits for my patient?
  - ➢Enable informed consent and shared decisions
- Population level
  - ➢Risk stratification
  - ➢Identification of rank ordered list of patients for recall or reassurance
- GP systems integration

Published & validated scores

scores	outcome	Web link
QRISK2	***CVD***	***www.qrisk.org***
QDiabetes	Type 2 diabetes	**www.qdiabetes.org**
QStroke	***Ischaemia stroke***	***www.qstroke.org***
QKidney	Moderate/severe renal failure	**www.qkidney.org**
QThrombosis	***VTE***	***www.qthrombosis.org***
QFracture	Osteoporotic fracture	**www.qfracture.org**
QIntervention	***Risks benefits interventions to lower CVD and diabetes risk***	***www.qintervention.org***
QCancer	Detection common cancers	**www.qcancer.org**
QAdmissions	***Emergency admissions***	***www.qadmissions.org***

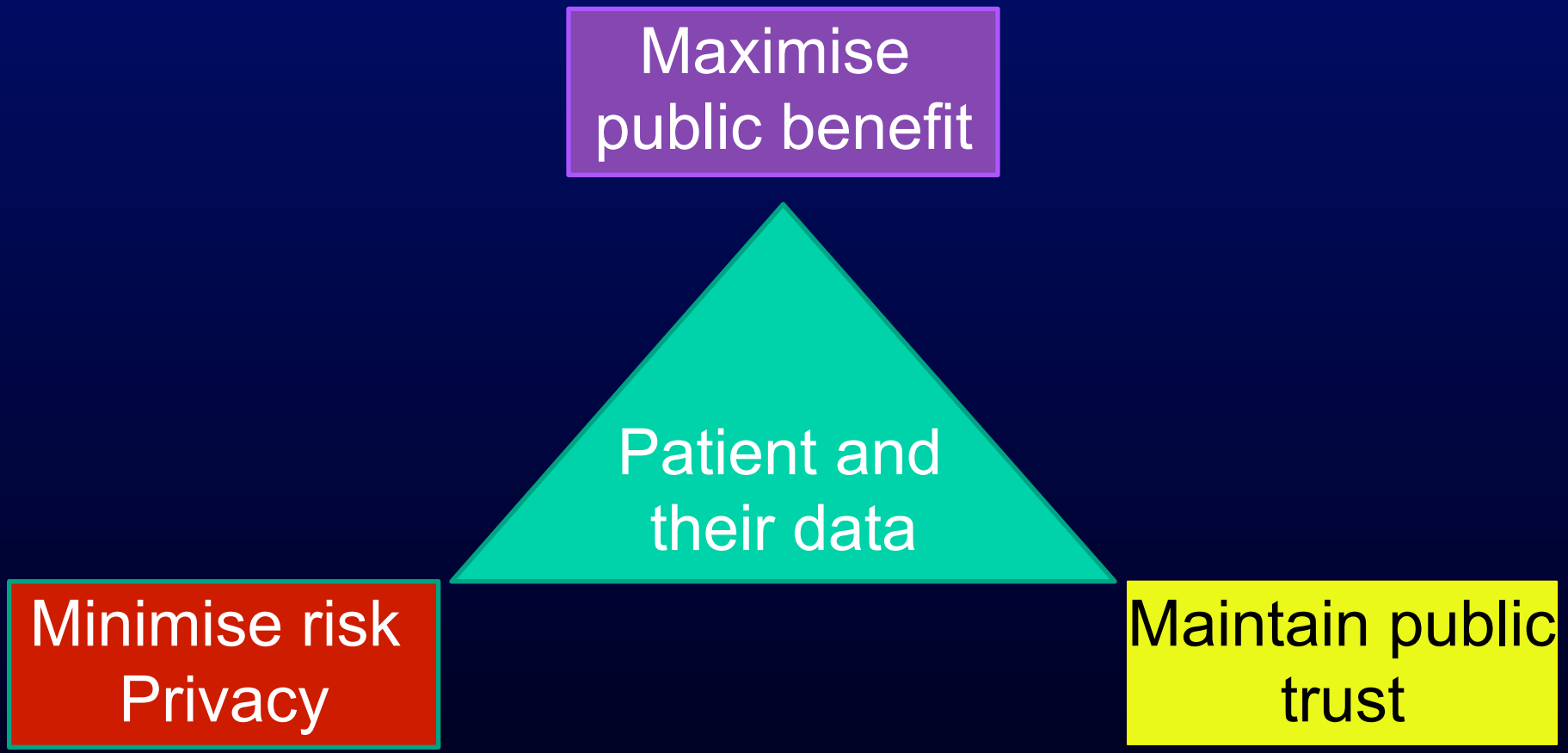
QResearch Data Linkage Project

- QResearch database already linked to
  - deprivation data in 2002
  - cause of death data in 2007
- Very useful for research
  - better definition & capture of outcomes
  - Improved performance of QRISK2 and similar scores
- Wanted to extend to linkages to
  - HES
  - Cancer registries
  - Other datasets
- Needed new open source technique for data linkage using pseudonymised data

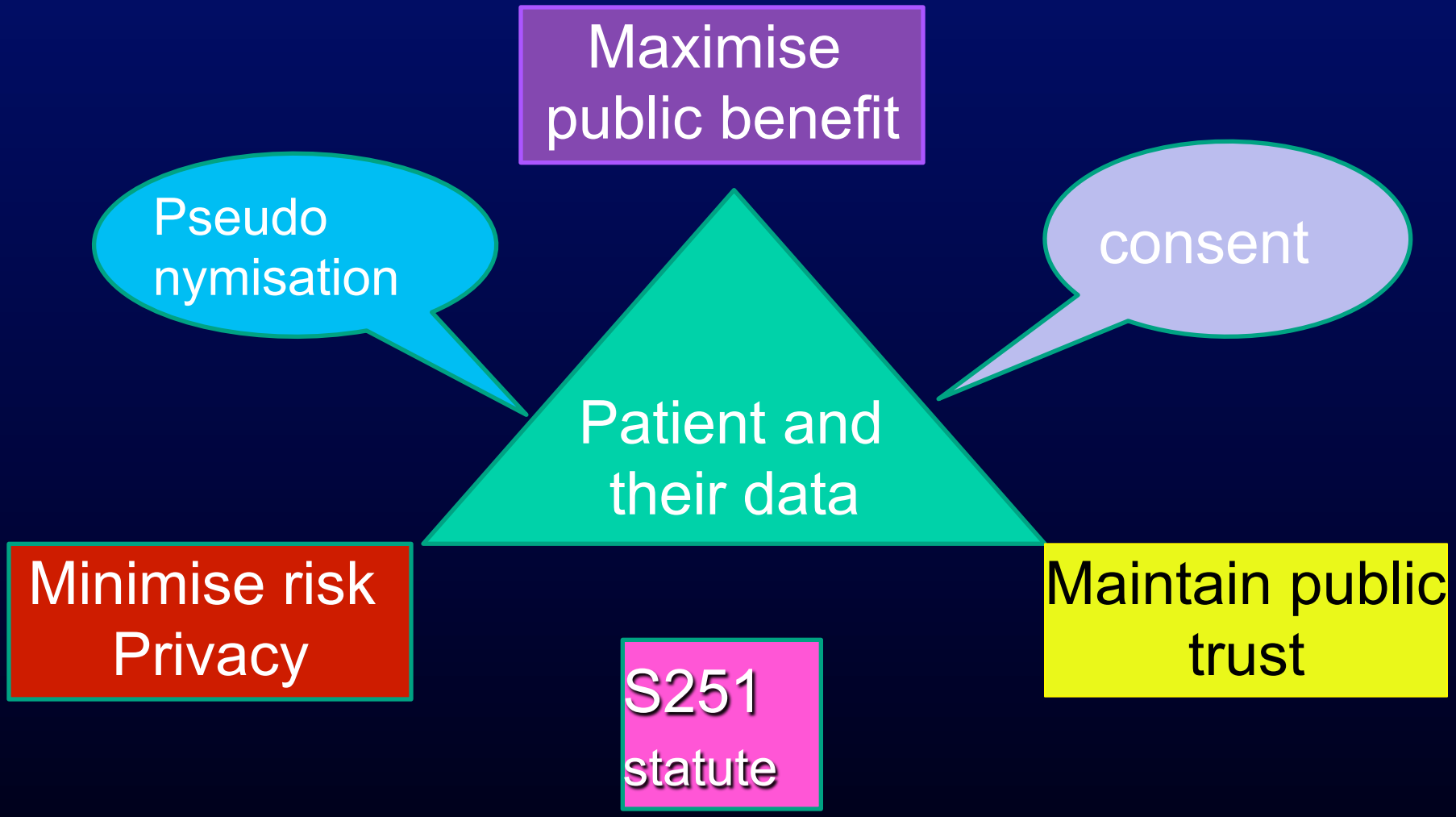
Objectives

- Open common technical approach for pseudonymisation
- allows individual record linkage BETWEEN organisations
- WITHOUT disclosure strong identifiers
- Inter-operability
- Voluntary ‘industry’ specification
- One of many approaches

