

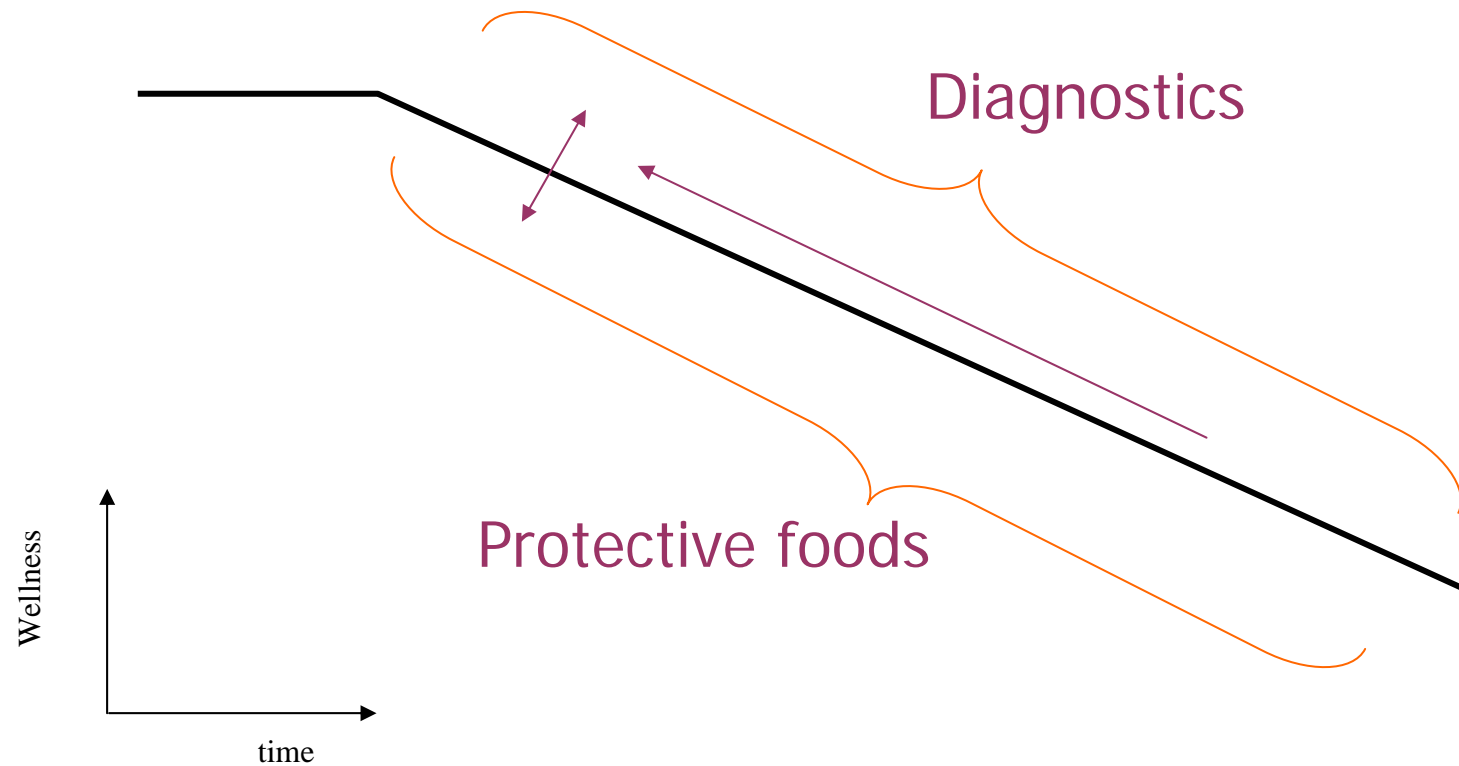
Trends in Health Data Integration

Dr David Hansen
Team Leader, HDI Project, EHRC

- What and Why of Health Data Integration
- Current efforts in Health Data Integration
- Data Integration Technology
- HDI - the EHRC Health Data Integration product
- Open questions and issues

CSIRO p-Health Flagship Approach - Colorectal Cancer

www.e-hrc.net



Health data integration/ Lifestyle determinants

- \$15 million JV between CSIRO and Queensland Government
- Undertake applied research and conduct clinical trials in conjunction with State Health agencies
- Research Program : “Improving Disease Management – providing information and knowledge where it is needed, when it is needed and in the form it is needed to support clients, patients and clinicians.”



Queensland
Government



CSIRO

E-HEALTH
RESEARCH CENTRE

What and Why of Health Data Integration

Environmental Drivers

www.e-hrc.net

- Restructuring of the Health System (world wide) to meet the demands of ageing populations (major economic impact)
- Funding agencies driving multi-disciplinary research initiatives → solutions to national healthcare priorities → data sharing needs
- Increasing pace of growth of the scientific knowledge base (e.g. genomics data) → increased need for linking data across disciplines
- Development and application of high-cost medical technologies → need for cost/benefit analyses (global benchmarking)
- Declining availability of specialist skills (“baby boomers” leaving the workforce) → improving productivity through advanced systems becomes a priority
- Private-sector healthcare providers taking over more routine procedures from Public providers → data access issues

- To answer a health related question requires
 - Identification of the real question
 - Design of the study
 - Collection of the data
 - Analysis of the data

- Does the data exist currently?
 - In many situations some or all of the data required exists across several data sources

