A generic informatics platform which provides opportunities for collaboration across organisations and expansion to other research areas.
Why link databases?

• **Record linkage for Clinical purposes:**
  – Instant sharing of information across treatment centres (hospitals, GP’s, etc)

• **Research power:**
  – Increase the sample size
  – Increase the potential for research collaborations

• **Main Issues:**
  – Lack of a common identifier at national or even state level (unlike many countries)

• **Key Question for Linkage:**
  – What level of accuracy is acceptable for each type of linkage?
BioGrid - Privacy and Ethics:

- Required approval from all sites to use their identifying information for linkage purposes. This includes the information being copied to a central location so it can be “matched” against patients from other sites.

- Required a legal opinion that we could use the last 5 digits of the Medicare number as part of the linkage algorithm. There is a provision in the Medicare Act that the whole number cannot be used for identification purposes (a legacy of the Australia Card?)

- Data available in de-identified form only, and therefore only suitable for research.

- Health data always kept separate from identifying data.
Authorised researchers query the Federated Data Repository for analysis.

Public Data Sources

The BioGrid Model
Authorised researchers query the Federated Data Repository for analysis.

Institute-specific data loaded into institute-specific Local Research Repository nightly.

The BioGrid Model

- Hospital # A
  - Multiple Source Databases
  - ETL
  - LRR
  - VPN

- Hospital # B
  - Multiple Source Databases
  - ETL
  - LRR
  - VPN

De-identified Linked data

Federated Data Integrator

Unique Subject Identifier

Metadata Repository

Internet

SAS
- Queries
- Statistical analysis
- Reports

Public Data Sources
- GenBank
- UniProt
- LocusLink
- PubMed

Nightly upload of data To computer at the site
Institute-specific data loaded into institute-specific Local Research Repository nightly.

Authorised researchers query the Federated Data Repository for analysis.

No identified health data leaves the site.

The BioGrid Model
Authorised researchers query the Federated Data Repository for analysis.

Institute-specific data loaded into institute-specific Local Research Repository nightly.

Hospital # A
Multiple Source Databases

Hospital # B
Multiple Source Databases

No health data stored centrally

The BioGrid Model

Public Data Sources
- GenBank
- UniProt
- LocusLink
- PubMed

Internet

SAS
Queries
Statistical analysis Reports

The Bio21 Cluster
Melbourne Health
Institute-specific data loaded into institute-specific Local Research Repository nightly

Authorised researchers query the Federated Data Repository for analysis.

The BioGrid Model

- **Individual records linked**
  - 6 id fields sent along secure lines to secure computer
  - Secure computer generates the linkage key
  - Linkage key sent to site’s computer
  - Linkage key store is ‘encrypted’
  - Health data only available with key, not with identifiers

**Public Data Sources**
- GenBank
- UniProt
- PubMed
- LocusLink
- SAS
- Metadata Repository
- Internet

**Hospital # A**
- Multiple Source Databases
  - ETL
  - LRR
  - VPN

**Hospital # B**
- Multiple Source Databases
Authorised researchers query the Federated Data Repository for analysis.

Institute-specific data loaded into institute-specific Local Research Repository nightly.

**Hospital # A**
- Multiple Source Databases
- ETL
- LRR
- VPN
- Metadata Repository

**Hospital # B**
- Multiple Source Databases
- ETL
- LRR
- VPN

**Federated Data Integrator**
- Unique Subject Identifier

**RESEARCH**
- Authorised access only
- De-identified data only sent
- No health data stored centrally

**SAS**
- Queries
- Statistical Analysis
- Reports

**Public Data Sources**
- GenBank
- UniProt
- LocusLink
- PubMed

**The BioGrid Model**
Authorised researchers and applications query the Federated Data Repository for analysis.

Data cached on computers at the sites - owned and controlled by site.
**Scope - Clinical Datasets**

**Cancer:**
- Colorectal
- Brain
- Breast
- Lung
- Sarcoma
- Gynaecology
- Prostate
- Head & Neck
- Upper GI
- Melanoma
- Renal
- Prostate
- Rare

**Neuroscience:**
- Epilepsy
- MS
- Stroke
- Neuropsychiatry

**Diabetes:**
- Type 1
- Type 2
- Gestational

**Other:**
- Crohns
- Cystic Fibrosis
- Bone Density
- Well Women’s
- MRI and PET

**BioGrid**

De-identified data

Researcher
Types of Linkage:

- Probabilistic
- Hashing (exact)
Probabilistic:

- Requires the two sets of identifiers to be brought together for comparison
- Weights are assigned to each identifier
- Often includes options for soundex, transposition detection, exclusion of dummy patients ("Babe 1")
- Final result is based on total of weighted matched fields, minus unmatched fields
- User must define a "threshold value" that determines a match
Linkage using eIndex
Operation of eIndex (1)

- Patient matches in BioGrid are determined by using these basic demographic fields:
  - Surname
  - Given name
  - Middle name / initial
  - Date of Birth
  - Gender
  - Digits 5 to 9 of the Medicare Number (Note that Medicare legislation does not allow the use of the full Medicare Number for identification purposes. However this partial use has been approved by the BioGrid Ethics Committee and legal advice.)
The matching process includes logic to handle many common types of errors, e.g.

- Names, dates and numbers which differ in only one position or in the transposition of two adjacent characters are given a weight for a possible match.

- Names are also checked for “soundex” matches and given a positive weight if this is the case.

