

BioGrid Australia

(formerly BioGrid)



BIOGRID
AUSTRALIA
Health through information

Record Linkage

A Victorian
Government
initiative

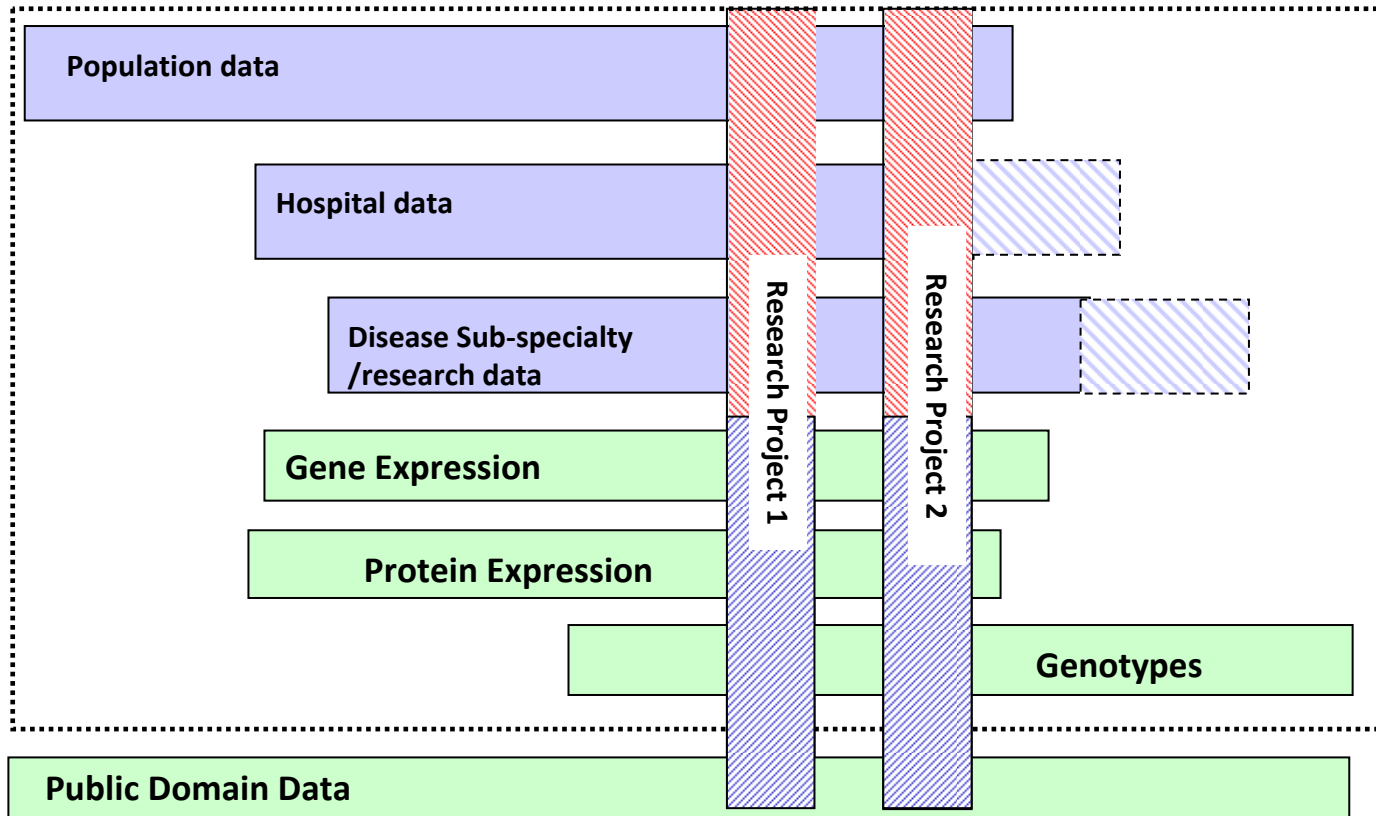
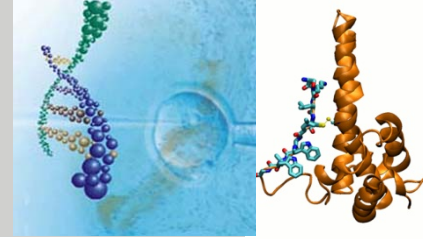


Australian Government

Department of Education, Science and Training

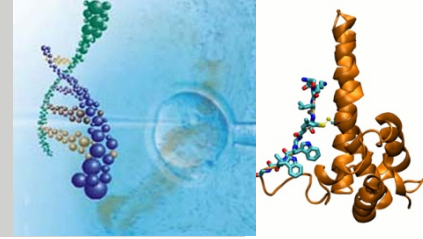


The Vision – remove the ‘silos’



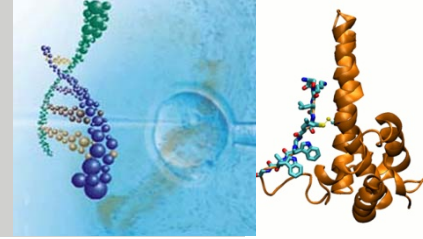
A generic informatics platform which provides opportunities for collaboration across organisations and expansion to other research areas.

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?

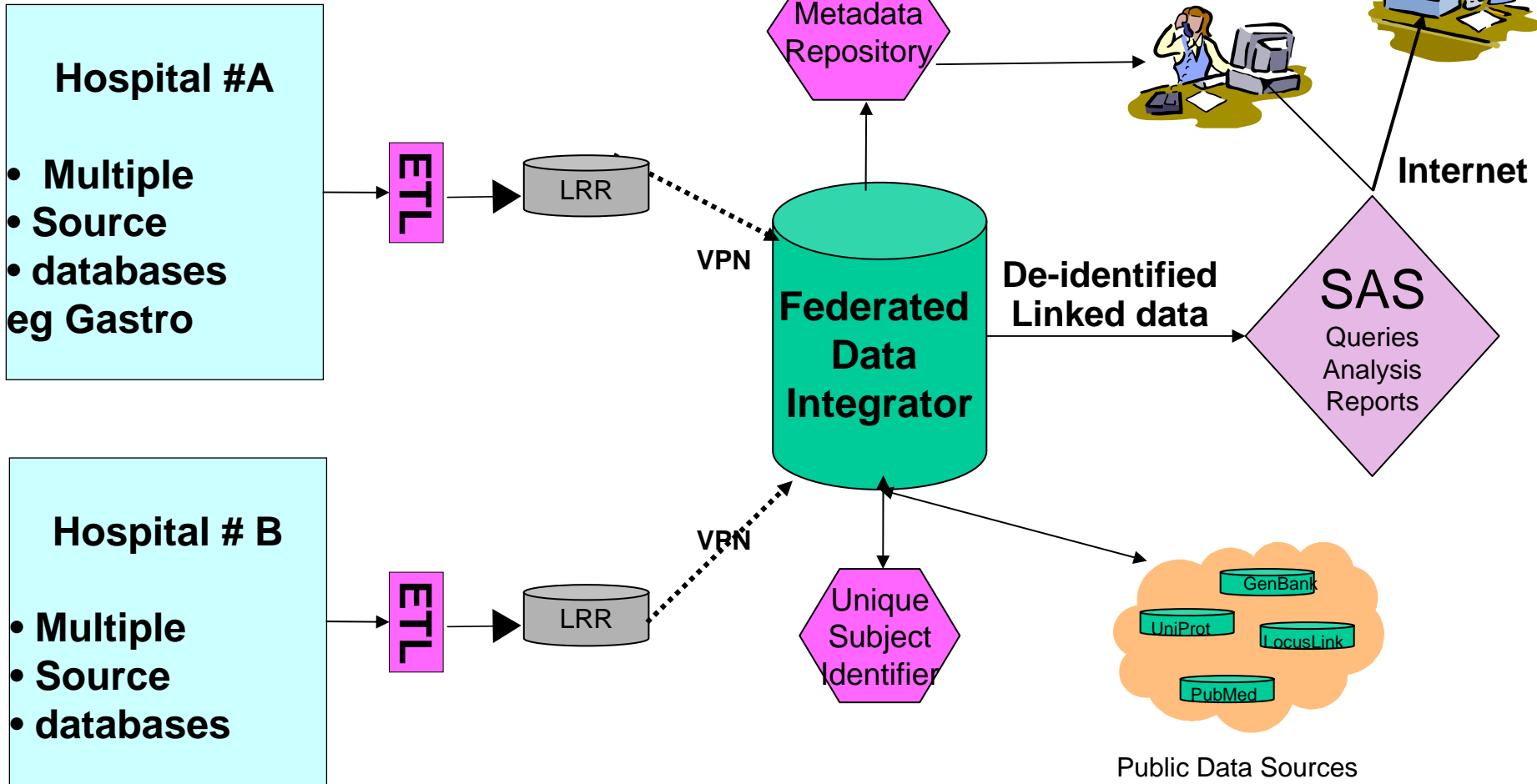
BioGrid - Privacy and Ethics:



