

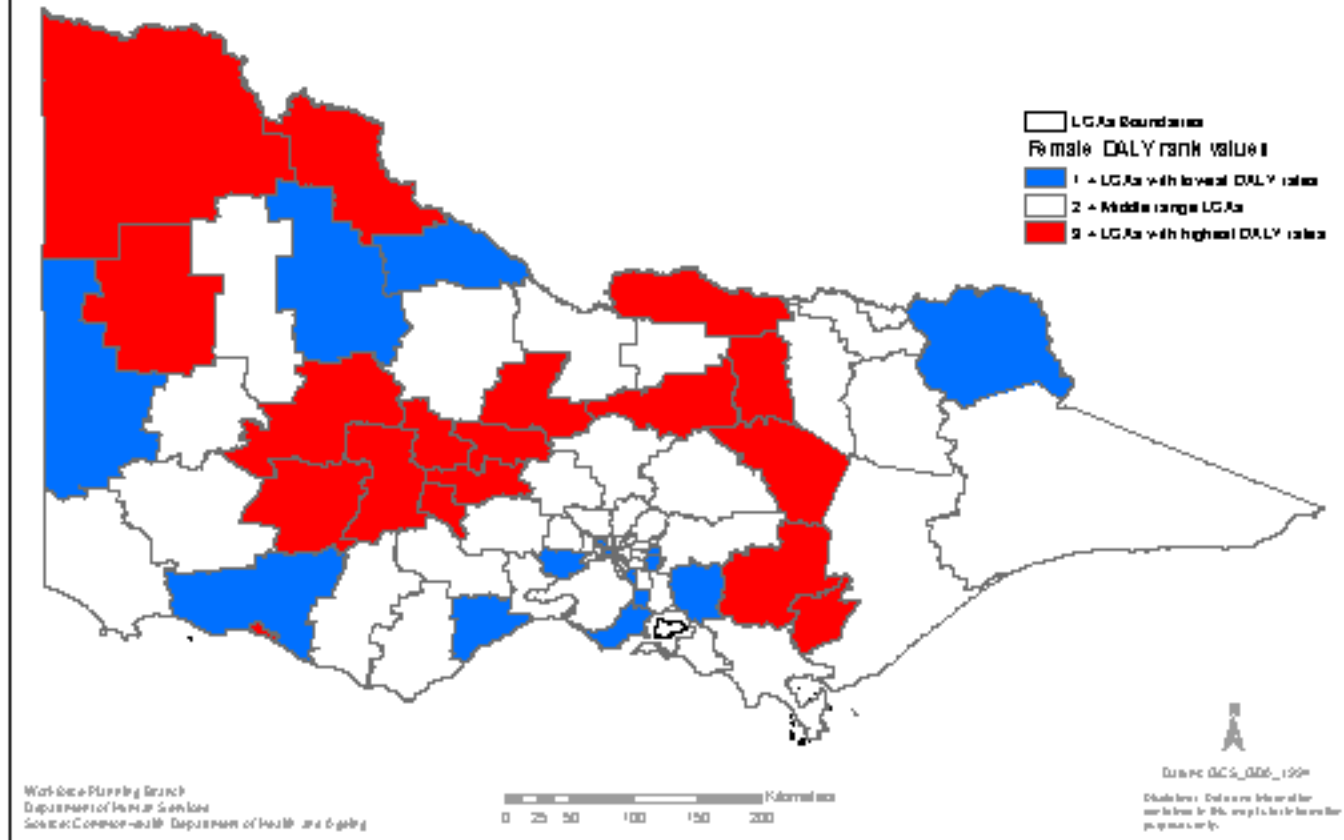
CRE Patient Safety Registry Interest Group
June 2008



Record Linkage at the Victorian Department of Human Services: past and future

Dr. Vijaya Sundararajan
Senior Medical Advisor
Funding Health Information Policy Branch