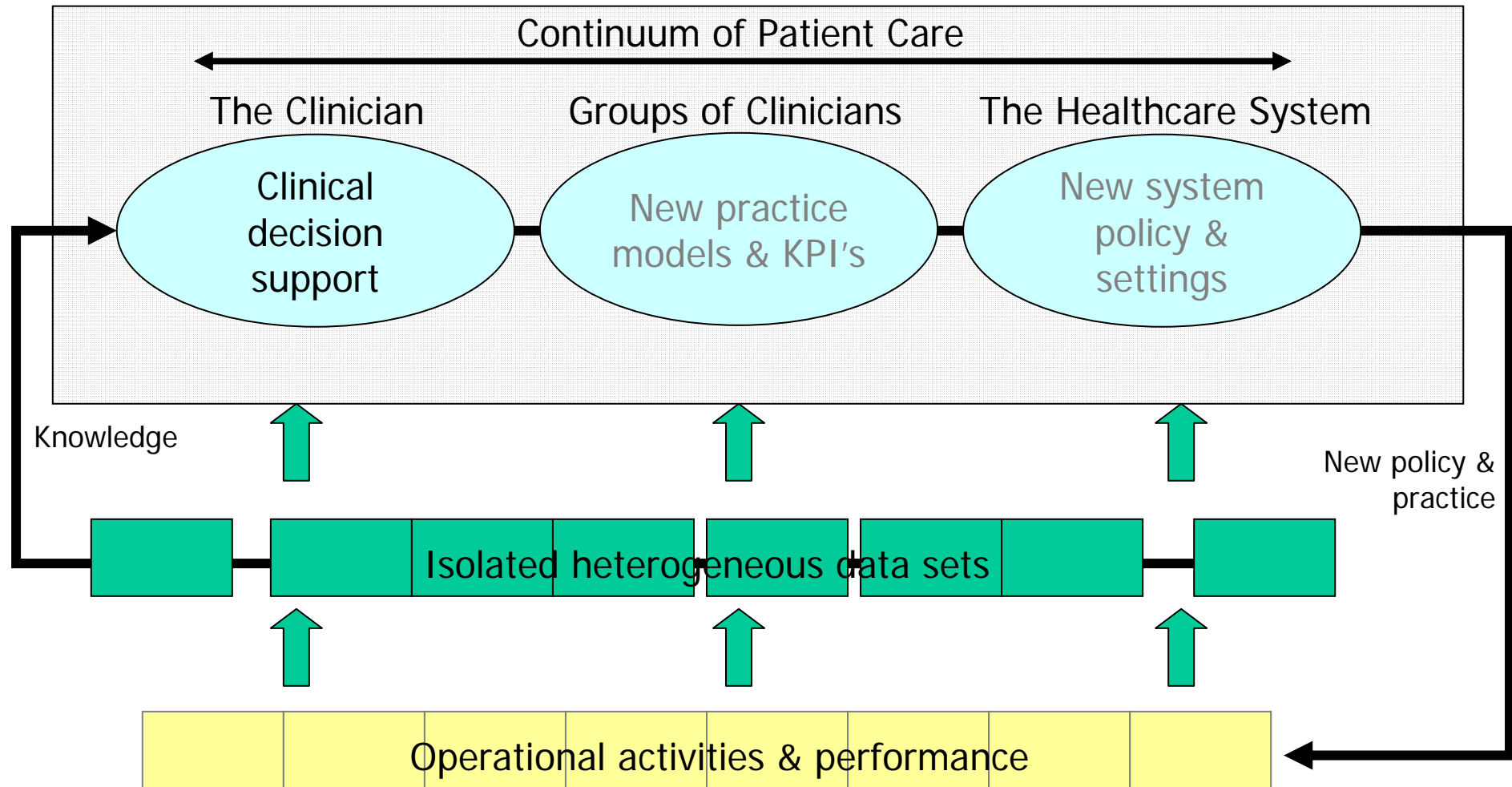
Issues with accessing and using current patient data

www.e-hrc.net

- **Patient Data spread across several databases**
 - Hospital administration and clinical databases
 - Pathology and Pharmaceuticals databases
- **No common person identifier**
 - New identifying number or ID per institution or service provider
- **Privacy and Security**
 - Patient concerns
 - Legislative requirements
 - Data ownership concerns
- **No way to easily manage access to data in multiple databases**
 - Involves significant time and manual handling of data between computer systems
- **Data quality and consistency**
 - Data entry errors
 - Non-standard coding and formats

HDI Vision – Fulfilling the “Unmet Need”

www.e-hrc.net



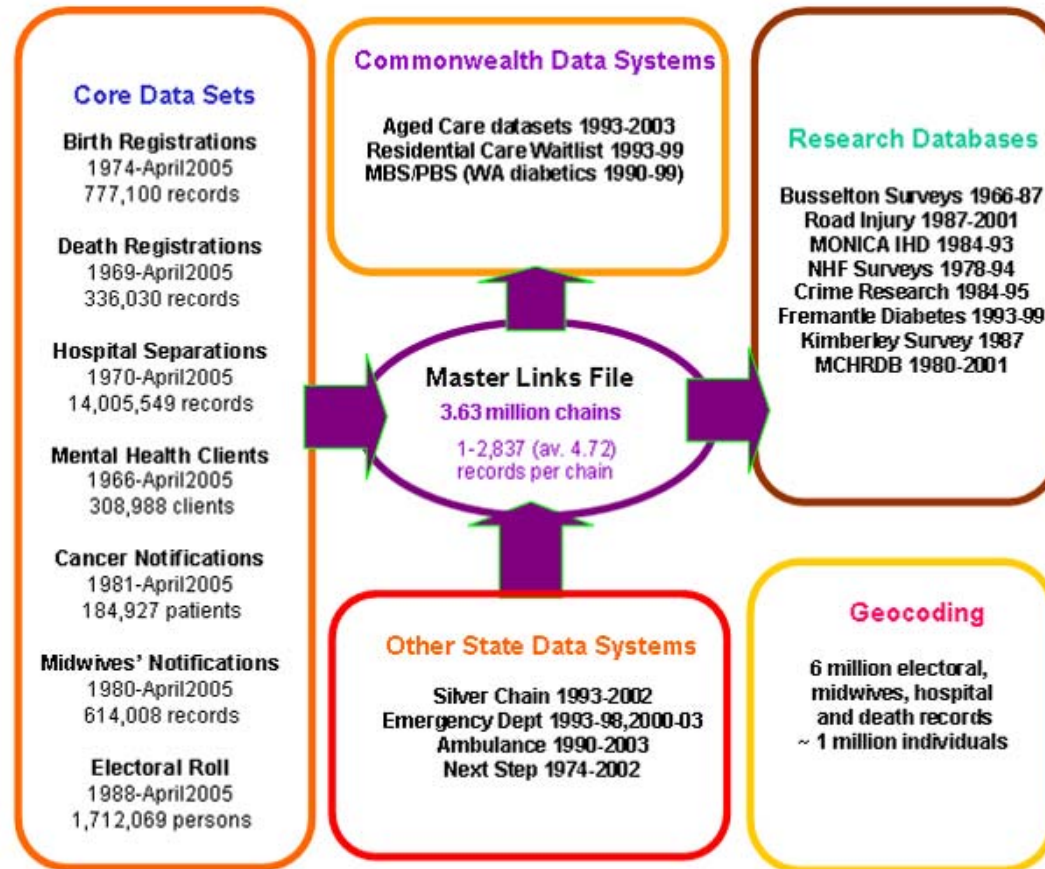
Current efforts in Health Data Integration

- **Use the data we have better**
 - Data Linkage Units
 - National Data Network (NDN)
 - Various software solutions
- **Collect the required data**
 - Electronic Health Records
 - Clinical Information Systems

- Data Linkage units provide linked data sets for epidemiological studies examples
 - Western Australia
 - Incorporates data linkage information from 67 data sources
 - Canadian Data Linkage Units
 - Manitoba Population Health Research Data Repository
 - NSW
 - Tender currently open to
Identify issues in and options for a data linkage facility in NSW

The Western Australia Data Linkage Unit

www.e-hrc.net



WA DATA LINKAGE SYSTEM
April 2005

- **National Data Network**
 - Currently being developed by the Australian Bureau of Statistics
 - Efforts aimed at standardising the collection of data so it can be integrated across the internet
 - Offers benefit of inter-agency integration of data

Electronic Health Records

www.e-hrc.net

- **UK**
 - National Program for IT – NPfIT
 - 6 billion pounds to put electronic health records throughout the NHS
 - 30,000 GPs and 300 Hospitals will be connected
 - Patients will have access to their Health Care Information
- **USA**
 - Appointed a National Health Information Technology Coordinator for *for widespread deployment of health information technology within 10 years to realize substantial improvements in safety and efficiency*
- **Australia**
 - Health Connect
 - Several trials around Australia
 - NEHTA
 - Accelerate the uptake of e-health in Australia
 - Standard for Patient Summaries and Electronic Medical Records
- **Generic Commercial solutions exist**