- 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

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

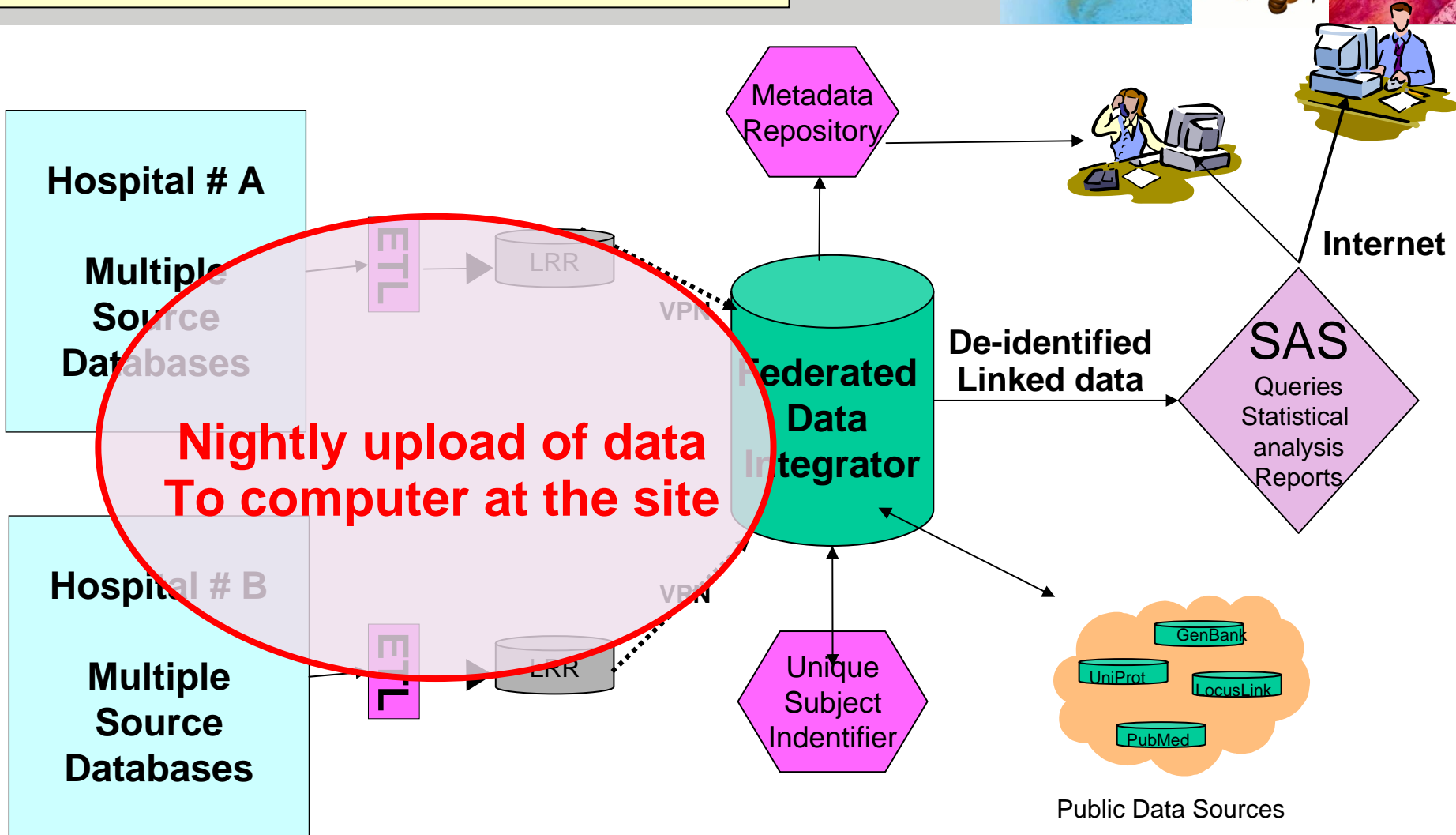
Authorised researchers query the Federated Data Repository for analysis.



The BioGrid Model

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

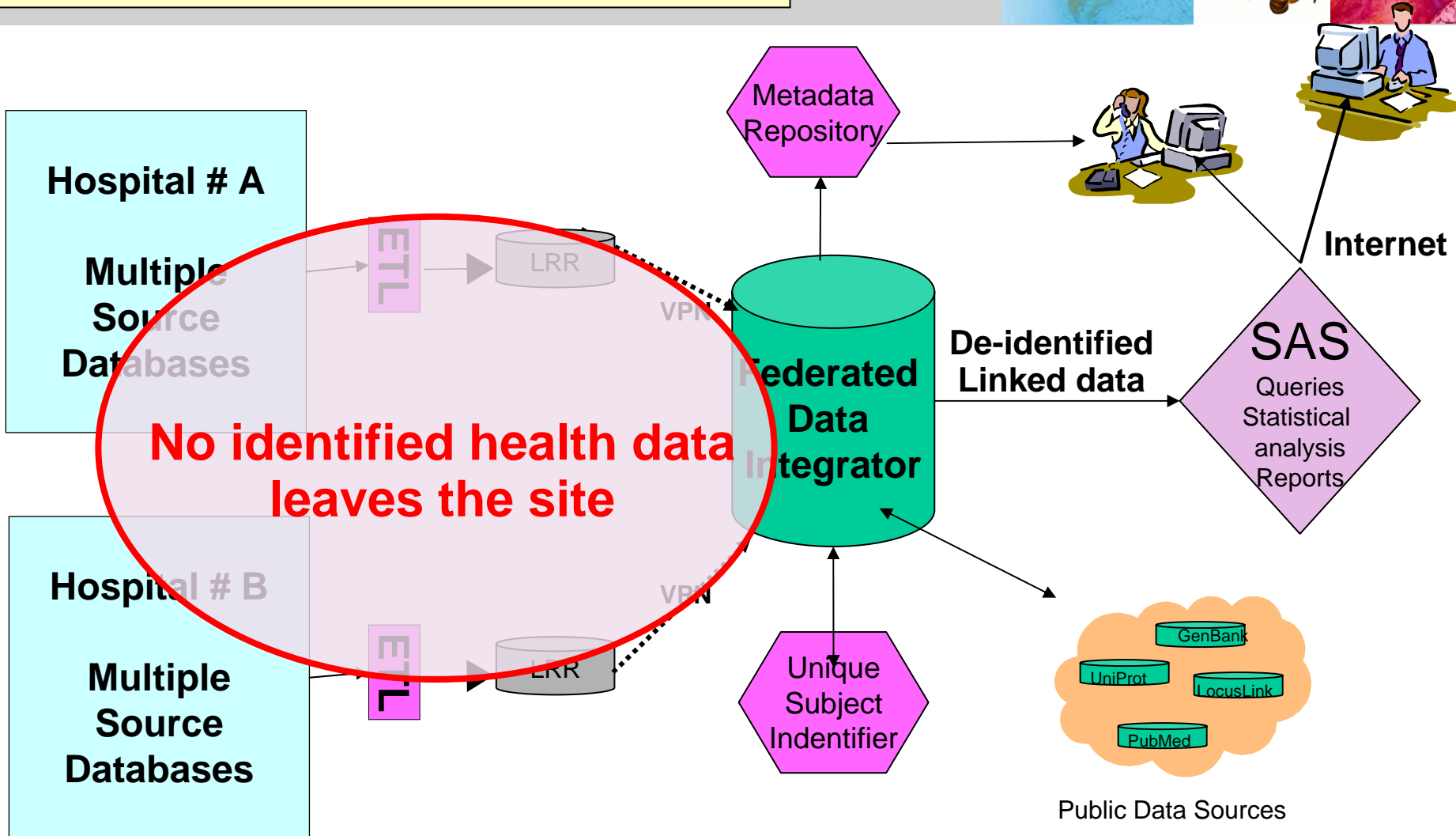
Authorised researchers query the Federated Data Repository for analysis.



