Databases and platforms for data analysis from NGS of MTB

Derrick Crook

MMM Consortium

UKCRC Modernising Medical Microbiology

UNIVERSITY OF OXFORD

Public Health England

MRC Medical Research Council

wellcome trust

Oxford Biomedical Research Centre
Enabling translational research through partnership

NHS
National Institute for Health Research

MMM Consortium

- Linking Clinical record systems and NHS databases
- Translating next generation sequencing for patient benefit
- Sequenced > 25,000 isolates:
  - 7,500 *C. difficile*
  - 7,500 *S. aureus*
  - **2,600 TB**
  - 2,500 *E. coli/Klebsiella spp*
  - 2000 Grp B Streptococcus
  - ~ 3000 other (including viruses)

Storing, searching and analysing the data locally

- Growing problems with managing sequence data and linking it to meta-data (quality statistics and organism/patient specific data)
- Obstacles to automated processing on a large and even national scale
  - Just SQL database storage system not suitable
  - Replaced with a non-SQL system using Casandra hosted on an "Amazon cloud like technology" (i.e. Eucalyptus) including Hadoop and MapReduce
  - Store minimal de-identified metadata on the genomics data-store and have architecture to link back to clinical records
- Plan to follow the model being proposed by Genome England (GeL) for the 100000 genomes project

Open access and sharing of data?

1. What data to deposit
2. When to deposit
3. Where to deposit

- How to use the publically accessible data
  - Some questions need little data (these may be the most important now)
  - Other questions need rich data (these are more important in the future)

- There is good progress on depositing next generation sequencing data through established portals

A group assembled by the FDA and DTU are doing what we need

About Global Microbial Identifier

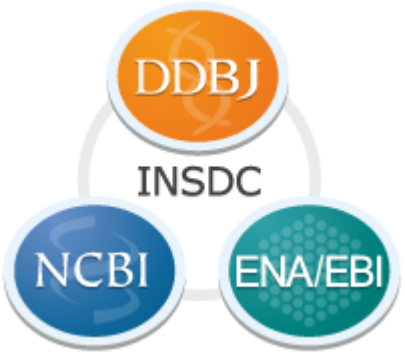
http://www.globalmicrobialidentifier.org

- The genomic epidemiological database for global identification of microorganisms or global identifier of microorganisms is a platform for storing whole genome sequencing (WGS) data of microorganisms, for the identification of relevant genes and for the comparison of genomes to detect outbreaks and emerging pathogens.
- The database holds two types of information: 1) genomic information of microorganisms, linked to, 2) metadata of those microorganism such as epidemiological details.

Three funded international archives

- GenBank ® NIH genetic sequence database (NCBI)
- ENA European Nucleotide Archive (EMBL – EBI)
- DDBJ DNA Data Bank of Japan