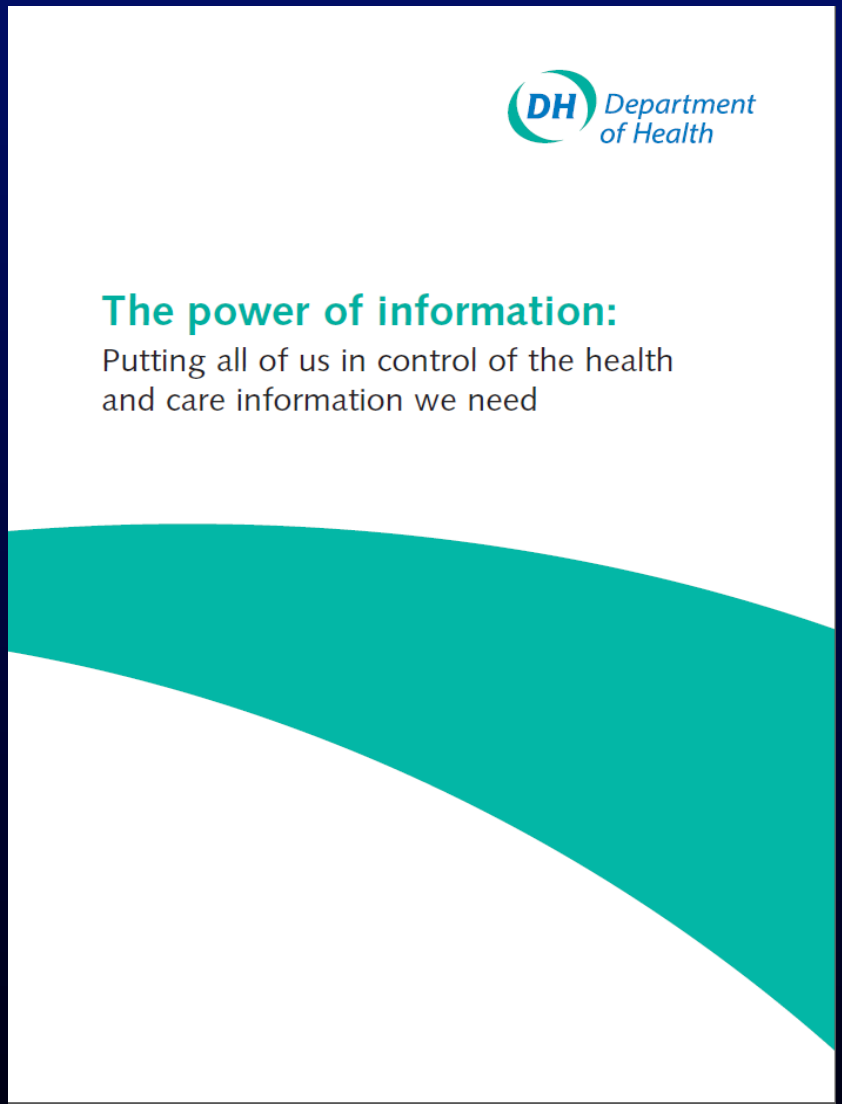
Key objectives for safe data sharing



Three main options for data access



Policy context



- Transparency Agenda
- Open Data
- Caldicott2
- Benefits of linkage for (in order from document)
  - Industry
  - Research
  - commissioners
  - Patients
  - service users
  - public

Attendances at 3 workshops

- East London CSUs
- GP suppliers – TPP, EMIS, INPS, iSoft
- NHS England, HSCIC, ISB, ONS, DH
- Cerner, CSC
- CPRD, THIN, ResearchOne, IMS
- PHCSG, BMA, RCGP, GP system user groups, Various universities
- Some pseud companies (Oka Bi, Sapior etc)