The BioGrid Model

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

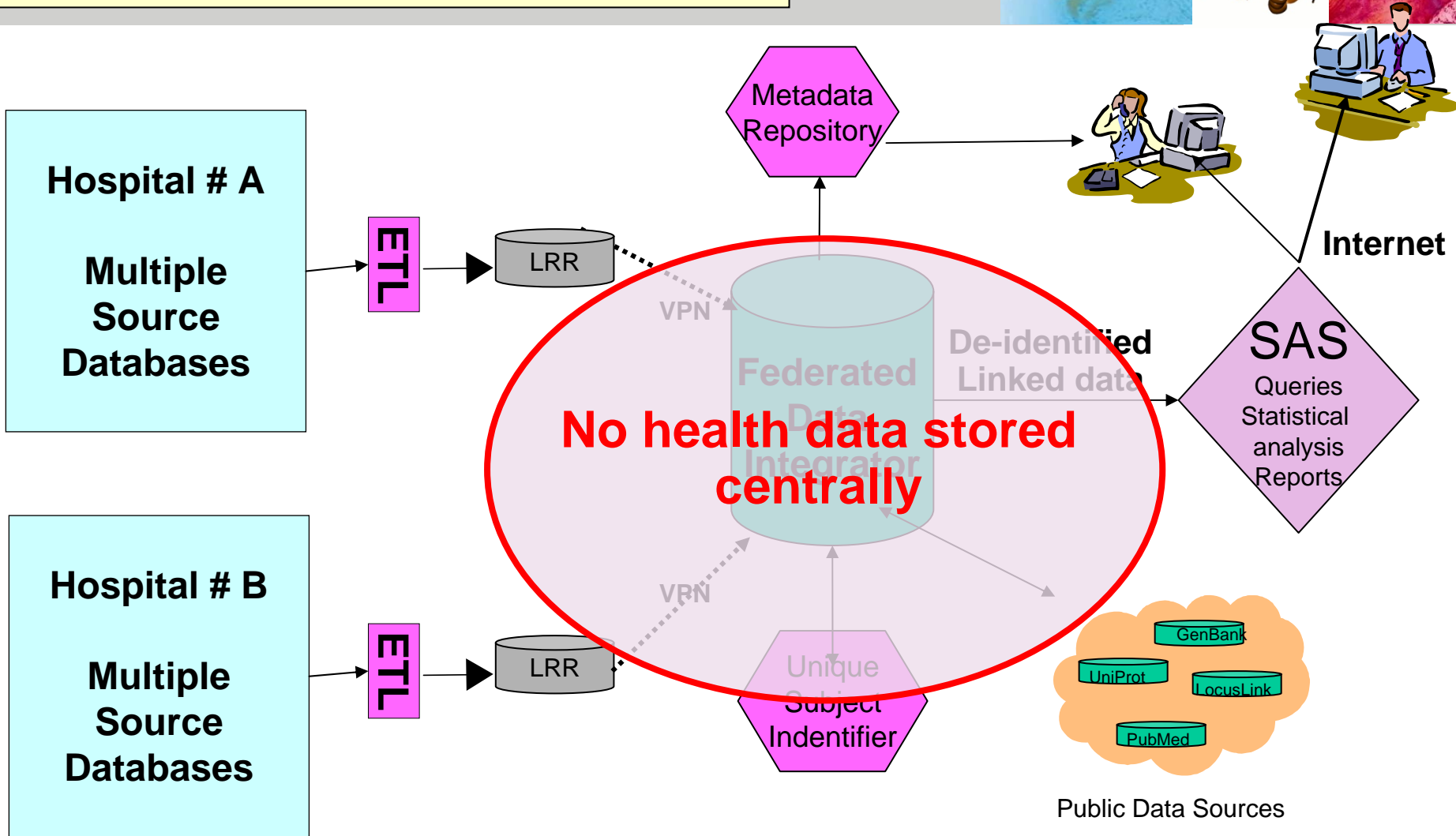
Authorised researchers query the Federated Data Repository for analysis.



The BioGrid Model

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

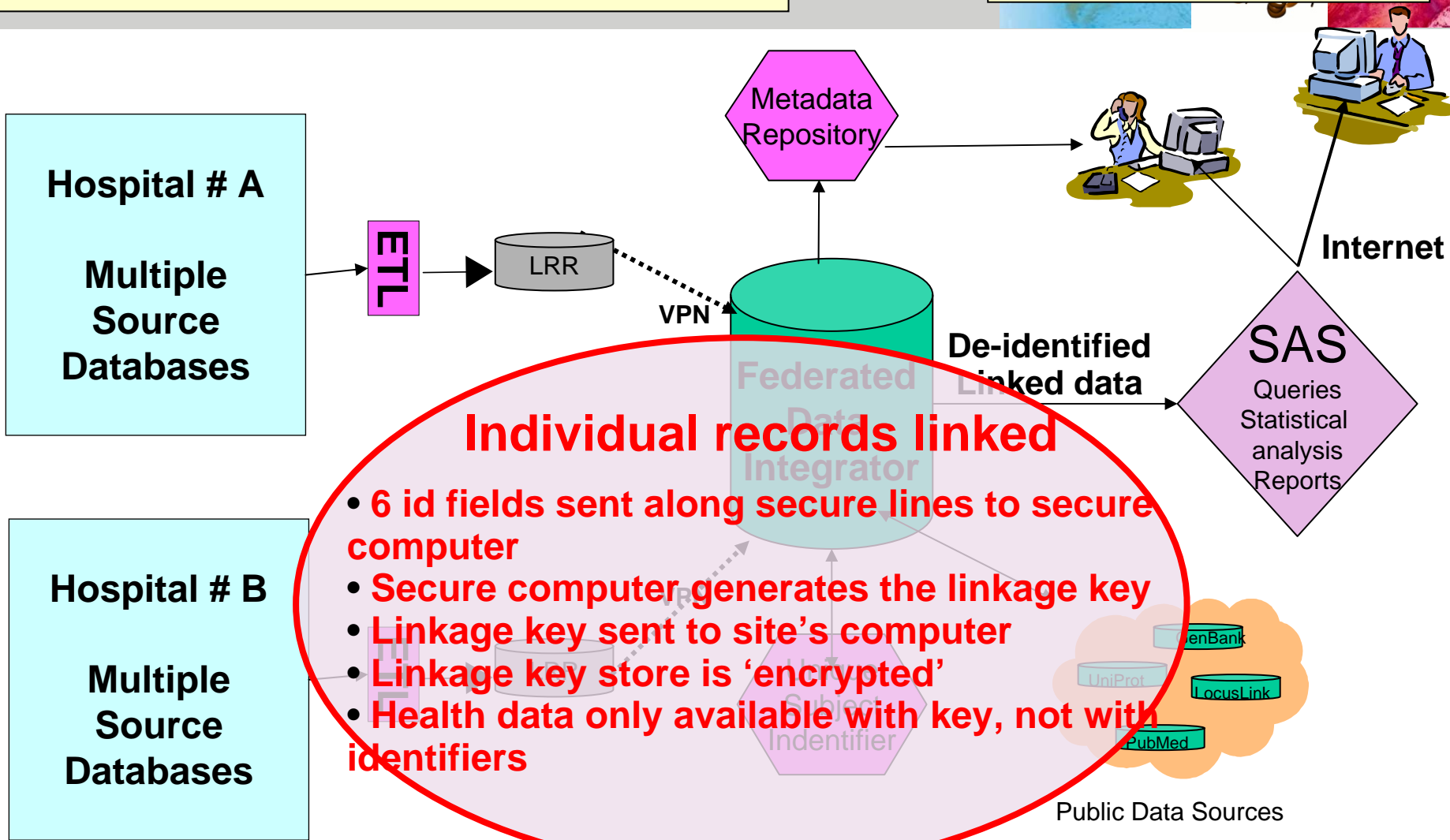
Authorised researchers query the Federated Data Repository for analysis.