Data Integration Technology

- **Web Services**
 - Easy(ier) federation of data and services
- **XML**
 - Mechanism for text based exchange of data
- **Ontology**
 - Relating concepts using semantic descriptions
 - Discipline based
 - Need domain expertise to define the ontology of data sources and fields
 - Removes ambiguity and provides base line agreement on terms

- **Integration of heterogeneous data sources**
 - Wrapper technology allows any data or service to be incorporated
 - Query optimizer
 - Targeted at life sciences

- **Note: Oracle gateway is a similar technology**
 - Primarily an enabler for customers to migrate from other data sources into Oracle

- The bio21 consortium of 21 hospitals and academic institutes in Melbourne
- MMiM
 - 6 hospitals, 3 disease states
 - Federation of data across the hospitals to provide data in the same format
 - Uses IBM Discovery Link to query the data sources at each hospital and bring back data
- Small step towards integration of genomic and health data
 - Some of the data was mutation and SNP data

- Integrates data sources and analysis tools
 - Flat File, XML and relational data base
 - Any command line analysis tool
- Integrates over 1000 data sources and 250 analysis tools
- Wide usage in biotech and pharmaceutical industries
- Meta data driven approach
- Web service based with powerful web service API and web interface

Use Meta Data!!

www.e-hrc.net

- Use meta data to describe
 - the data source
 - how data elements interact with the system
 - the behaviour of data
 - relationship between data
 - etc

HDI - the EHRC Health Data Integration product

To enable effective and efficient interpretation of data to support innovation in health care for Australia

- By providing effective linkage of health data in a secure environment with ease-of-use in multiple applications

Development of software tools for linking, integration and analysis of data for cancer patients

- Qld Health Cancer Control and Analysis team - improving clinical and patient outcomes for Lung Cancer
- MoU National Bowel Cancer Program - Qld Health, Royal Melbourne Hospital & Flinders Medical Centre

Intent is to develop HDI as the industry standard for secure data linking

What is HDI?

www.e-hrc.net

- Tool that *integrates disparate sets of data* through sophisticated algorithms to *match patient records* while protecting the *privacy and security of patients' data*
- Enables all users to work with authorised data sets through a *standardised layer of metadata* over the source databases
- Provides for both *data integration as well as analysis and reporting on that data in the one tool*

Innovation and Features of HDI

www.e-hrc.net

- New sophisticated algorithms and probabilistic matching used to determine common patient records across databases
 - Matching on encrypted identifiable patient data
- No warehousing of data
 - Data Custodian retains control and security
 - All identified data encrypted before leaving data custodian
- Metadata layer linked to industry standards to provide a common language and across data bases – increasing usability
- Ability to perform both complicated analysis and to establish commonly run reports
 - Including ability to easily change constraints or variables in the data to manipulate and interrogate results

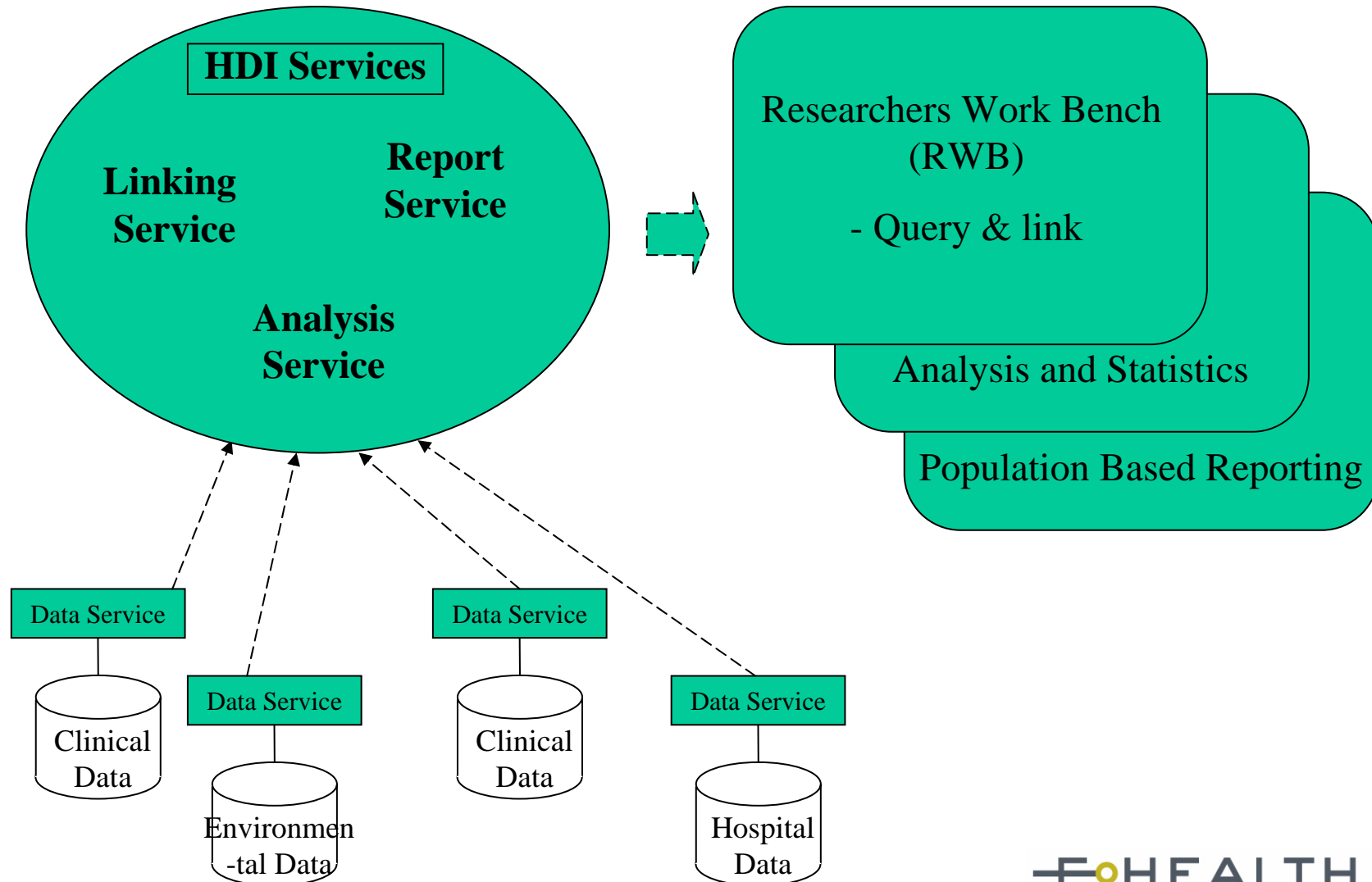
How can HDI be used?

www.e-hrc.net

Through the integration of previously ‘un-linkable’ data:

- Integrate and analyse larger sets of data for research
 - Survival analysis and treatment outcomes
 - Discovering previously unknown links

- Examine population based sets of data – possibly through standard reports
 - Key Performance Indicators, Clinical Indicators and Benchmarking
 - Public health and policy making decisions and monitoring (eg access to services, waiting times)