Female DALY rates by LGA, Victoria 2001



The Victorian Burden of Disease Study


Overview



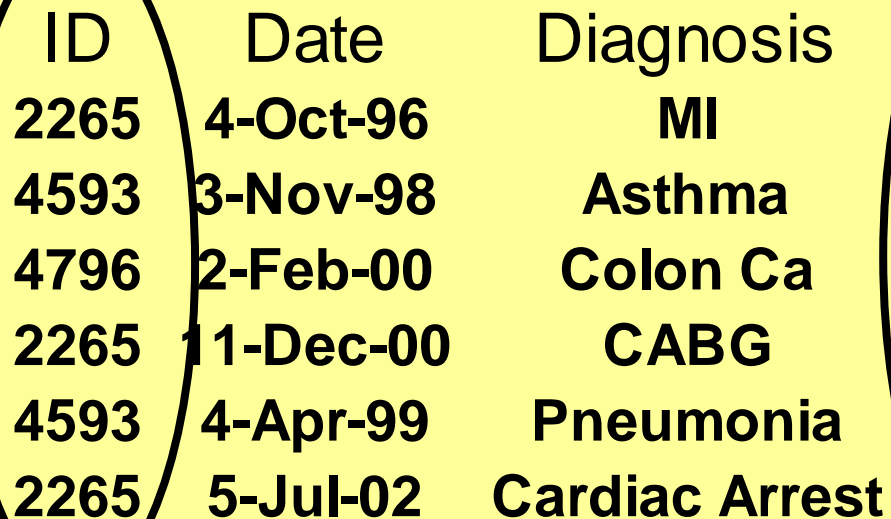
- Review of record linkage concepts
- Victorian linkage results
- Privacy, DHS policy, Ethics
- New record linkages
 - Process
 - Governance
- NCRIS
- Expanded Victorian Data Linkage Unit

Overview

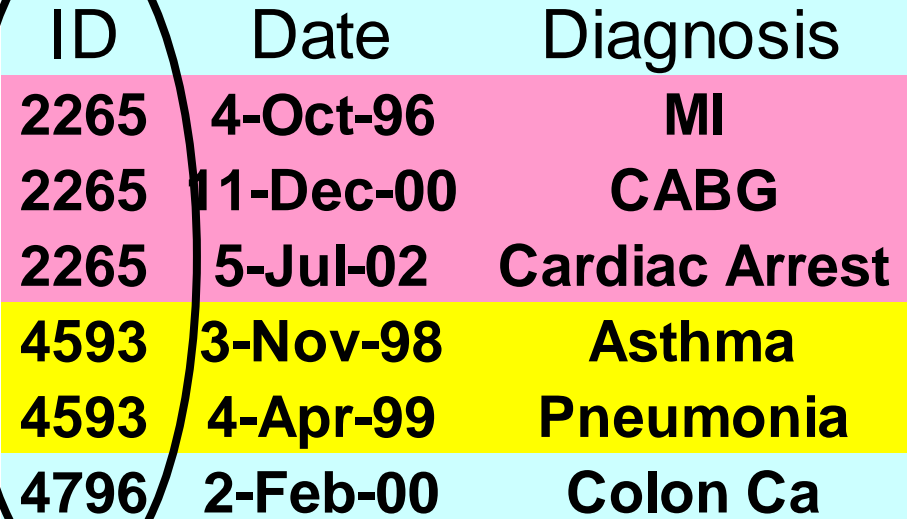


- Review of record linkage concepts 
- Victorian linkage results
- Privacy, DHS policy, Ethics
- New record linkages
 - Process
 - Governance
- NCRIS
- Expanded Victorian Data Linkage Unit

Data Linkage: Organising ONE dataset



ID	Date	Diagnosis
2265	4-Oct-96	MI
4593	3-Nov-98	Asthma
4796	2-Feb-00	Colon Ca
2265	11-Dec-00	CABG
4593	4-Apr-99	Pneumonia
2265	5-Jul-02	Cardiac Arrest