The BioGrid Model

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

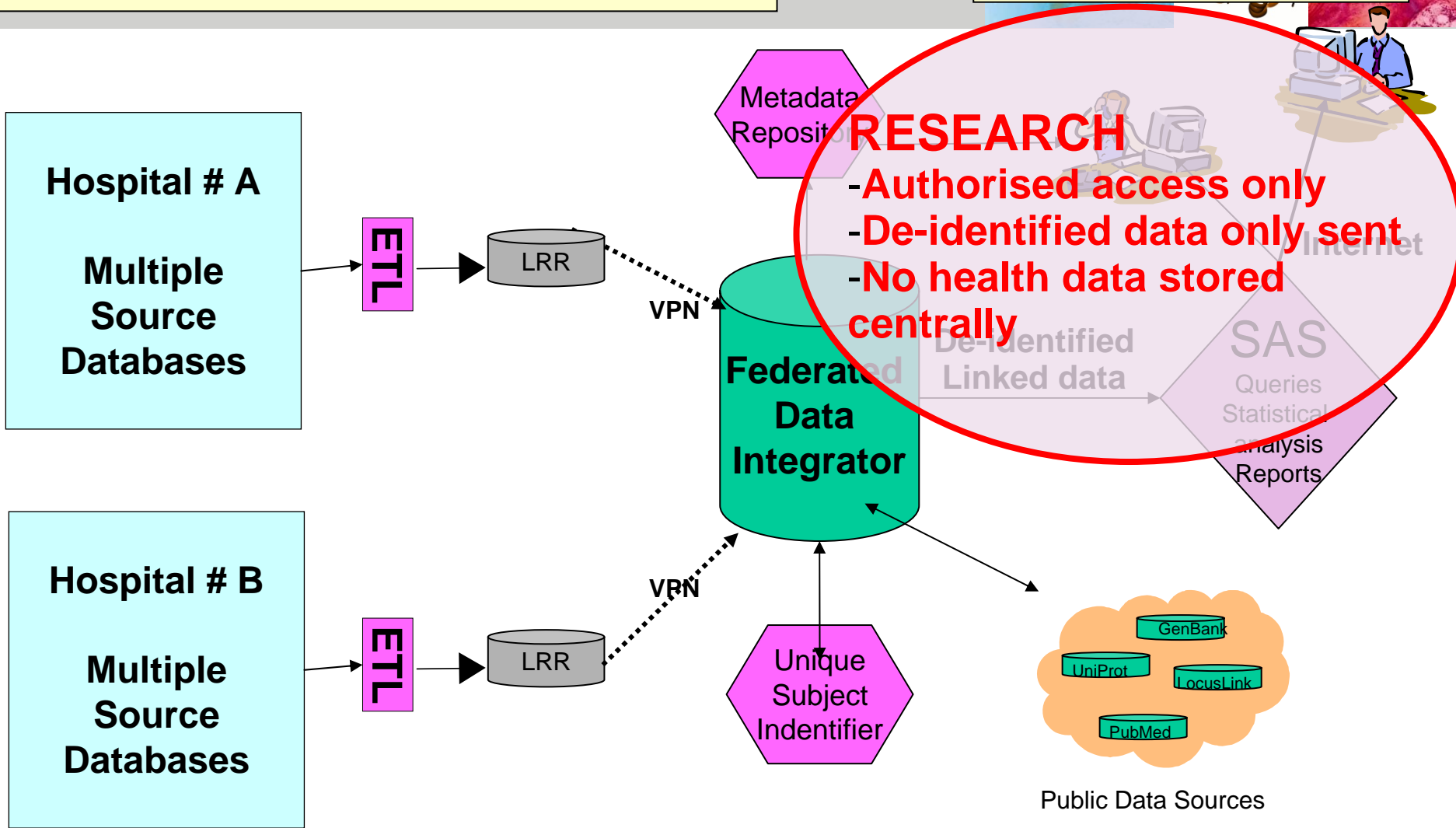
Authorised researchers query the Federated Data Repository for analysis.



The BioGrid Model

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

Authorised researchers query the Federated Data Repository for analysis.