International Nucleotide Sequence Database Collaboration



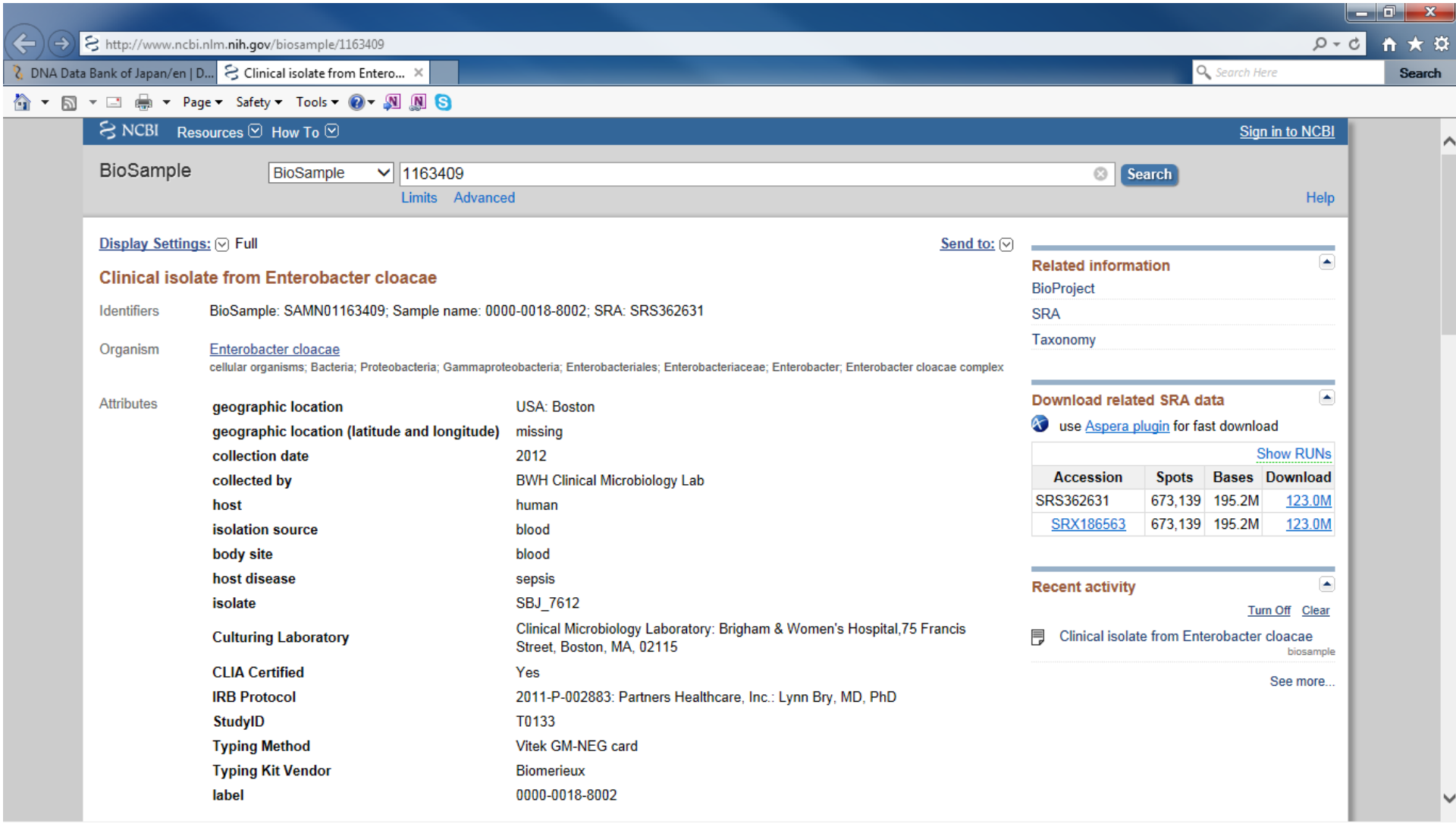
- NCBI is developing a automated uploader of data with meta-data; will be incorporated by ENA and DDBJ