ID	Date	Diagnosis
2265	4-Oct-96	MI
2265	11-Dec-00	CABG
2265	5-Jul-02	Cardiac Arrest
4593	3-Nov-98	Asthma
4593	4-Apr-99	Pneumonia
4796	2-Feb-00	Colon Ca

Data Linkage:

Linking two different datasets

LOS	Diagnosis	ID
5	CABG	1239
4	MI	4596
3	PTCA	2236

ID	Cardiac rehab
1239	No
2236	Yes

LOS	Diagnosis	ID	Cardiac rehab
5	CABG	1239	No
3	PTCA	2236	Yes

Record	Year of birth	Month of birth	Day of birth	First name	Medicare number	Country of birth	Postcode
1	1952	9	25	missing	missing	Australia	3065
2	1992	9	15	Jane	45599203	England	3777
3	1952	9	25	Fred	34594789	Australia	3065
4	1945	7	25	Fred	44850961	England	3065

	A				B			
	Medicare number	Yr Brith	D Birth	M Birth	M Birth	D Birth	Yr Brith	Medicare number
Perfect Match	12467898	1976	23			23	1978	12467898
Random Match	1256758	1956	11			25	1978	2256378

Agreement in links: probability that a matching variable agrees exactly given that A and B are a perfect match (also known as "m" probability)

Agreement in non-links: probability that a matching variable agrees by chance given A and B do not match, that is are a forced match ("u" probability)

$$\text{Agreement weight } w = \log_2(m/u)$$

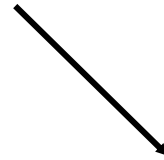
$$\text{Disagreement weight } w = \log_2((1-m)/(1-u))$$

Data characteristics

Variable	Present on every record?	Frequency of agreement in perfect matches	Frequency of agreement in random links	Agreement weight	Disagreement Weight
Year of birth	Yes	0.92	0.01	7.11	-3.63
Month of birth	Yes	0.85	0.08	3.35	-2.61
Day of birth	Yes	0.8	0.03	4.59	-2.27
First name	Yes	0.7	0.00	12.77	-1.74
Medicare number	In 70%	0.95	0.00	16.54	-4.32
Contry of birth	Yes	0.85	0.02	5.64	-2.71
Postcode	Yes	0.8	0.00	8.91	-2.32

Probabilistic record linkage

Cumulative weight w_t = summation of agreement and disagreement weights for each of the potential matching variables



Match	Record1	Record2	Sum of weights
1	1	2	-7.59
2	1	3	29.60
3	1	4	4.54
4	2	1	-7.59
5	2	3	-13.65
6	2	4	-11.25
7	3	1	29.60
8	3	2	-13.65
9	3	4	12.99
10	4	1	4.54
11			
12			

The higher the value of w_t , the greater the chance that the records belong to the same person

Stepwise deterministic linkage

Record	Year of birth	Month of birth	Day of birth	First name	Medicare number	Country of birth	Postcode
1	1952	9	25	missing	missing	Australia	3065
2	1992	9	15	Jane	45599203	England	3777
3	1952	9	25	Fred	34594789	Australia	3065
4	1945	7	25	Fred	44850961	England	3065

Linkage steps	Record at beginning of step	Records linked
Step 1. Link all records that match on the yearbirth linkage composite	1,2, 3, 4	1 and 3
Step 2. Of those records unlinked by the first linkage composite, link by medicare linkage composite (of those records with have both of these variables)	2,4	none
Step 3. Remaining unlinked records considered unique records	2,4	

Stepwise deterministic linkage

- Matching based on unique sets of identifiers predetermined by the researcher
- Combinations of identifiers tend to have equal weight; however, the order chosen in the algorithm gives them an implicit ordering of weights
- Identifiers are chosen based on the intuitions of the experienced researcher
- Subjective impression

Probabilistic linkage

- Based on the probabilities of agreement or disagreement between the identifiers
- All identifiers do not have equal weight
- Linkage is mainly dependent on the amount of discriminating power inherent in the variables common to the records that need to be matched
- Mathematically oriented


Matches by linkage	True match	False match
Non- matches by linkage	False non- match	True non- match
	True matches	True non- matches

One person or two?