The BioGrid Model

RCH/ MCRI

- Respiratory
- Diabetes
- Crohns

Monash MC

- Oncology
- Tissue Bank
- Cystic Fibrosis

Box Hill

- Oncology

St Vincents

- Neuroscience
- Diabetes
- Oncology

RWH

- Oncology
- Diabetes

PeterMac

- Oncology
- Tissues Bank
- PET images

Alfred

- Cystic Fibrosis
- Neuroscience
- Oncology

Austin

- Oncology
- Tissues Bank
- Diabetes

Melbourne + Western

- Oncology
- Tissues Bank
- Diabetes
- Neuroscience
- MRI Images

South Aust

RAH

Q Elizabeth

Flinders MC

- Oncology
- Cystic Fibrosis

Tasmania

RHH

- Diabetes
- Oncology

ACT

Canberra

- Oncology

NSW

St Vincents

POW

- Oncology
- Tissue Bank

Queensland

Brisbane

- Oncology
- Epilepsy

Bendigo

- Oncology

Hume

- Oncology

Gippsland

- Oncology

Grampians

- Oncology

Barwon

- Oncology

Peninsula

- Oncology
- Diabeties

Cabrini

Oncology

Epworth

- Oncology

DHS

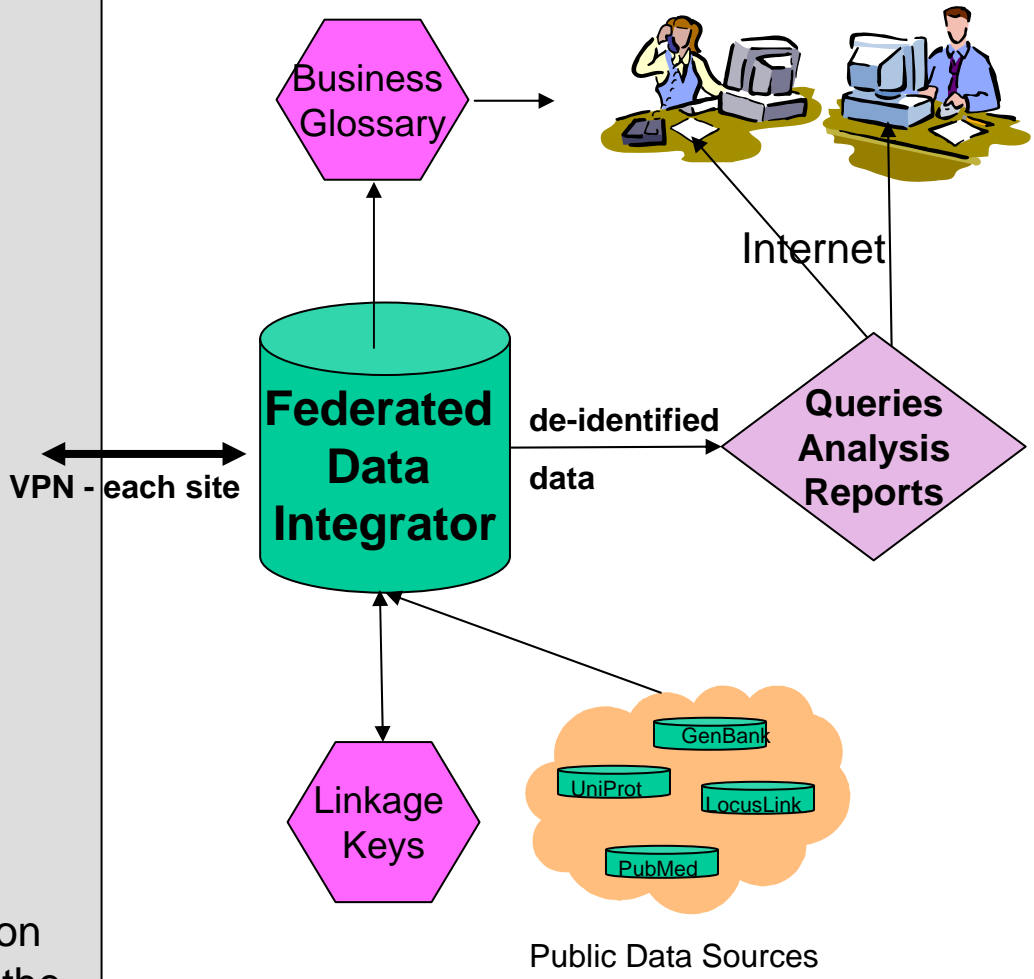
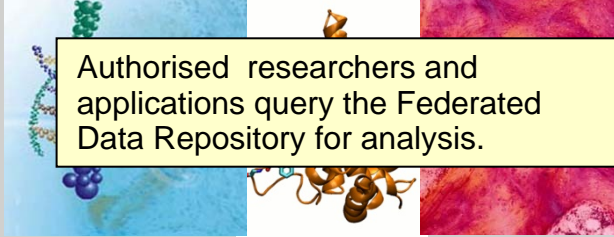
- VAED etc

AIHW @ MH

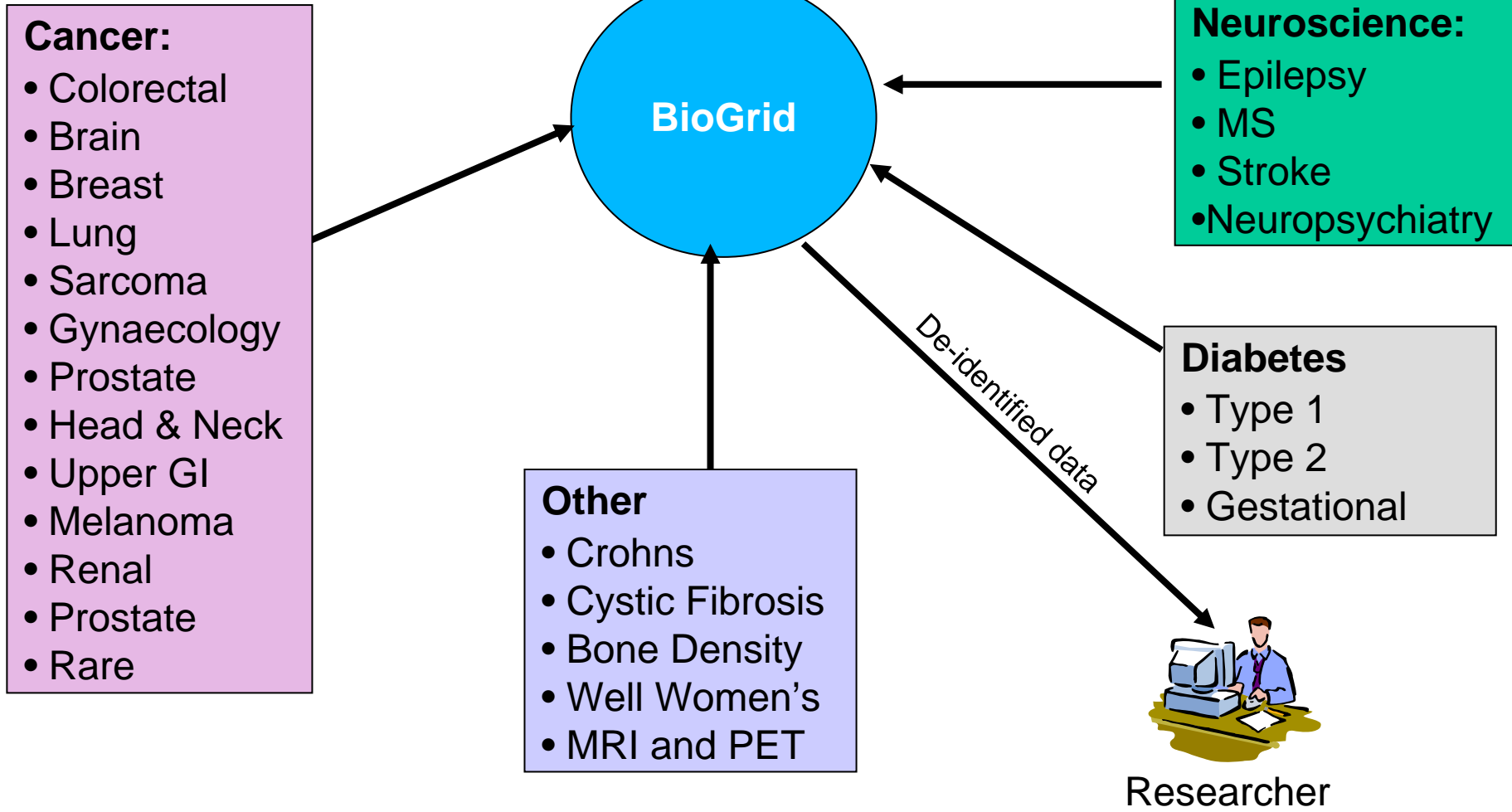
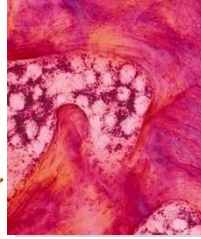
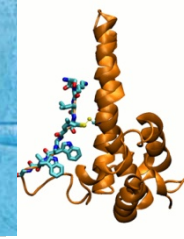
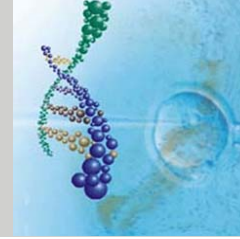
- Death + cause



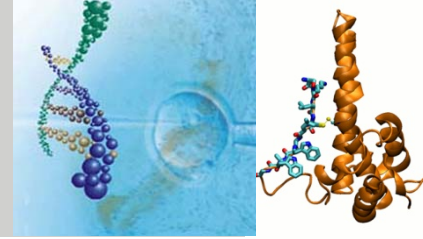
Data cached on computers at the sites - owned and controlled by site.