Ground rules: all outputs from workshop

- Published
- Open
- Freely available
- Can be adapted & developed
- Complement existing approaches

Big Data or Big Headache

- Need to protect patient confidentiality
- Maintain public trust
- Data protection
- Freedom of Information
- Information Governance
- ‘safe de-identified format’

Assumptions

- Pseudonymisation is desired "end state" for data sharing for purposes other than direct care
- Legitimate use of data
  - legitimate purpose
  - legitimate applicant or organisation
- Ethics and governance approval in place
- Appropriate data sharing agreements

Working definition of pseudonymisation

- Technical process applied to identifiers which replaces them with pseudonyms
- Enables us to distinguish between individual without enabling that individual identified
- Either reversible or irreversible
- Part of de-identification

Identifiable information

- person identifier that could identify a person. Examples include:
  - Name
  - Address
  - Dob
  - Postcode
  - NHS number
  - telephone no
  - Email
  - (local GP practice or trust number)

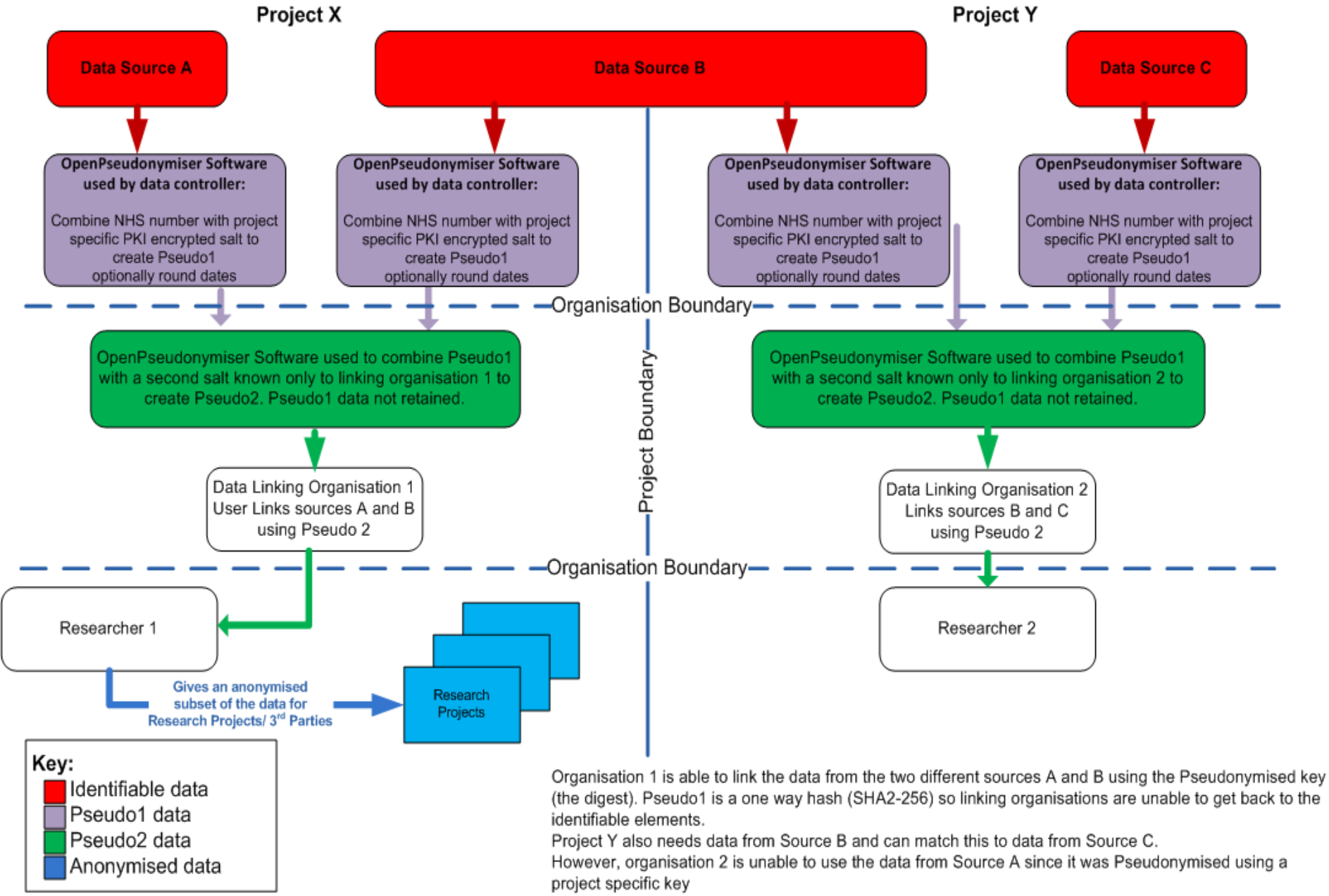
Open pseudonymiser approach

- Need approach which doesn't extract identifiable data but still allows linkage
  - Legal ethical and NIGB approvals
  - Secure, Scalable
  - Reliable, Affordable
  - Generates ID which are Unique to project
  - Can be used by any set of organisations wishing to share data
  - Pseudonymisation applied as close as possible to identifiable data ie within clinical systems

Pseudonymisation: method

- Scrambles NHS number BEFORE extraction from clinical system
  - Takes NHS number + project specific encrypted ‘salt code’