HDI™ Case Study – The Data

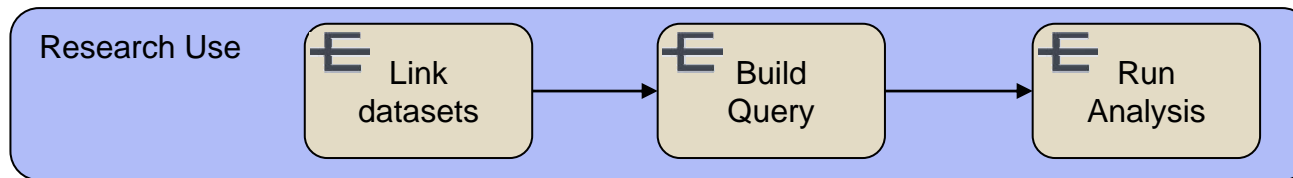
www.e-hrc.net

Three synthetic data sets:

- MUD – clinical patient summaries (based on a ‘typical’ cancer like disease)
- RAD – radiotherapy information
- MUDE – air pollution information

HDI™ Case Study - Summary

www.e-hrc.net



For example:

- Disease incidence trends
- Treatment outcomes
- Discovering previously unrelated links

Report Use



For example:

- Key performance indicator reports
- Policy decision support reports
- Benchmarking reports

HDI™ Case Study – Link

www.e-hrc.net

Research Use



For example:

Patients whose basic clinical cancer information is contained in the MUD clinical database may also have more detailed radiotherapy treatment information in the RAD database .

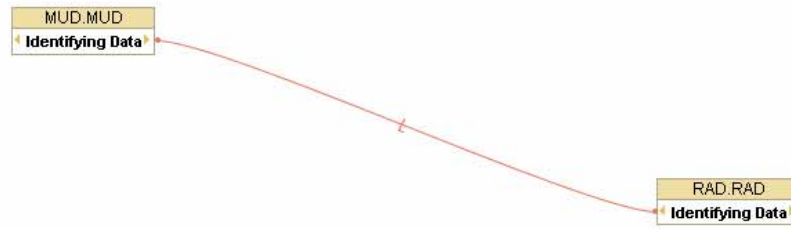
However there is no common identifier between the 2 databases.

Link these 2 databases using 'probabilistic matching' to determine which patient records match.



- Data Sources
- MUD
 - MUD
 - RAD
 - RAD

Link Results





Data Sources

- MUD
 - MUD
- RAD
 - RAD

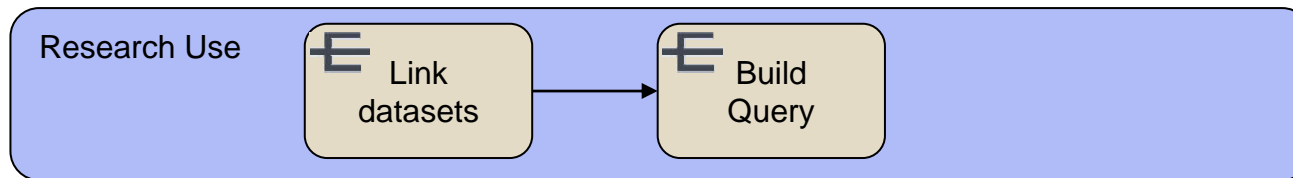
Link Results

There were 642 matches - This data has been saved to HDI Linkage data source.

Row	IDA	IDB
1	0652-4821-5122	979.0
2	0652-4821-5125	2038.0
3	1354-4826-4686	675.0
4	1354-4826-4688	1483.0
5	1354-4826-4690	2890.0
6	1354-4826-4699	698.0
7	1354-4826-4703	792.0
8	1354-4826-4720	1762.0
9	1354-4826-4721	1815.0
10	1354-4826-4726	2084.0
11	1354-4826-4728	2368.0
12	1354-4826-4730	2512.0
13	1354-4826-4736	2828.0
14	1354-4826-4737	2891.0
15	1354-4826-4740	38.0
16	1354-4826-4744	87.0
17	1354-4826-4745	193.0
18	1354-4826-4751	652.0
19	1658-6428-1009	446.0
20	1658-6428-1017	1669.0
21	1658-6428-1021	2825.0
22	1658-6428-1023	701.0
23	1658-6428-1025	1085.0
24	1658-6428-1028	2451.0
25	1658-6428-1034	2453.0
26	1658-6428-1036	2908.0
27	1658-6498-5216	667.0
28	1658-6498-5217	1271.0
29	1658-6498-5218	1489.0
30	1658-6498-5229	2233.0
31	1658-6498-5230	2303.0
32	1658-6498-5231	2613.0
33	1658-6498-5232	27.0
34	1658-6498-5235	516.0
35	1658-6498-5252	2741.0
36	1658-6498-5253	2847.0
37	1856-4826-5584	2031.0
38	1856-4826-5585	2090.0
39	1856-4826-5587	2223.0
40	1856-4826-5588	2346.0
41	1856-4826-5592	2472.0
42	1856-4826-5598	2868.0
43	1856-4826-5601	2978.0
44	1856-4826-5605	117.0
45	1856-4826-5624	1142.0
46	1856-4826-5632	1441.0
47	1856-4826-5633	1494.0
48	1856-4826-5635	1565.0
49	1856-4826-5647	2432.0
50	1856-4826-5651	2764.0
51	1856-4826-5652	2787.0
52	1856-4826-5661	830.0
53	1856-4826-5670	1531.0
54	1856-4826-5672	1703.0

HDI™ Case Study – Query

www.e-hrc.net



For example:

We would like to see if different radiotherapy regimes have an impact on survival of cancer patients.

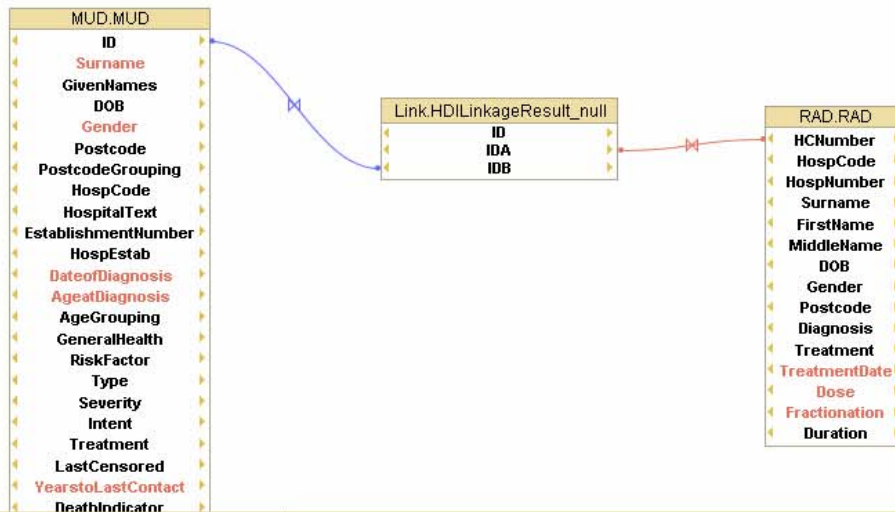
1. Query on linked MUD clinical cancer information database and RAD radiotherapy database to get a dataset of information on cancer patients, their current status and their radiotherapy treatments.