Scope - Clinical Datasets

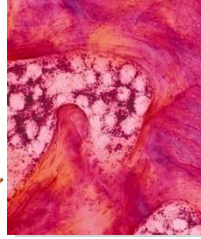
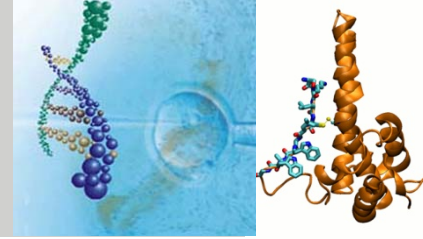


Types of Linkage:



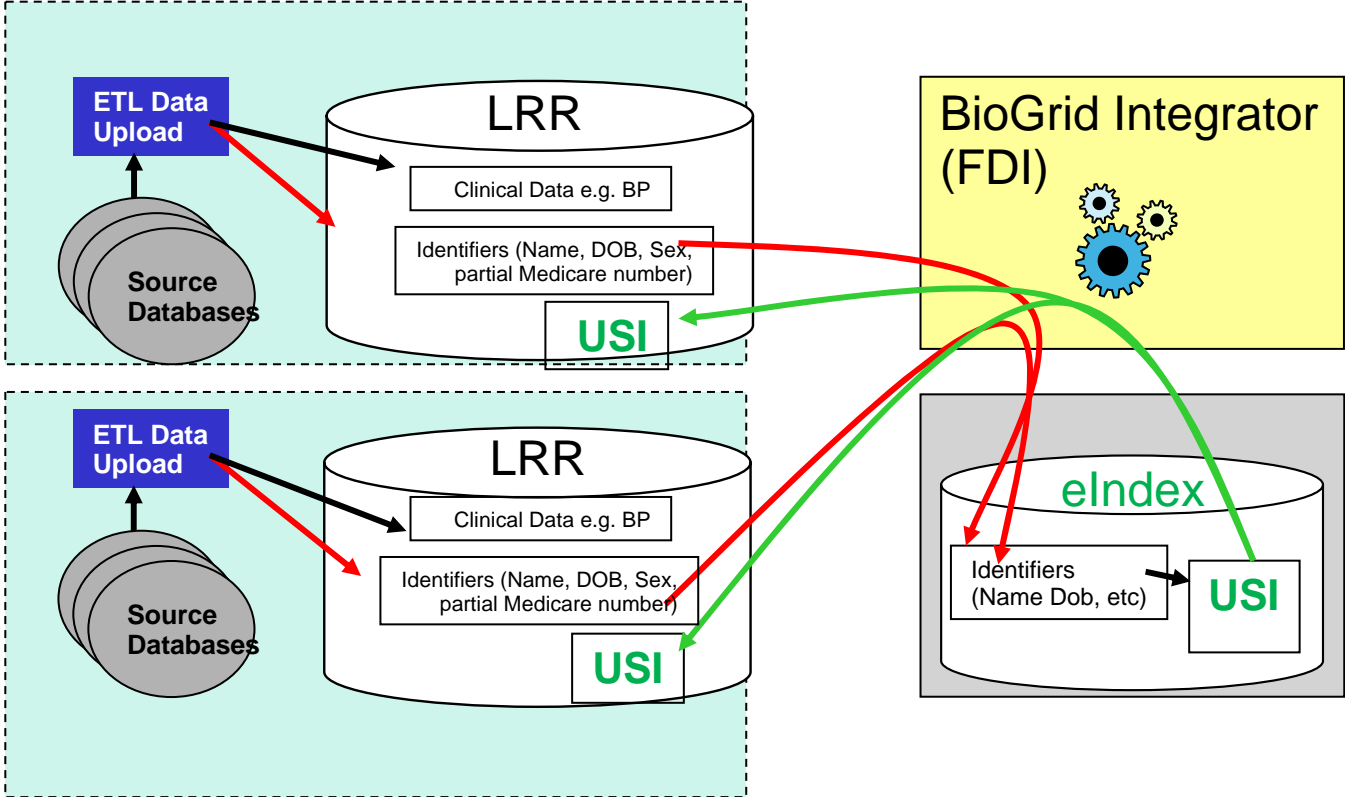
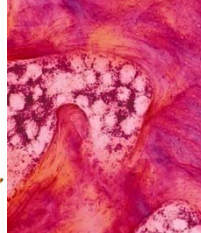
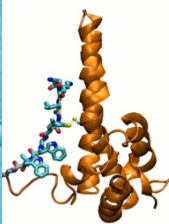
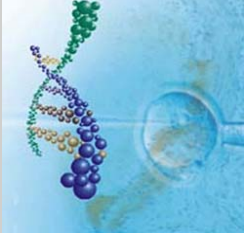
- **Probabilistic**
- **Hashing (exact)**

Probabilistic:

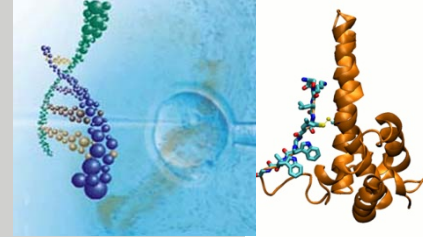


- Requires the two sets of identifiers to be brought together for comparison
- weights are assigned to each identifier
- Often includes options for soundex, transposition detection, exclusion of dummy patients (“Babe 1”)
- Final result is based on total of weighted matched fields, minus unmatched fields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

Linkage using eIndex

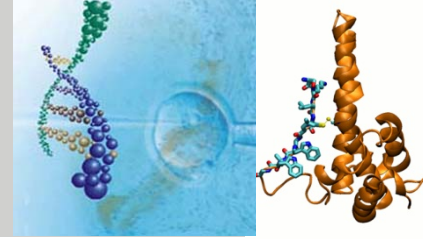


Operation of eIndex (1)



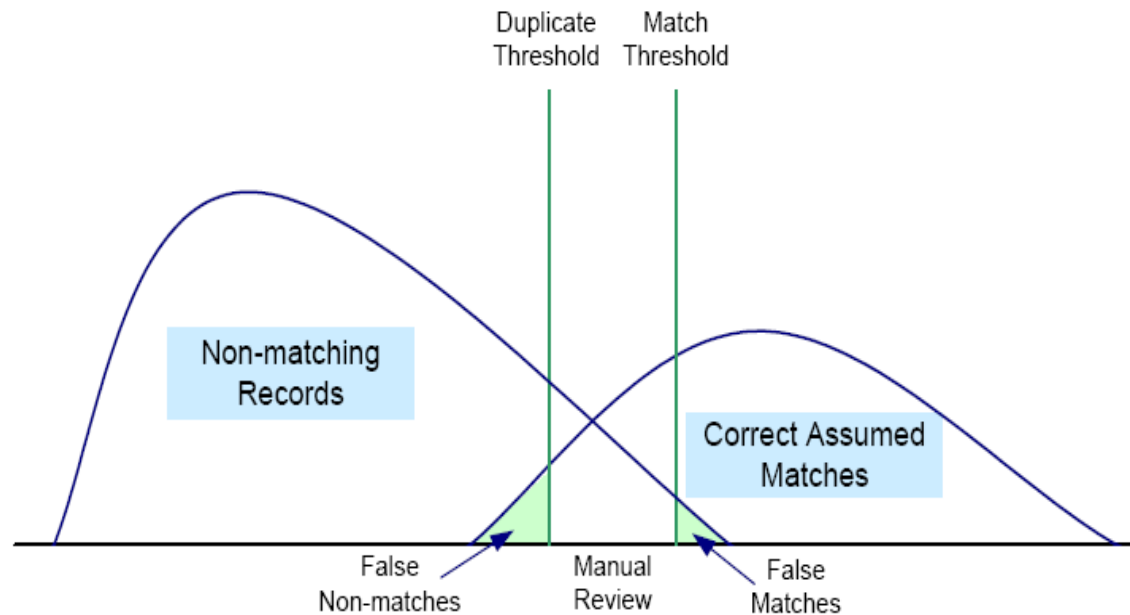
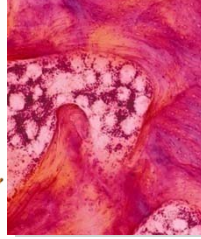
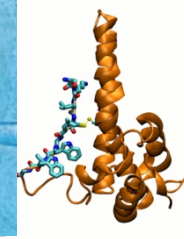
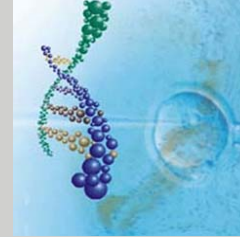
- Patient matches in BioGrid are determined by using these basic demographic fields:
 - Surname
 - Given name
 - Middle name / initial
 - Date of Birth
 - Gender
 - Digits 5 to 9 of the Medicare Number (Note that Medicare legislation does not allow the use of the full Medicare Number for identification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