NCBI prototypic schema

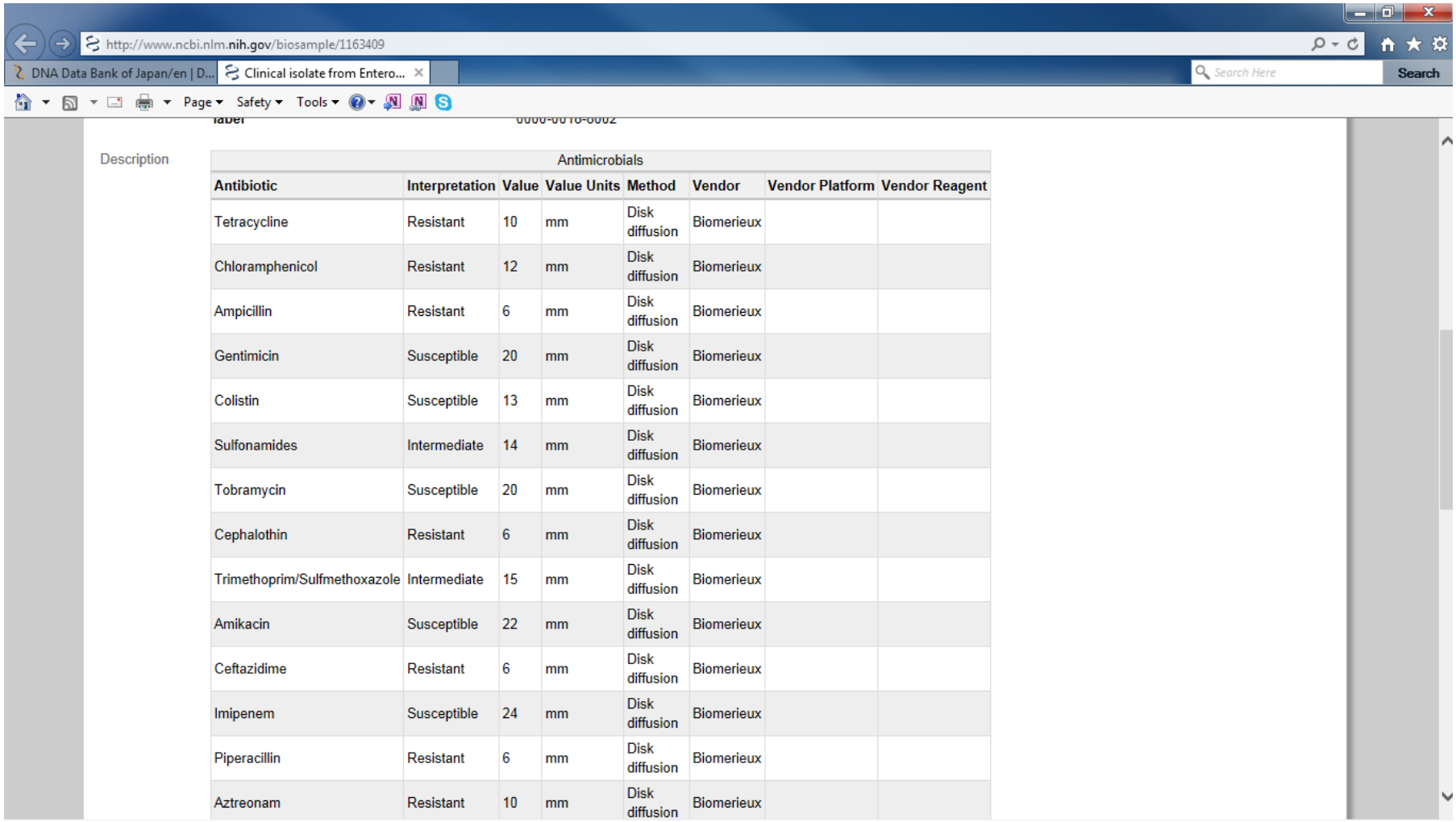
sample name	*unique ID for the sample*
attribute package	*Indicate the type of pathogen.* *Allowed values are "clinical or host-associated pathogen" or "environmental, food or other pathogen".* *Value provided in this field drives validation of other fields.*
organism	*scientific name of the organism that provided the sequenced genetic material- expect genus species*
strain	*strain/isolate from which sequence was obtained*
collection_date	*Date of sampling, in "DD-Mmm-YYYY", "Mmm-YYYY" or "YYYY" format (single instance, eg., 05-Oct-1990, Oct-1990 or 1990) or ISO 8601 standard "YYYY-mm-dd" or "YYYY-mm-ddThh:mm:ss" (eg. 1990-11-05 or 1990-11-05T14:41:36)*
collected-by	*Name of the person or lab who collected the sample.*
isolation-source	*Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived.*
geo_loc_name	*Geographical origin of the sample*
lat_lon	*Report values in decimal degrees and in WGS84 system*
specific_host	*Required for 'clinical or host-associated pathogen' sample type- Taxid or organism name of host*
host-disease	*Required for 'clinical or host-associated pathogen' sample type- Name of relevant disease, e.g. Salmonella gastroenteritis. Controlled vocabulary, http://bioportal.bioontology.org/ontologies/1009 or http://www.ncbi.nlm.nih.gov/mesh*