Data Sources

- [-] MUD
 - [x] MUD
- [-] RAD
 - [x] RAD
- [-] MUDE
- [-] Link
 - [x] HDILinkageResult_null

Metadata Query Results



Set linking operator

Attribute 1	Operator	Attribute 2
Link.HDILinkageResult_null.IDA	=	RAD.RAD.HCNumber

Join Union



Data Sources

- [-] MUD
 - [+] MUD
- [-] RAD
 - [+] RAD
- [-] MUDE
- [-] Link
 - [+] HDILinkageResult_null

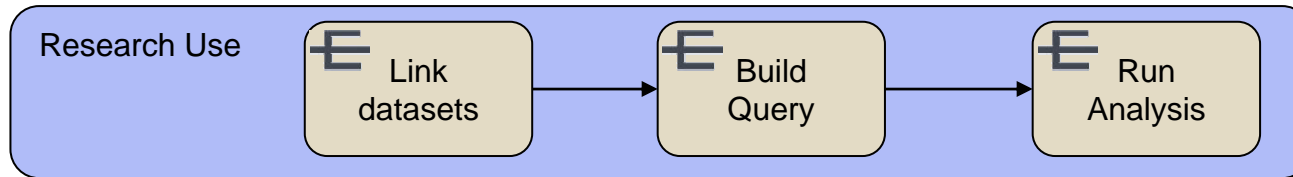
Metadata Query Results

The query returned 643 rows: Page 1 of 7

Row	Surname	Gender	DateofDia...	AgeatDiag...	YearstoLa...	DeathIndic...	Treatment...	Dose	Fractionati...
1	zu3up+juE...	F	2002-02-1...	86.117808...	0.3863013...	1.0	2002-03-0...	40.0	15.0
2	zbwFRAbv9...	F	2002-04-1...	64.430136...	1.8082191...	1.0	2002-04-3...	72.0	1.2
3	zbwFRAbv9...	F	2002-04-1...	64.430136...	1.8082191...	1.0	2002-11-0...	81.6	1.2
4	Jk+8xZrdB...	M	2002-09-1...	67.783561...	1.6931506...	1.0	2002-09-2...	70.0	2.0
5	wrxUii7i4F...	F	2001-12-2...	70.238356...	0.4794520...	1.0	2001-12-2...	63.0	1.8
6	jbyQRaNE...	F	2001-12-0...	77.068493...	0.7232876...	0.0	2001-12-2...	24.0	2.0
7	K0xRXUQ2...	M	2002-03-1...	65.901369...	1.2575342...	1.0	2002-03-3...	70.0	2.0
8	hd9NERvT...	M	2001-01-2...	78.153424...	0.4821917...	1.0	2001-02-0...	68.4	1.8
9	o0Fc+qseZ...	M	2002-09-0...	65.663013...	1.7260273...	1.0	2002-09-1...	24.0	7.0
10	H7kKW4Tx9...	M	2002-02-2...	85.463013...	0.7178082...	1.0	2002-03-0...	68.4	1.8
11	EsVy8Kfjmj...	F	2002-09-1...	77.093150...	1.0821917...	1.0	2002-09-2...	72.0	1.8
12	sYBjae4U...	M	2003-05-3...	76.498630...	0.1123287...	0.0	2003-06-0...	68.4	1.8
13	zEauRvbxIY...	M	2003-09-0...	66.013698...	0.6273972...	0.0	2003-09-1...	72.0	1.2
14	DTJMjdJ4y...	M	2001-08-0...	77.504109...	2.6657534...	1.0	2001-08-1...	28.0	7.0
15	0SSQutEp...	M	2002-12-3...	62.452054...	0.9534246...	1.0	2003-01-1...	81.6	1.2
16	Vkw9jeXRT...	F	2002-05-0...	82.460273...	0.3698630...	1.0	2002-05-0...	60.0	2.0
17	EH6i72Oe...	M	2003-09-2...	75.986301...	0.2219178...	1.0	2003-09-2...	68.4	1.8
18	/o12FfCSA...	F	2003-07-0...	86.931506...	0.9068493...	1.0	2003-07-1...	28.0	7.0
19	gP0/Ar20T...	F	2003-11-0...	88.915068...	0.6191780...	1.0	2003-11-2...	40.0	8.0
20	xORsOWrL...	M	2000-09-0...	66.435616...	0.2438356...	1.0	2000-09-0...	72.0	1.8
21	NHIL9MzS...	M	2002-08-1...	82.328767...	1.6054794...	0.0	2002-08-2...	63.0	1.8
22	SOMXfvd+h...	F	2001-08-2...	75.446575...	1.3972602...	1.0	2001-09-1...	69.6	1.2
23	lo2Gj8eMN...	F	2001-11-0...	71.673972...	0.8602739...	1.0	2001-11-2...	70.0	2.0
24	ossZ4v2W...	M	2002-11-2...	74.493150...	0.9342465...	1.0	2002-12-1...	63.0	1.8
25	149qo9i1q...	M	2000-09-1...	70.515068...	0.9315068...	1.0	2000-09-2...	70.0	2.0
26	G3Amcv+p...	M	2001-03-0...	60.150684...	0.9315068...	0.0	2001-03-1...	68.4	1.8
27	QXlyEvmjv...	F	2003-01-2...	79.279452...	0.9506849...	1.0	2003-02-0...	24.0	7.0
28	Kdaz7YtgG...	F	2000-08-1...	74.490410...	0.7068493...	1.0	2000-08-2...	70.0	2.0
29	cObMe0QJ...	M	2003-07-0...	74.230136...	0.8054794...	0.0	2003-07-1...	24.0	7.0
30	utGUmGd+...	M	2002-02-1...	74.156164...	0.9589041...	1.0	2002-02-1...	70.0	2.0
31	zxEbBBIUaf...	M	2002-08-1...	78.717808...	0.9671232...	1.0	2002-08-2...	24.0	2.0
32	7IIUG4lyNo...	M	2001-01-1...	61.452054...	0.8767123...	1.0	2001-02-0...	70.0	2.0
33	62FtbzRqsf...	M	2001-05-3...	72.383561...	0.9315068...	1.0	2001-06-0...	68.4	1.8
34	MOY5trHL...	M	2003-04-1...	66.219178...	0.7780821...	0.0	2003-04-2...	28.0	7.0
35	Yv9mkiyt52...	M	2001-07-1...	64.887671...	1.3835616...	1.0	2001-07-2...	70.0	2.0
36	Et0mnrVws...	F	2001-01-0...	68.364383...	2.2849315...	1.0	2001-01-1...	40.0	15.0
37	/lscteZvifa...	M	2000-04-0...	75.334246...	1.2739726...	0.0	2000-04-1...	63.0	1.8
38	jrmZkkDQ...	F	2002-03-1...	52.416438...	7.6712328...	1.0	2002-04-0...	28.0	7.0
39	iMhcNDQu...	F	2002-02-1...	69.950684...	1.6191780...	0.0	2002-02-2...	70.0	2.0
40	hON6PkMx...	F	2001-05-1...	60.854794...	0.3095890...	1.0	2001-05-2...	40.0	15.0
41	+5mv1Ck2...	M	2000-08-1...	73.263013...	2.1506849...	0.0	2000-08-1...	70.0	2.0
42	uuge4IZWi...	M	2003-04-1...	74.852054...	0.4520547...	1.0	2003-04-2...	68.4	1.8
43	21ewz81jN...	M	2000-06-1...	67.819178...	1.6191780...	1.0	2000-06-1...	72.0	1.8
44	oU+mjenif...	M	2002-03-1...	79.030136...	0.5068493...	1.0	2002-03-1...	36.0	15.0
45	6lfexxNTW...	M	2003-04-0...	77.589041...	1.0465753...	1.0	2003-04-1...	60.0	2.0
46	GJDm+O/5...	M	2001-08-3...	43.180821...	0.5095890...	1.0	2001-09-0...	24.0	7.0
47	xORsOWrL...	F	2000-08-1...	43.706849...	0.7780821...	0.0	2000-08-2...	72.0	1.8
48	64/Oaiaf4...	F	2001-03-0...	76.512328...	1.0794520...	1.0	2001-03-0...	70.0	2.0
49	bwYt8ytkYG...	M	2001-12-1...	65.484931...	1.2821917...	1.0	2001-12-1...	63.0	1.8
50	toCwn5ul...	F	2000-03-0...	67.161643...	1.5945205...	1.0	2000-03-0...	60.0	2.0
51	yAdNfO7O+...	F	2002-04-0...	75.789041...	0.8054794...	0.0	2002-04-1...	24.0	2.0
52	T7SJEJP1S...	M	2002-01-1...	65.865753...	1.3561643...	1.0	2002-01-2...	24.0	7.0
53	4n3GzP04v...	M	2001-02-1...	77.454794...	2.1506849...	1.0	2001-02-2...	81.6	1.2
54	HUL57ykd...	F	2001-11-2...	76.498630...	1.8739726...	1.0	2001-12-0...	60.0	2.0
55	N67HIGP...	F	2002-12-1...	66.642925...	0.5242465...	1.0	2002-01-0...	72.0	1.2