Overview



- Review of record linkage concepts
- Victorian linkage results 
- Privacy, DHS policy, Ethics
- New record linkages
 - Process
 - Governance
- NCRIS
- Expanded Victorian Data Linkage Unit

Victorian data.....

DOB		
COB	URNO	Medicare
Postal code	Hospital code	FN3
Gender		

Hospital separations

Year	Separations	Case-groups
1996-1997	1,395,135	786,109
1997-1998	1,448,933	798,120
1998-1999	1,503,277	803,436
1999-2000	1,561,673	818,464
2000-2001	1,645,992	848,200
2001-2002	1,706,686	866,786
2002-2003	1,835,216	906,499
2003-2004	1,904,710	922,204
Total: 1996-2004	13,001,622	3,418,602

Emergency Department Visits

Year	Separations	Case-groups
1999-2000	1,561,673	517,879
2000-2001	1,645,992	535,022
2001-2002	1,706,686	607,322
2002-2003	1,835,216	636,815
2003-2004	1,904,710	632,959
Total: 1999-2004	4,774,761	2,017,485

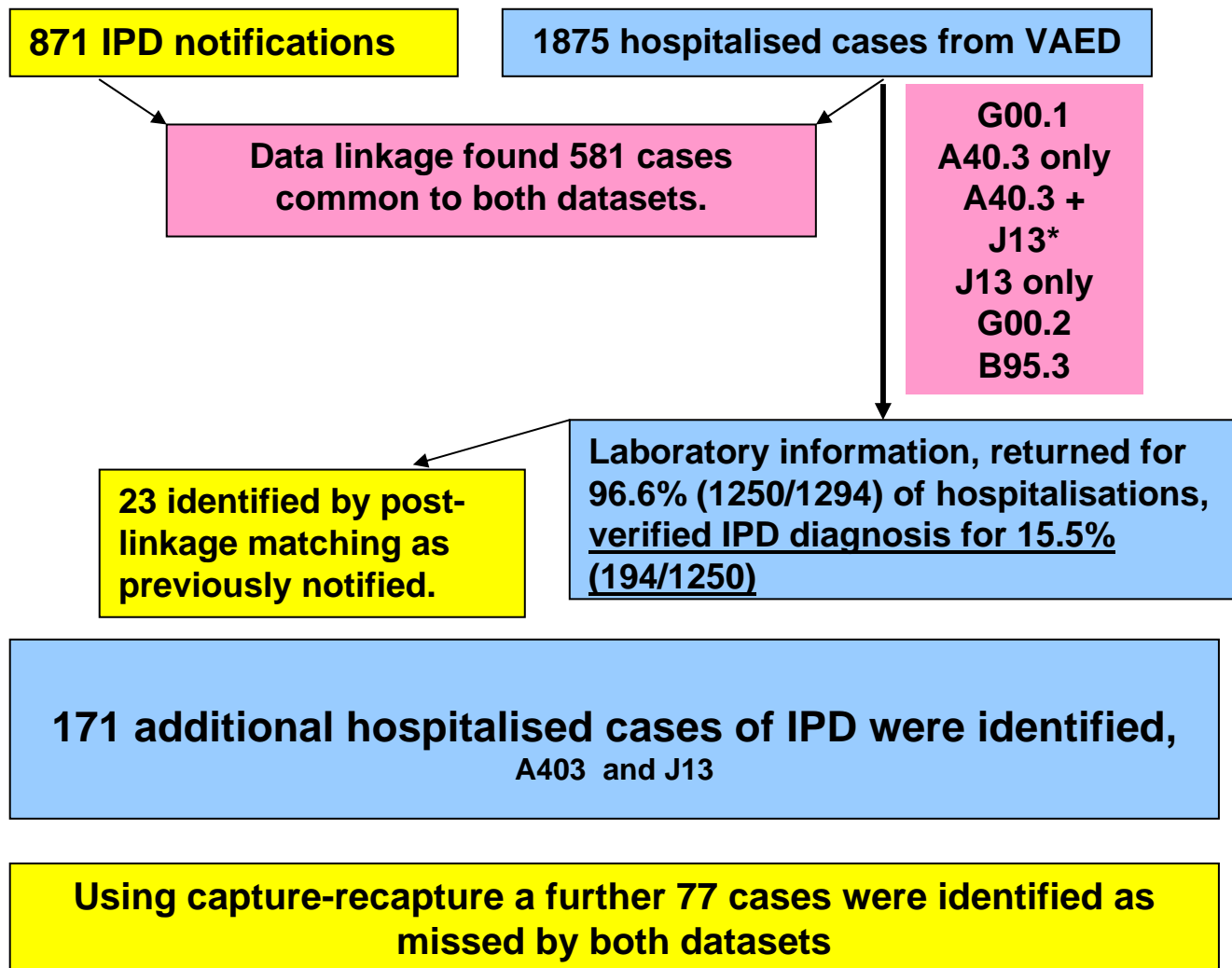
Characteristics of DVAID sample in comparison to VAED and VEMD