Operation of eIndex (2)

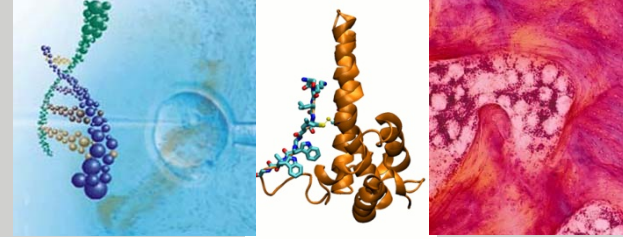


- The matching process includes logic to handle many common types of errors, e.g.
 - Names, dates and numbers which differ in only one position or in the transposition of two adjacent characters are given a weight for a possible match
 - Names are also checked for “soundex” matches and given a positive 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)

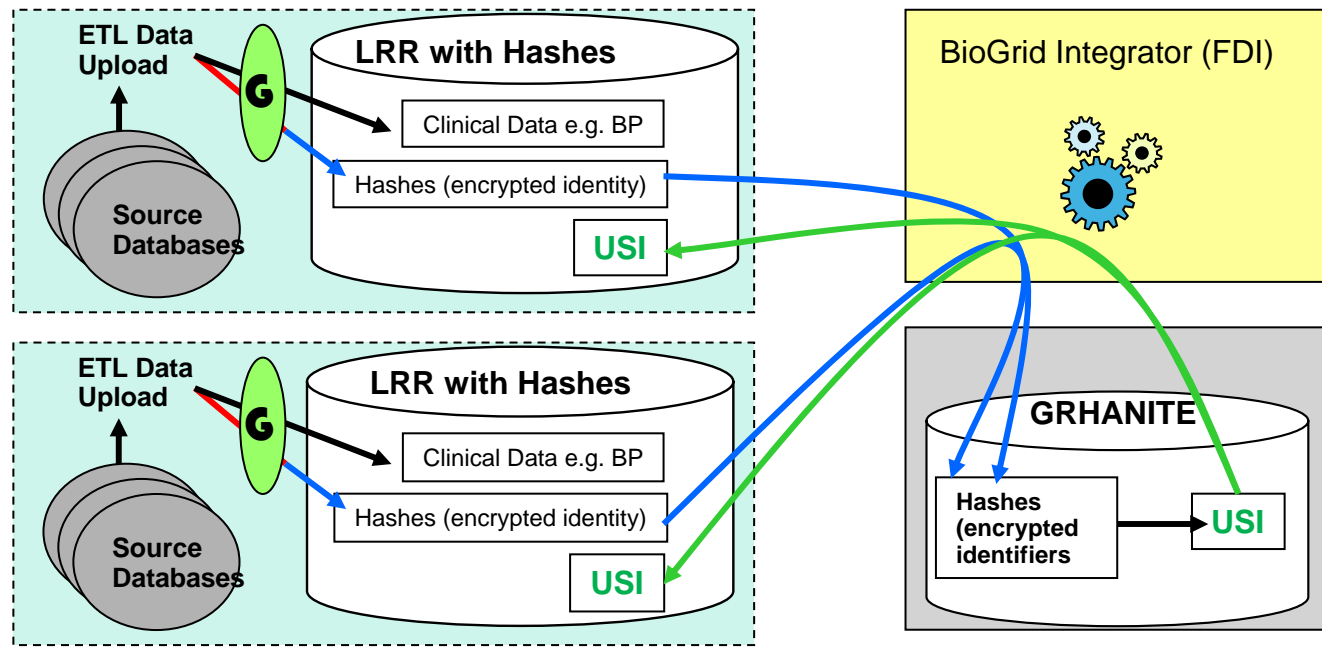
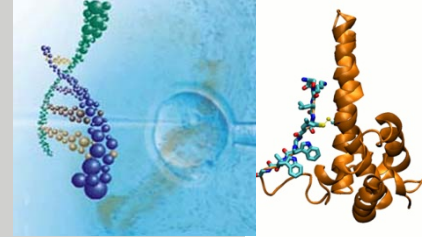


Hashing (“exact”):

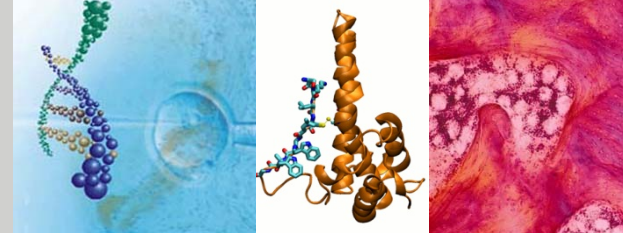


- A non-reversible hash is generated at each site and these are brought together for comparison
- Therefore does not require the 2 sets of identifiers to be brought together. This can be very important if the ethics or legal requirements preclude identifiers leaving both sites.
- In a simple hash system, the match must be exact – all or nothing. This makes them almost useless in real-life situations

Linkage using GRHANITE™ (1)



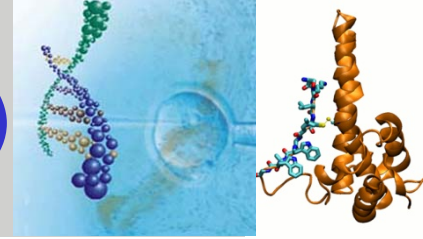
Linkage using GRHANITE™ (2)