HDI™ Case Study - Analyse

www.e-hrc.net



For example:

We would like to see if different radiotherapy regimes have an impact on survival of cancer patients.

1. Query on linked MUD clinical cancer information database and RAD radiotherapy database to get a dataset of information on cancer patients, their current status and their radiotherapy treatments.
2. Use this set of data to perform a Kaplan-Meier survival analysis – investigating if different types of treatment have different survival outcomes.



- Analysis Methods
 - Survival Analysis
 - Nonparametric
 - Kaplan-Meier

Fit Kaplan-Meier Survival Curves

Select the data source:
demo.csv

Select the variable indicating end time:
YearstoLastContact

Select the variable indicating survival status:
DeathIndicatorNumeric

Select variables to be treated as factor(s):

TreatmentDate	Add >>	Fractionation
Dose		
Surname	<< Delete	
Gender		
DateofDiagnosis		
AgeatDiagnosis		

Select the level for two-sided confidence interval(s):

0.90 0.95 0.99

Compute the standard errors for the model fit:

TRUE FALSE

Type of survival curve to be fitted:

kaplan-meier

Formula for the error:

greenwood tsiatis

The way confidence intervals are to be calculated:

none log log-log plain

Select the way the lower confidence interval is to be modified:

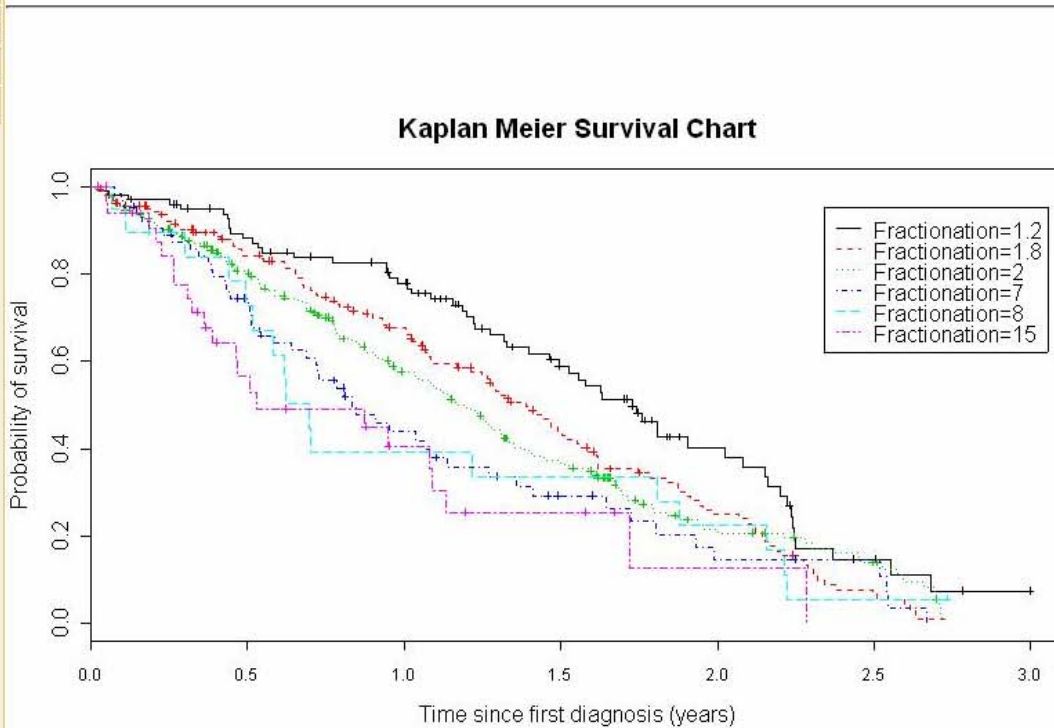
usual

Specify a title for the survival chart:

Kaplan Meier Survival Chart



- Analysis Methods
 - Survival Analysis
 - Nonparametric
 - Kaplan-Meier



Summary of Survival Analysis

```
Call: survfit(formula = survfitformula, data = survfitdata, na.action = na.action,
  conf.int = conf.int, se.fit = se.fit, type = type, error = error,
  conf.type = conf.type, conf.lower = conf.lower)
```

Fractionation=1.2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	66	20	0.779	0.0438	0.698	0.870
2	18	25	0.404	0.0606	0.301	0.542

Fractionation=1.8

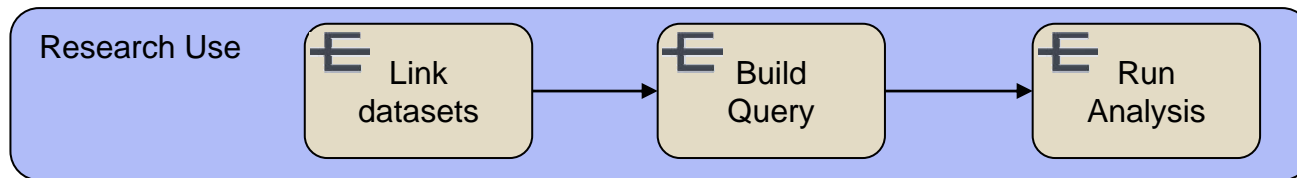
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	84	45	0.676	0.0400	0.602	0.759
2	24	46	0.250	0.0419	0.180	0.347

Fractionation=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
------	--------	---------	----------	---------	--------------	--------------

HDI™ Case Study - Report

www.e-hrc.net



For example:

For ongoing policy and practice management reporting, basic information on the treatments and intent of treatment for each stage and type of cancer is required. Statistics of types of disease, their risk factors and related severity are also commonly monitored. As these types of reporting are run on a regular basis, standard reports can be developed for this purpose.