Bill Klimke is the contact person: Klimke, Bill (NIH/NLM/NCBI) [E] <klimke@ncbi.nlm.nih.gov>

NCBI prototypic schema



NCBI prototypic schema

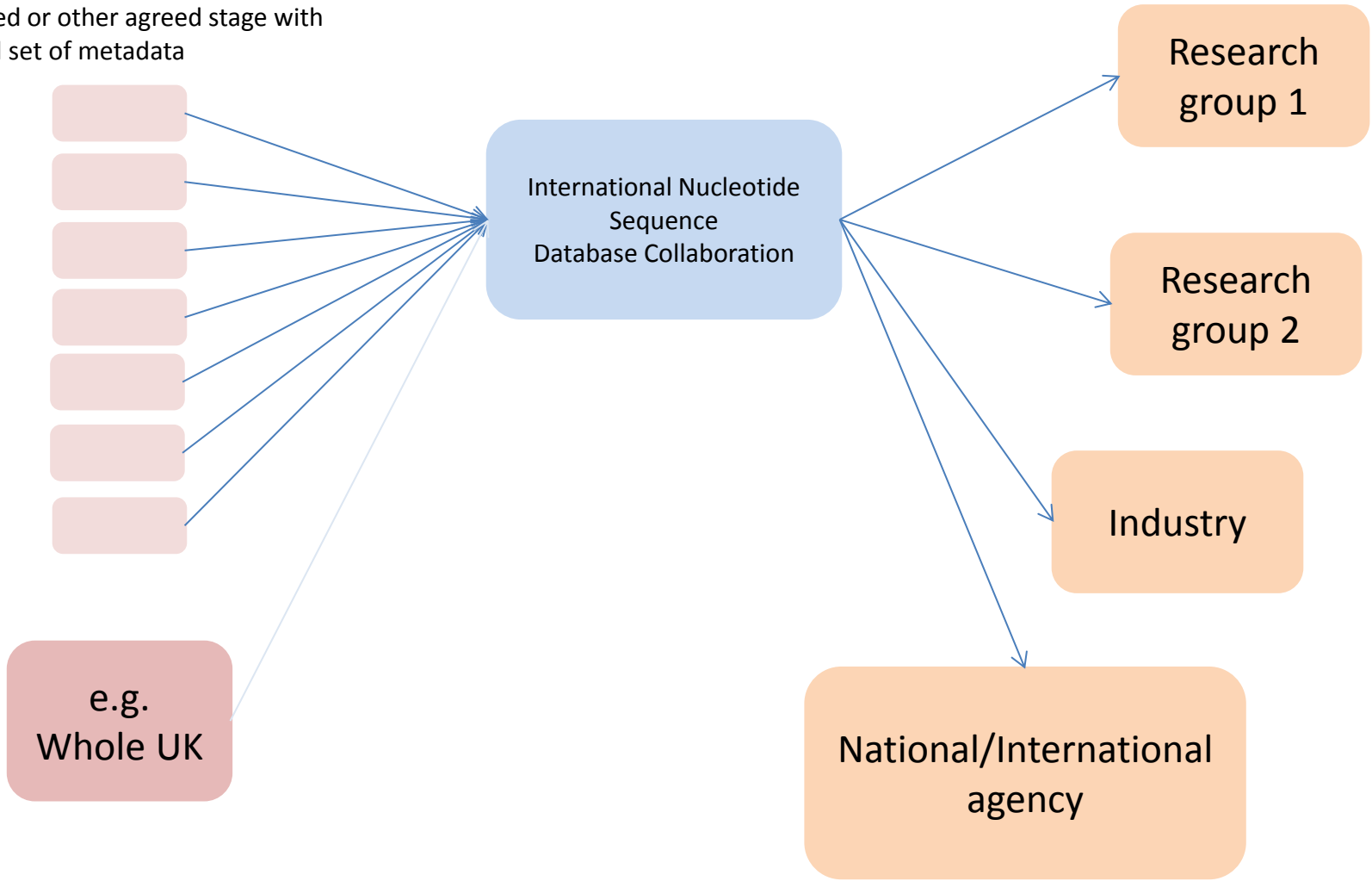


Antimicrobials

Antibiotic	Interpretation	Value	Value Units	Method	Vendor	Vendor Platform	Vendor Reagent
Tetracycline	Resistant	10	mm	Disk diffusion	Biomerieux		
Chloramphenicol	Resistant	12	mm	Disk diffusion	Biomerieux		
Ampicillin	Resistant	6	mm	Disk diffusion	Biomerieux		
Gentimicin	Susceptible	20	mm	Disk diffusion	Biomerieux		
Colistin	Susceptible	13	mm	Disk diffusion	Biomerieux		
Sulfonamides	Intermediate	14	mm	Disk diffusion	Biomerieux		
Tobramycin	Susceptible	20	mm	Disk diffusion	Biomerieux		
Cephalothin	Resistant	6	mm	Disk diffusion	Biomerieux		
Trimethoprim/Sulfmethoxazole	Intermediate	15	mm	Disk diffusion	Biomerieux		
Amikacin	Susceptible	22	mm	Disk diffusion	Biomerieux		
Ceftazidime	Resistant	6	mm	Disk diffusion	Biomerieux		
Imipenem	Susceptible	24	mm	Disk diffusion	Biomerieux		
Piperacillin	Resistant	6	mm	Disk diffusion	Biomerieux		
Aztreonam	Resistant	10	mm	Disk diffusion	Biomerieux		

One possible approach

Source data for upload to archive when published or other agreed stage with minimal set of metadata



Platforms for analysis

- Simple objectives e.g. :
  - Species identification
  - Resistance prediction
  - Relatedness (genomic match)

- Each needs a knowledge base

- Each will have a design, which will vary according to e.g. question, sequencing platform, method of assembly (mapped or *de novo*), method of querying the knowledge base etc.