  - One way hashing algorithm (SHA2-256) – no collisions and US standard from 2010
  - Applied twice - before leaving clinical system & on receipt by next organisation
- Apply identical software to second dataset
- Allows two pseudonymised datasets to be linked
- Cant be reversed engineered

Data Flow Diagram illustrating the OpenPseudonymiser process to enable sharing and linkage of two discrete data sources



Organisation 1 is able to link the data from the two different sources A and B using the Pseudonymised key (the digest). Pseudo1 is a one way hash (SHA2-256) so linking organisations are unable to get back to the identifiable elements.
Project Y also needs data from Source B and can match this to data from Source C.
However, organisation 2 is unable to use the data from Source A since it was Pseudonymised using a project specific key

Web tool to create encrypted salt: proof of concept

- Web site private key used to encrypt user defined project specific salt
- Encrypted salt distributed to relevant data supplier with identifiable data
- Public key in supplier's software to decrypt salt at run time and concatenate to NHS number (or equivalent)
- Hash then applied
- Resulting ID then unique to patient within project

www.openpseudonymiser.org

- Free website with tools & utilities for implementation, evaluation and testing
  - Desktop application
  - DLL for integration
  - Test data
  - Documentation
  - Utility to generate encrypted salt codes
  - Source code GNU LGPL
  - Some support

Current implementations

- EMIS – 56% of GP practices
- TPP – 20% GP practices
- Cerner
- Office National Statistics
- HSCIC
- ISoft/CSC (in progress)
- 111 service (in progress)
- CSU's (in progress)
- United Health (in progress)