	VAED, 98-04		VEMD, 00-04	
	DVAID	All data	DVAID	All data
N	210,714	10,157,554	63,307	3,922,486
% of total	2%		2%	
No Medicare number, N	115,528	2,183,358	36,273	504,523
No Medicare number, %	55%	21%	57%	13%
Mean age	79	50	62	40
Gender, % women	33	54	39	49

	VAED, 1998-2004	VEMD, 2000-2004
Total observations, N	279,614	82,124
Unique groups from record linkage, N	67,988	39,083
Agreement	97.61 (97.5, 97.73)	99.58 (99.51, 99.64)


BASE	Routine linkages established	Current work	Future
<p>Hospital admissions 1996-2007</p>	<p>Emergency Department Visits, 1999-2007</p> <p>Victorian Registry of Deaths, 1996-2007</p> <p>AIHW Victorian Death Index, 1995-2005</p> <p>Aged Care, 2005-2007</p>	<p>BioGrid</p>	<p>MBS/PBS</p> <p>Cancer Registry?</p> <p>Perinatal?</p> <p>Outpatients</p> <p>Chronic Disease Management</p> <p>GP data?</p> <p>Other Registries</p>

Invasive pneumococcal disease in Victoria: a better measurement of the true incidence



Overview



- Review of record linkage concepts
- Victorian linkage results
- Privacy, DHS policy, Ethics 
- New record linkages
 - Process
 - Governance
- NCRIS
- Expanded Victorian Data Linkage Unit

DHS policy on health information and privacy

- Records held in data warehouse are anonymised (no patient names or addresses)
- Identity of individual patients cannot be reasonably ascertained.
- Reduces risk of identification of subjects
- Obligation to minimise risk by adhering to the Victorian Privacy Principles to make information on the health of Victorians and the performance of Victorian health services widely available.

Collaboration of DHS Ethics and Health Information

Two step approval process for data requests:

1. Victorian DHS Health Research Ethics Committee

Ethics committee approval – particularly for linked data/any unit level data/aggregate data which has small cell size

2. Health Information Unit

Works with researcher to limit potential identifiers in data – for example:

- 3 digit ICD-10-AM Diagnostic codes
- Age in 5 year groups
- LGA rather than SLA, post code is rarely released