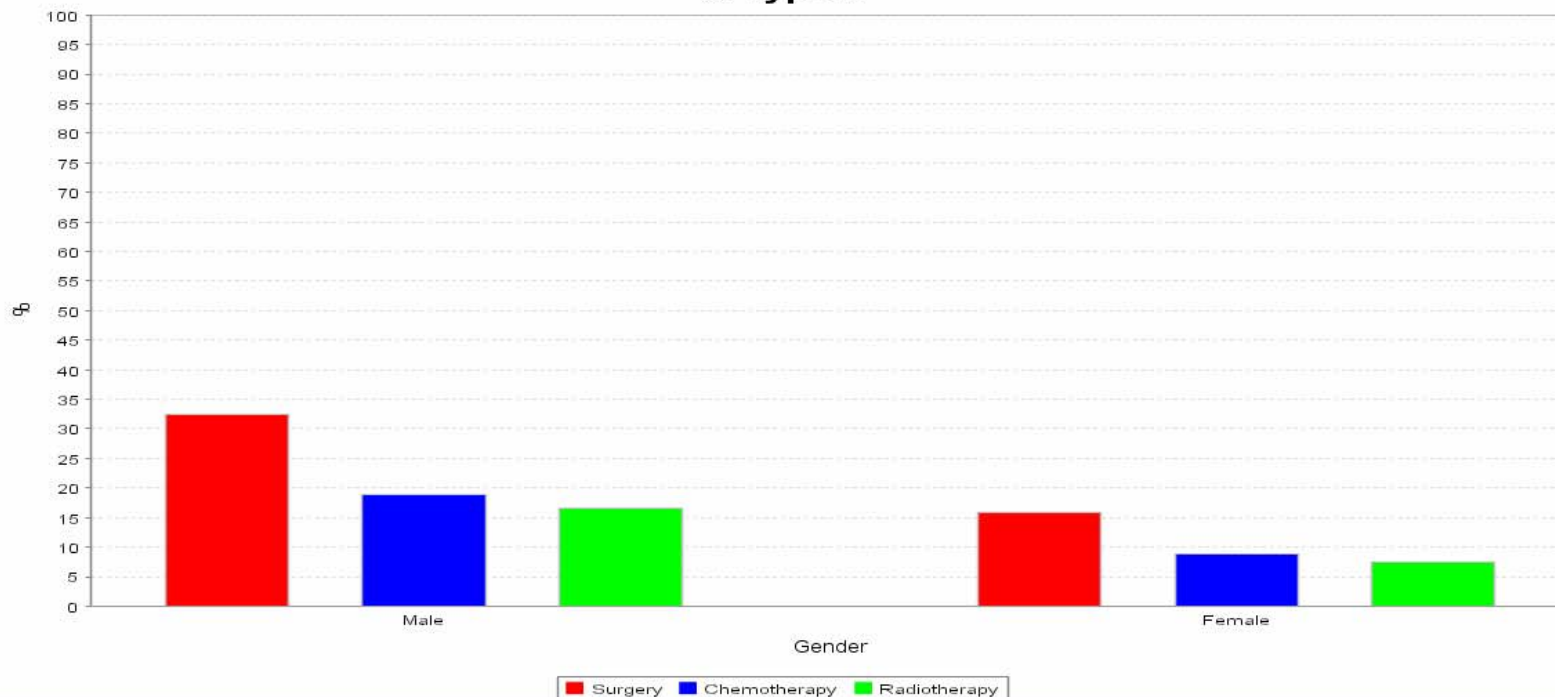
Run a standard report, and interrogate the data as required through the provided variables.



Quick Reports

- Demonstration Reports
 - Percentage Reports
 - % Curative vs Palliative
 - % Different Treatments
 - % Type L
 - % Type M
 - % Type B
 - Statistical Analysis
 - Count Reports

% Type M



		% Denominator	% Numerator	Percent
Surgery	Male	771	251	32.55
Surgery	Female	771	122	15.82
Chemotherapy	Male	771	145	18.8
Chemotherapy	Female	771	68	8.81
Radiotherapy	Male	771	128	16.6
Radiotherapy	Female	771	57	7.39