Platforms for analysis

- These, at present, will be software needing high performance computing (assembly, statistical genetic and machine learning methodologies)

- Research vs Service

- Three key endeavours for success for public health
  - Resistance prediction with high sensitivity and specificity which is continuously updated
  - Rapid identification of transmission chains
  - Rapid cheap processing using light weight compute

More complex questions

- Strain specific factors determining disease manifestation e.g.
  - Latent vs active
  - Pulmonary vs other
  - Species adaptation (e.g. M. bovis)
  - Will need development of new statistical genetic methodologies
  - Clinical response
  - etc

- Analysing vast amounts of data i.e. many thousands or even millions of genomes
- This moves into the field of “Big Data”

Acknowledgements

PHE and Gel

- **Jim Davies** (Director informatics Gel)
- **Philip Monk** (lead for TB 100000 genomes project)
- **Grace Smith** (head TB reference lab Birmingham)

International advisor/s

- Stefan Niemann

Oxford University

Analysis:

- Zamin Iqbal
- Daniel Wilson
- David Clifton
- Sarah Walker
- Tim Walker
- Tim Peto

High performance computing:

- Jim Davies
- Charles Crichton

UKCRC Modernising Medical Microbiology

University of Oxford

Public Health England

MRC Medical Research Council

wellcome trust

Oxford Biomedical Research Centre
Enabling translational research through partnership

NHS National Institute for Health Research