Researcher required to sign conditions of release


Collaboration of DHS Ethics and Health Information

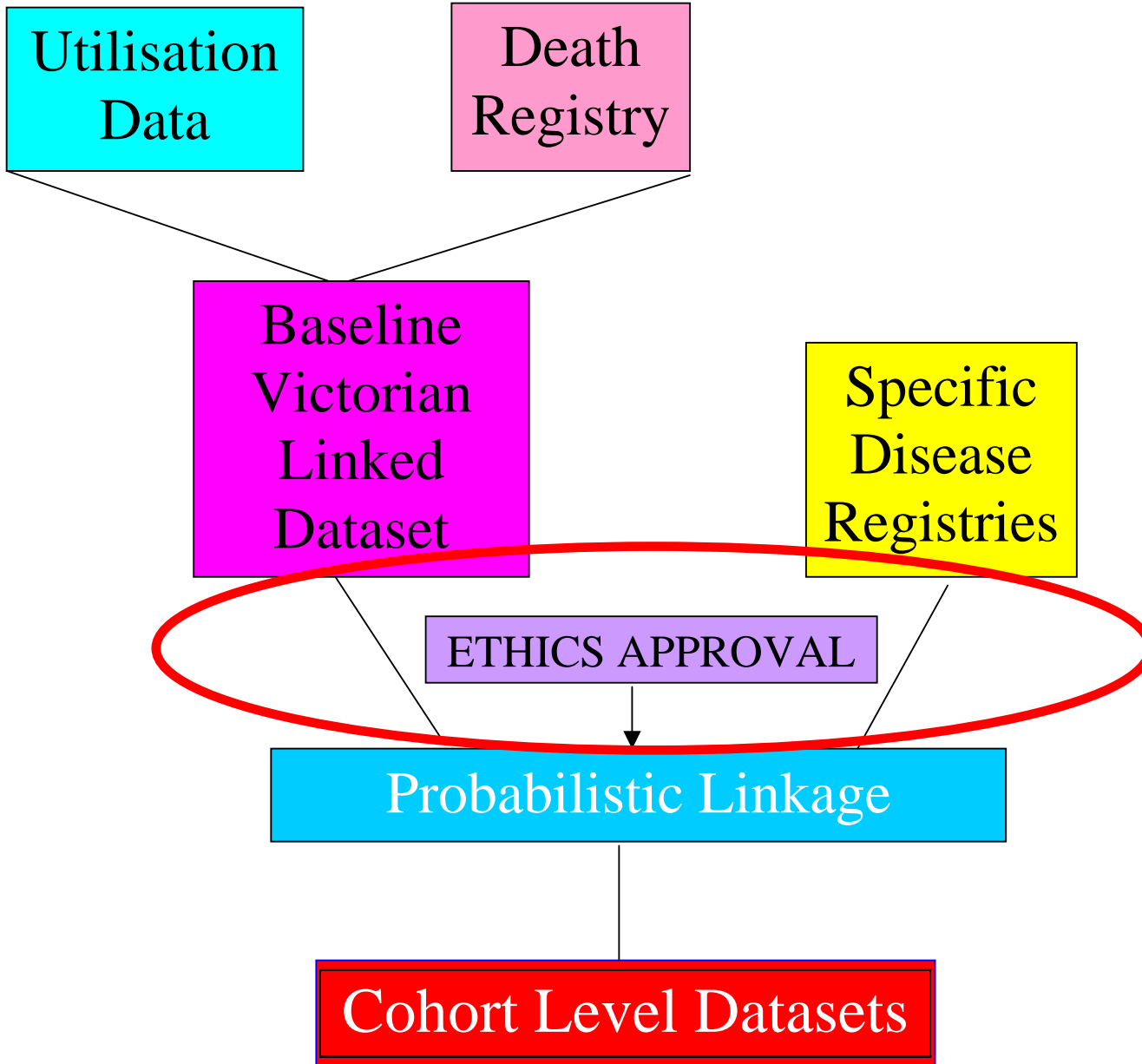
- Confirmation that researcher has no intention to identify individual patients or impact in any way on their care.
- Restricted, study-specific access to linked data
- Tight data disposal timelines
- Creation of an 2nd generation encrypted ID number for cases linked across episodes unrelated to any real identifier.
- **Removal of ANY and ALL identifiers** such as DOB/Medicare number and suffix/hospital record (UR) number