- “Dummy” names are ignored. These include “UNKNOWN” and commonly used hospital names such as “TWIN 1”, “TRIPLET 1”, “BABE” and so on.
Operation of eIndex (3)
Hashing ("exact"):

- A non-reversible hash is generated at each site and these are brought together for comparison.

- Therefore does not require the 2 sets of identifiers to be brought together. This can be very important if the ethics or legal requirements preclude identifiers leaving both sites.

- In a simple hash system, the match must be exact – all or nothing. This makes them almost useless in real-life situations.
Linkage using GRHANITE™ (1)
Some data has additional encryption because it has been deemed sensitive. Date of Birth has been extracted here, but can only be decrypted for research use if a specific study requires age-related information.

Linkage using GRHANITE™ (2)

This form of encryption is non-reversible and heavily protected against dictionary attack.
“Hashing” creates a very long character string based on the incoming identifiers. The BioGrid-Grhanite system uses the SHA-256 algorithm, which cannot be reversed to reveal the original identifiers.

The hash values are spread across the possible output values in a pseudo-random fashion. This means that incoming names which are close together (e.g. SMITH and SMITHE) will have hash values that are random and not necessarily close to each other.

Grhanite uses not one but multiple hashes. This includes hashes based on:
- The full set of identifiers
- The soundex version of the identifiers
- A version of the identifiers with digits in the date of birth transposed

Each hash is assigned a weight in the final score. The hashes can therefore be thought of as "pseudo-identifiers" and are matched against another patient’s hashes in the same probabilistic fashion as real identifiers would be.

This method catches matches even if there are minor data entry errors. In our testing, Grhanite has achieved error rates of less than 1% for both false positives and false negatives (tested on 45,000 BioGrid patients). This is comparable to the best commercially available probabilistic algorithms.
Matching Comparison

- There are about 45000 patient records in our files and about 35000 unique patients, meaning that there are about 10000 matches.

- About 100 “matches” in Grhanite that were not found in eIndex
- About 100 “matches” in eIndex that were not found by Grhanite
- “False positive” and “false negative” is not quite the right terminology, since neither system is 100% correct.

- In each case, he has done a visual (i.e. human) check of the discrepancies and found that both systems are wrong about 50% of the time in cases where it is possible to make a determination. This means that each system has an error rate of about 100 or 1%, but for different patients in each.
## Linkage software (1)

<table>
<thead>
<tr>
<th>Software</th>
<th>Price</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun eIndex</td>
<td>High</td>
<td>Includes fuzzy logic – various types</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Includes high and low threshold values and grey area</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Has tools to store manually decided matches / non-matches</td>
</tr>
<tr>
<td>Dataflux (SAS)</td>
<td>High?</td>
<td>Can also do data matching, including fuzzy logic – not clear how sophisticated this is</td>
</tr>
<tr>
<td>IBM Quality Stage</td>
<td>High</td>
<td>Can assign weights to fields</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not clear the extent of fuzzy logic or soundex – fuzzy logic example but no explanation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Seems to just assign records to groups but no mention of automatic assignment of a USI type number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Determines its own “discriminatory power” for each data field using frequency analysis</td>
</tr>
<tr>
<td>Winpure</td>
<td>Moderate</td>
<td>Basic (telephone numbers and emails) or advanced (names and addresses) – no mention of fuzzy logic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Does not seem to assign common numbers – looks like user must decide what to do with each one</td>
</tr>
</tbody>
</table>
## Linkage software (2)

<table>
<thead>
<tr>
<th>Software</th>
<th>Price</th>
<th>Description</th>
</tr>
</thead>
</table>
| Linkage Wiz       | From $550 to $4250 (USD) depending number of records to be linked | - Some data cleansing – mostly names, addresses and emails  
- User definable weights (weighted based on relative scarcity of the value, e.g. SMITH)  
- Has Soundex  
- No mention of fuzzy logic, transpositions, etc |
| Grhanite          | In development            | - No data cleansing – pure matching tool  
- Hashing → exact match, but does this on several combinations of variables to produce a pseudo-probabilistic match  
- Can use relative scarcity of values in surname, DOB, etc to modify weights |
| Febrl (from ANU)  | Free                      | - Sketchy details, but probabilistic matching - does include some fuzzy logic (Jaro, Winkler and other string distance operators)  
- Includes some data cleansing – e.g. standardisation of phone number format |
| The Link King     | Free                      | - But requires base SAS  
- Combination of probabilistic and deterministic algorithms  
- Handles Nicknames, soundex, approximate strings, transposed dates and digits in SSN  
- Looks quite good |
The Future for Record Linkage

The ultimate goal is to link in such a way that the information can be used for clinical purposes. As well as legal and ethical considerations, this also has practical difficulties.

Note that such a linkage will eventually be possible if the initiative by NEHTA to have a national plan for linkage of health data is to become a reality.
Phase 1 (pilot): funded by Victorian Government (STI) via Bio21 (A$1.6M)
- 5 hospitals + 2 Research Institutes
- 3 disease types: Oncology, Diabetes, Epilepsy

Phase 2: funded by Australian Government (DEST) via the University of Melbourne (A$4.4M)
- 7 hospitals + 2 Research institutes
- 4 diseases: Oncology, Neuroscience, Diabetes, Respiratory + Images
- Expires at the end of 2008

Phase 3: funded by Victorian Government DIIRD via the University of Melbourne (A$11M) for Victorian Anti-Cancer Council
- Numerous hospitals + 2 Research institutes
- Currently 4 diseases: Oncology, Neuroscience, Diabetes, Respiratory + Images