◆ Hospital = \${report.default.hospital}

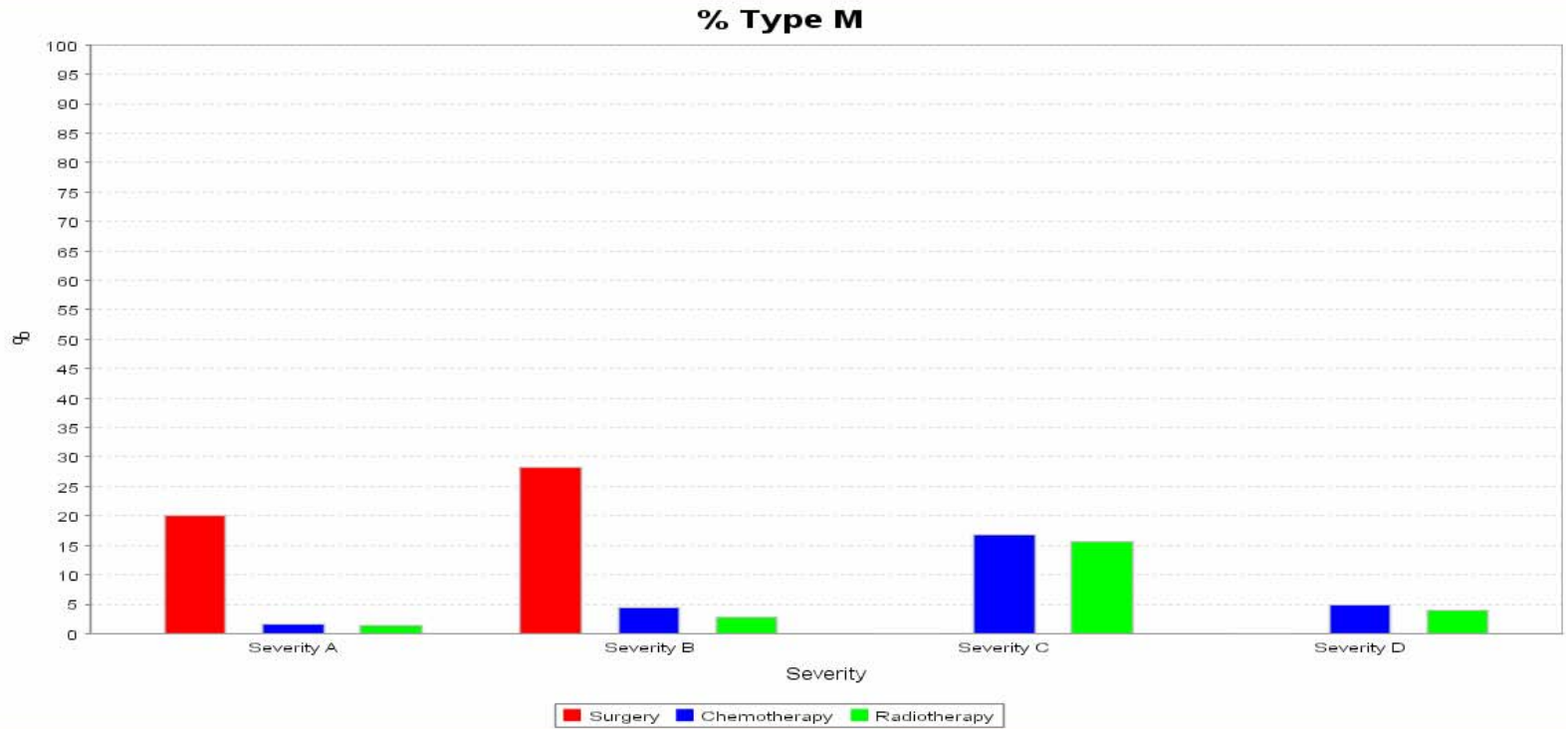
Axis Variable
 Gender

Legend Variable
 Treatment



Quick Reports

- Demonstration Reports
 - Percentage Reports
 - % Curative vs Palliative
 - % Different Treatments
 - % Type L
 - % Type M
 - % Type B
 - Statistical Analysis
 - Kaplan Meier Survival A
 - Type L Kaplan Meier S
 - Type M Kaplan Meier S
 - Type B Kaplan Meier S
 - Count Reports



		% Denominator	% Numerator	Percent
Surgery	Severity A	771	155	20.1
Surgery	Severity B	771	218	28.27
Surgery	Severity C	771	0	0.0
Surgery	Severity D	771	0	0.0
Chemotherapy	Severity A	771	13	1.68
Chemotherapy	Severity B	771	34	4.4
Chemotherapy	Severity C	771	129	16.73
Chemotherapy	Severity D	771	37	4.79

◆ Hospital = \${report.default.hospital}

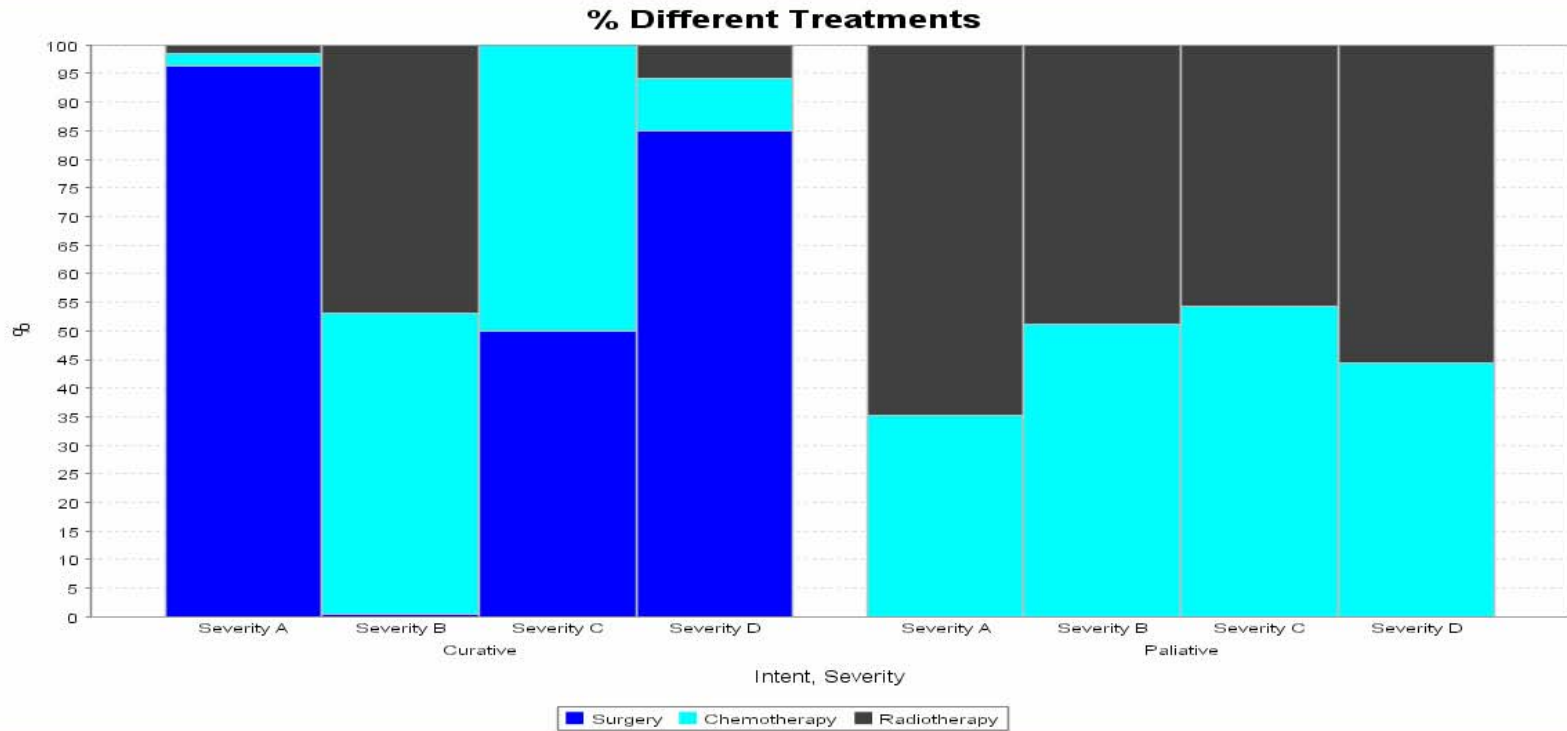
Axis Variable
Severity

Legend Variable
Treatment



Quick Reports

- Demonstration Reports
 - Percentage Reports
 - % Curative vs Palliative
 - % Different Treatments
 - % Type L
 - % Type M
 - % Type B
 - Statistical Analysis
 - Kaplan Meier Survival A
 - Type L Kaplan Meier S
 - Type M Kaplan Meier S
 - Type B Kaplan Meier S
 - Count Reports



			Total (N)	No. Receiving Treatment Type (n)	Percent Receiving Treatment Ty...
Curative	Severity A	Surgery	690	665	96.37
Curative	Severity A	Chemotherapy	690	14	2.02
Curative	Severity A	Radiotherapy	690	11	1.59
Curative	Severity B	Surgery	1050	892	84.95
Curative	Severity B	Chemotherapy	1050	97	9.23
Curative	Severity B	Radiotherapy	1050	61	5.8
Curative	Severity C	Surgery	758	3	0.39
Curative	Severity C	Chemotherapy	758	400	52.77

◆ Hospital = \${report.default.hospital}

Axis Variable
Intent

Legend Variable
Severity

Future of Health Data Integration

- **Translational medicine**
 - How can research information more readily be used in clinical situations
- **Genomic data**
 - Personalized genomic information
- **Environmental**
- **Life style**
- **Geographic**

- The semantic web

*The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.*

- Grid Technologies

- Several large grid projects now running in Europe to made data and services available

Future of Health Data Integration

www.e-hrc.net

- The aim is to extract more from current and future data and data sources for
 - Information to understand the factors affecting people's health
 - and knowledge to improve the system