Overview



- Review of record linkage concepts
- Victorian linkage results
- Privacy, DHS policy, Ethics
- New record linkages 
 - Process
 - Governance
- NCRIS
- Expanded Victorian Data Linkage Unit



Original DATA			
DHS	DHS ID	Linkage VAR	Other VAR
Registry Data	Registry ID	Linkage VAR	Other VAR

Linkage data		
DHS	DHS ID	Linkage VAR
Registry	Registry ID	Linkage VAR

Output of linkage		
DHS-registry Map	DHS ID	RegistryID

Senario 1

Dual custodianship

1. Researcher goes to both custodians and requests access
2. Researcher identifies cases of interest from Registry.
3. Registry sends to DHS a file of the Registry ID's of these cases and a new project ID
4. Registry sends to researcher a file of the Project ID's and registry data for the cases
5. DHS extracts the DHS data for the particular Registry ID's
6. DHS then send the researcher a file of the Project ID's and DHS data
7. The researcher can use the Project ID to connect registry and DHS data for relevant cases.

Data for researcher does not have any linkage variables - that is no potentially identifying variables

Senario 2

MOU delegating custodianship to registry (or DHS)

1. Researcher goes to single custodian to request access
2. Researcher identifies cases of interest from Registry
3. Registry sends data with new project ID for researcher which includes both (original) registry and DHS data (such as number o hospitalisations for XXX)



Crude and adjusted odds ratios for mortality following a first acute cardiovascular event in RA vs. non-RA subjects

Event Type	Crude OR (95% CI)	Adj* OR (95% CI)
Myocardial infarction		
30 Day Mortality - All cause	2.3 (1.6-3.1)	1.8 (1.3-2.6)
30 Day Mortality - Cardiovascular	2.3 (1.7-3.2)	1.9 (1.3-2.7)
Stroke		
30 Day Mortality - All cause	0.9 (0.6-1.6)	1 (0.6-1.8)
30 Day Mortality - Cardiovascular	1.1 (0.6-1.8)	1.2 (0.7-2.0)

OR: Odds Ratio *Adjusted for age, gender, risk factors (hypertension, smoking, hyperlipidaemia), type of intervention after cardiac event (percutaneous transluminal coronary angioplasty, coronary artery bypass graft) - for acute myocardial infarction on

Overview



- Review of record linkage concepts
- Victorian linkage results
- Privacy, DHS policy, Ethics
- New record linkages
 - Process
 - Governance
- NCRIS 
- Expanded Victorian Data Linkage Unit 

NCRIS

- CW funds for developing research infrastructure
- Population Health and Clinical Data Linkage - \$20 million
- Each State/Territory submitted bids
- Victoria has requested funds for a unit with expanded capability to sit under the DHS CIO's office, with collaborations including universities and Dept EDU.