	Patient_UUID	HashType	Hash
▶	88693834-84b5-478b-9c90-0058bbc6611c	5	j0OXm/un9hzOia5B2HO0pGACSTWT2Cl3VR6ez3kPJIEctRb5ICEdT/KMzoJXaI24dXENVdA18xc/h9X6boljdWTQ59qUR0mESadIKIng4=
	88693834-84b5-478b-9c90-0058bbc6611c	7	7WMmFzuuQjvKwFsO27+S3uUKRrxE+kW5vZ87Iv9Y86K2tmfK9EvXXOVcG5AN8Q9ngTBe400AzaUcoTtJ3rWKCvaf1zfzeqvi5CSoed55w5A=
	09442fe8-e4b4-4653-a750-0184f39d5f44	5	INTGcWAgvq8m+hWQxxsuHKUhrtp4TT8aDB5Eq4UgedmL9IUBvA9mDKZG4DfNqtWV/0X3BuGgdhEBOFv6PgU2zHpNCR1XMa/xaHi2JBsA7I0
	09442fe8-e4b4-4653-a750-0184f39d5f44	7	q8va4jildeq7R.Jhd5WEtHs0aeaNkMnYOXf7Y8JHa4VvJcZXXGhaV3WHbLBQ9aJ5yqSPTTZImAr/6jd/zFRw5XuPe4YPK4M17WtrgwON2o=
	09442fe8-e4b4-4653-a750-0184f39d5f44	8	WuWF65On43D7mYnUTa17EsQBjnUUgLDdJQ3cRowgheeOdhM7A3dlp27uI+8TliBX2Nf1dyPWUQOa7kr59nPfQnfHimwcPKLuba+M/YMCRhjQ
	a1b51347-91ef-40ff-990b-043ba2a29421	5	ynWWD2T57U5h+eKilaxOw00zRDB8apRfgaHyxu1pAubf5sIfRqDUJcMMjCSdchD+uXg3dVJruGi7/OM0qJsB8Jfdok/8uVVBVC/eBTOOE=
	a1b51347-91ef-40ff-990b-043ba2a29421	7	9B2k/Rw3ZDEPPCgJyOE4eNo5869feXvAGq4suZHdr91Zsj1t19OpDRwr5O8mKhTsan51tC0h2OrDe1O8cW58NMb7mqAwwN+5MxZCPNHPJi
	a1b51347-91ef-40ff-990b-043ba2a29421	8	rqz2XVN12uw6Wtg0nW1nI/LM1RsX0ntuUZGF+oqhBPZGQtVpuFHFYyUEhnBxt3ADxcUUGo/XIKJWSw4i35dy+5Yd8551toDt7Jsvya5c=
	3aa10887-016d-490e-8431-046e0292b1f1	1	9kmu3wWHTAOwsO45YR9F9JZGCDcP3ziYseRPYavBbYMDkIrc6iCiyzc1cD++qN1e7jqLIXbqe4TMDAHWY40YyrUkn/qToCY+OkAFCmHWA4=
	3aa10887-016d-490e-8431-046e0292b1f1	3	YDATI0Hy
	3aa10887-016d-490e-8431-046e0292b1f1	5	rJfm3GCM
	3aa10887-016d-490e-8431-046e0292b1f1	7	lk7203VV
	9765d1a0-65ae-4609-b830-047a51ca9f6f	5	j0OXm/un
	9765d1a0-65ae-4609-b830-047a51ca9f6f	7	7WMmFz
	c2a9ac18-125b-423e-838c-05167a0e340e	5	3aqUId/ks
	c2a9ac18-125b-423e-838c-05167a0e340e	7	BWUFnW
	c2a9ac18-125b-423e-838c-05167a0e340e	8	2QDF4de
	0aa28542-6831-4c88-a4cf-052c2697b100	5	SMCQ0bd
	0aa28542-6831-4c88-a4cf-052c2697b100	7	0ETkdHn1fGw486DBm07n58JXULD+I6aQiG8sflloWGCNG8ZMC4z2w2oICmaM3277n1XGoalThnH3CKWxR60FIK5YUpG5iRAl8Nky9mCAIKA=
	0aa28542-6831-4c88-a4cf-052c2697b100	8	ufMtnGEI2Jip54UPX+X0XdqYfWqNGm07JOZtTGSTWRV0hHof7/3KdhZ3YxGwqqZwMJJqsZ8bzT0dBppqIs3pok3gKRK8MSVKxeRNxHg47jg=
	1edaa12c-62ce-4d42-931f-05b8e06eab71	5	3R8U7gWRBjnfhybJcof6yZToUg8Yxa2q1EFZuNwZ5yWISxFHX2sgX+xlccI.NmfXLYcrbkEpe6wya2tXz1/78Od/lm/NQ0Rq55OjYiQUyos=
	1edaa12c-62ce-4d42-931f-05b8e06eab71	7	oFuK0oot769leG8zhtaOV5K9ef58D0q4mqnvQoU57o7cftmVC5BYHLxA+oV1Iw3qqMJTN2vU7mZ80AQ0BaBetsoZsuPYJuz+5+NMPZUE/0=
	1edaa12c-62ce-4d42-931f-05b8e06eab71	8	KoU6rMe9bf0TF1EU1yx3i6wu25oQQNtp5I5YVup0VF/NujeYuNYRqvr1X2Y4M7ZE10rZjtC0Q+bX2Vtm/g5fi1fgJ7QINucrD97WV5TjymA=
	d98a0c52-2109-40fa-af80-07a169cd93d2	5	7kiya5pu4WDMscoQPwc+yd12dytZIs2YiaI4P8mtmRtBqU+oGHInmYO1neAx2b4AYkwGex5s7R5nLxAg1uSrmH10B7P1FxdP6E1/RM9BwC=
	d98a0c52-2109-40fa-af80-07a169cd93d2	7	bQC/TrHtAaXWZ9tmjPmpkYhuyEY7yKzsozGQ3WRQPwRUFp5JxgXC36FQOv00iZPGbEY2yRJTtb51IVPNJYD+fZZYmKA/Lb5/mYB5ZwJrM=

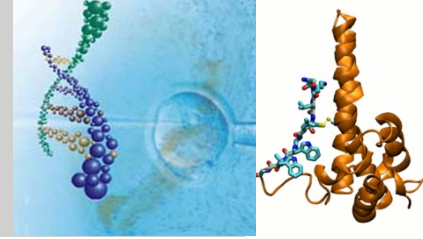
This form of encryption is non-reversible and heavily protected against dictionary attack

Linkage using GRHANITE™ (3)



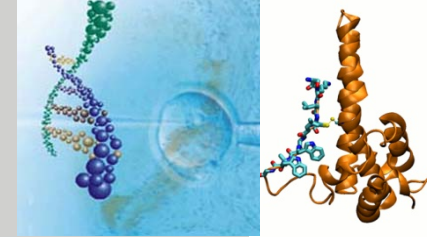
- “Hashing” creates a very long character string based on the incoming identifiers. The BioGrid-Grhanite system uses the SHA-256 algorithm, which cannot be reversed to reveal the original identifiers.
- The hash values are spread across the possible output values in a pseudo-random fashion. This means 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



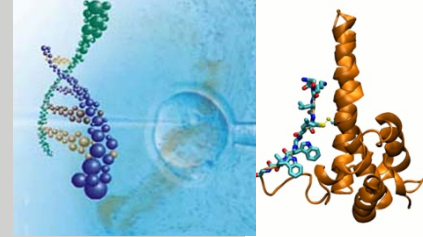
- There are about 45000 patient records in our files and about 35000 unique patients, meaning that there are about 10000 matches.
- About 100 “matches” in Grhanite that were not found in eIndex
- About 100 “matches” in eIndex that were not found by Ghranite
- “False positive” and “false negative” is 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)



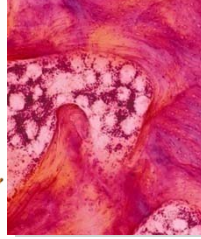
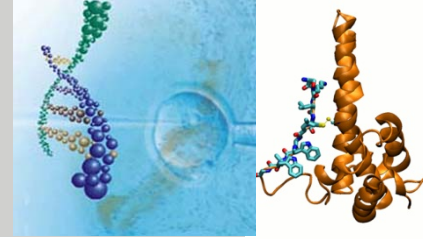
Software	Price	Description
Sun eIndex	High	<ul style="list-style-type: none"> - Includes fuzzy logic – various types - Includes high and low threshold values and grey area - Has tools to store manually decided matches / non-matches
Dataflux (SAS)	High ?	<ul style="list-style-type: none"> - Can also do data matching, including fuzzy logic – not clear how sophisticated this is
IBM Quality Stage	High	<ul style="list-style-type: none"> - Can assign weights to fields - Not clear the extent of fuzzy logic or soundex – fuzzy logic example but no explanation - Seems to just assign records to groups but no mention of automatic assignment of a USI type number - Determines its own “discriminatory power” for each data field using frequency analysis
Winpure	Moderate	<ul style="list-style-type: none"> - Basic (telephone numbers and emails) or advanced (names and addresses) – no mention of fuzzy logic - Does not seem to assign common numbers – looks like user must decide what to do with each one

Linkage software (2)



Software	Price	Description
Linkage Wiz	From \$550 to \$4250 (USD) depending on number of records to be linked	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> - But requires base SAS - Combination of probabilistic and deterministic algorithms - Handles Nicknames, soundex, approximate strings, transposed dates and digits in SSN - Looks quite good

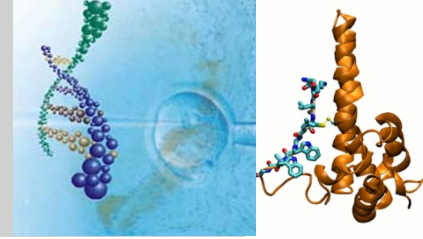
The Future for Record Linkage



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

A word from our sponsors ...



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*