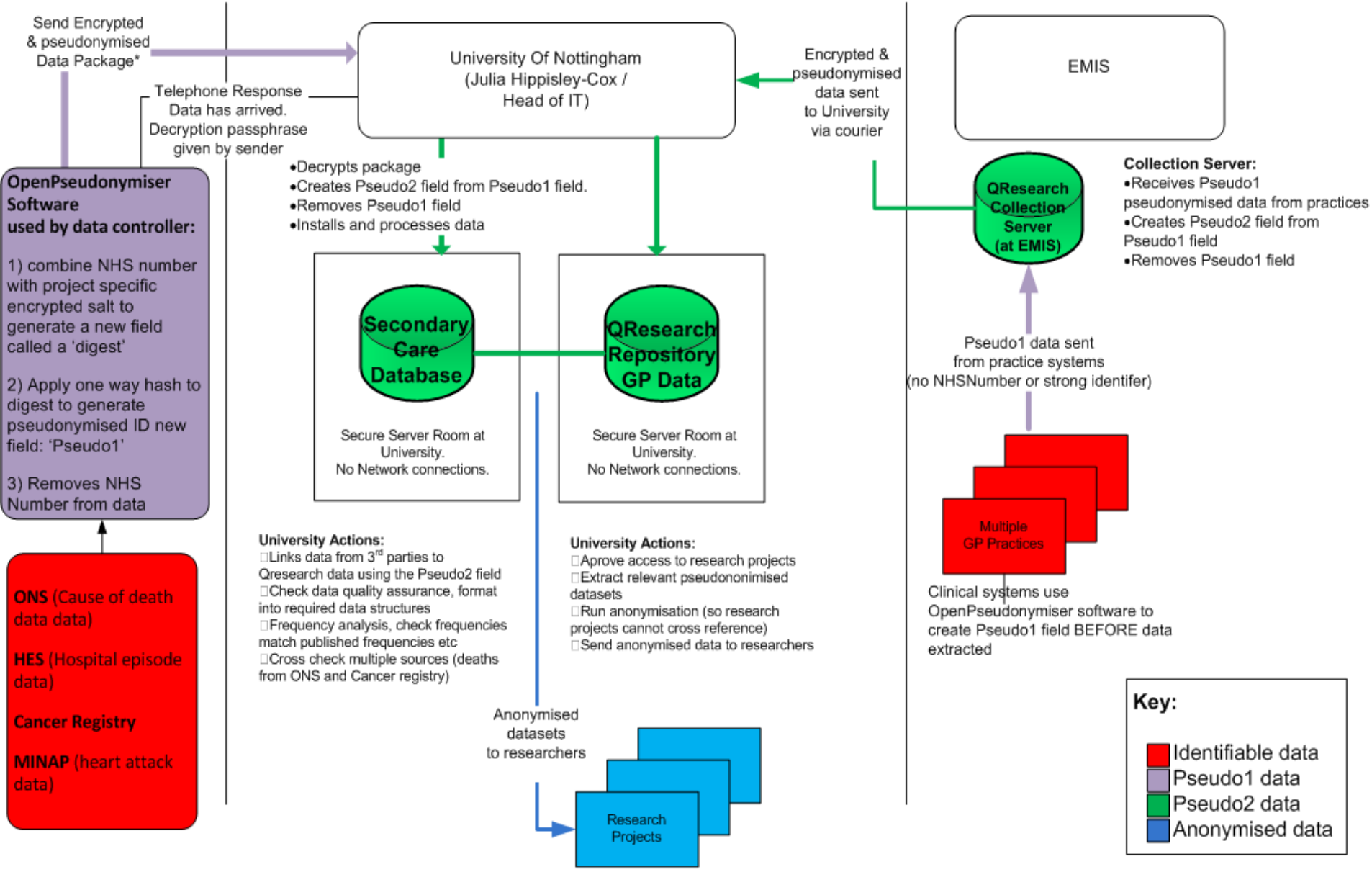
ResearchOne Pseudonymisation

- OpenPseudonymiser is the standard for ResearchOne pseudonymisation
- Runs on the database via the open source script
- SystmOne ID → automatic pseudonymisation → ResearchOne ID
- ResearchOne ID + project-specific salt → Project ID

Qresearch data linkage projects

- Link HES, Cancer, deaths to QResearch
- NHS number complete and valid in > 99.7%
- Successfully applied OpenP
- - Information Centre
- - ONS cancer data
- - ONS mortality data
- - GP data (EMIS systems)
-

Data Flow Diagram illustrating the OpenPseudonymiser process to enable linkage secondary care data to the QResearch database



QAdmissions

- New risk stratification tool to identify risk emergency admission
- Modelled using GP-HES-ONS linked data
- Can apply to linked data or GP data only
- NHS number complete & valid 99.8%
- 97% of dead patient have matching ONS deaths record
- High concordance of year of birth, deprivation scores

Benefits pseudonymisation-at-source

- Better for patient confidentiality
- Better for practice and public confidence
- Better to enforce security in data than rely on trust
- No need for s251
- No need to handle Subject Access Requests
- Can retain data longer & hold more data.
- Don't need to handle opt outs and delete data from live systems backups

Additional benefits pseudonymisation-at-source

- Better longitudinal data from General Practice
- More complete
- Less likely to be mis-intepreted & misused
- Better for research
- Better for informing policy
- Likely to lead to better decisions
- More likely to increase public benefit
- No additional work\cost for GPs

Key points pseudonymisation-at-source

- Instead of extracting identifiers and storing lookup tables/keys centrally, then technology to generate key is stored within the clinical systems
- Use of project specific encrypted salted hash ensures secure sets of ID unique to project
- Full control of data controller
- Can work in addition to existing approaches
- Open source technology so transparent & free
- Practical
- Available

So key question is

- Why not pseudonymise-at-source when technology available and its better for patient confidentiality?