Possible unit structure

Program Director

**Data Linkage
Methods**

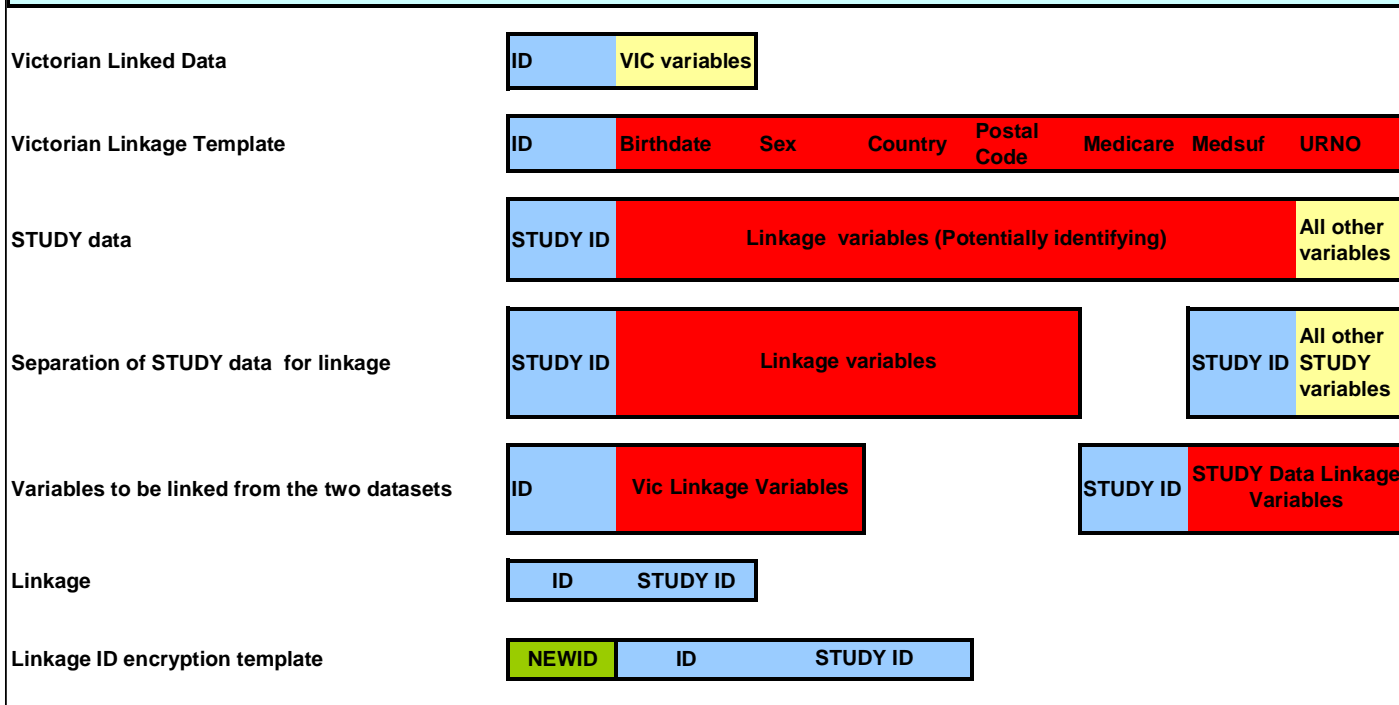
Data Services

**Governance,
privacy,
communication
and performance**

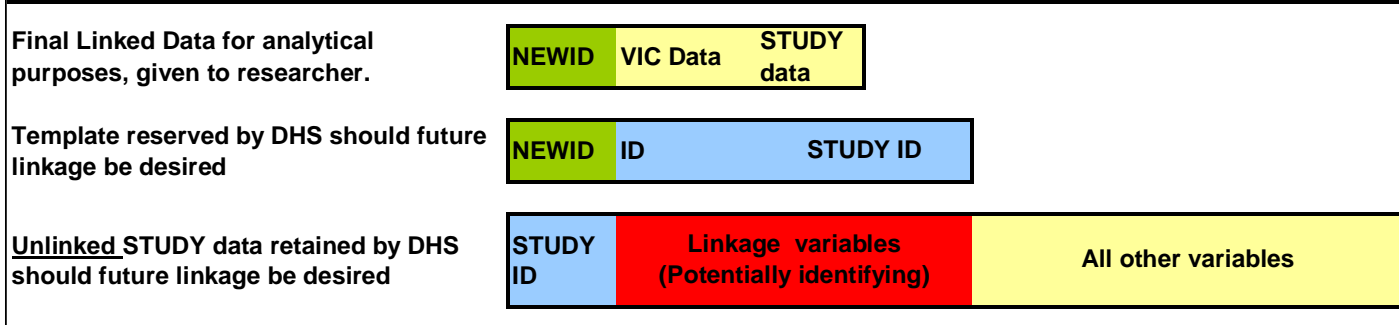
**Capacity for 4 part time research fellows to
come and work with unit on a project of their
own design.**

VIC data=Victorian Linked Data

STUDY data=dataset to be linked



Linkage complete



Colour legend of variable types

- Original ID's from each dataset
- Linkage variables (potentially identifying)
- All other variables
- Linkage ID