

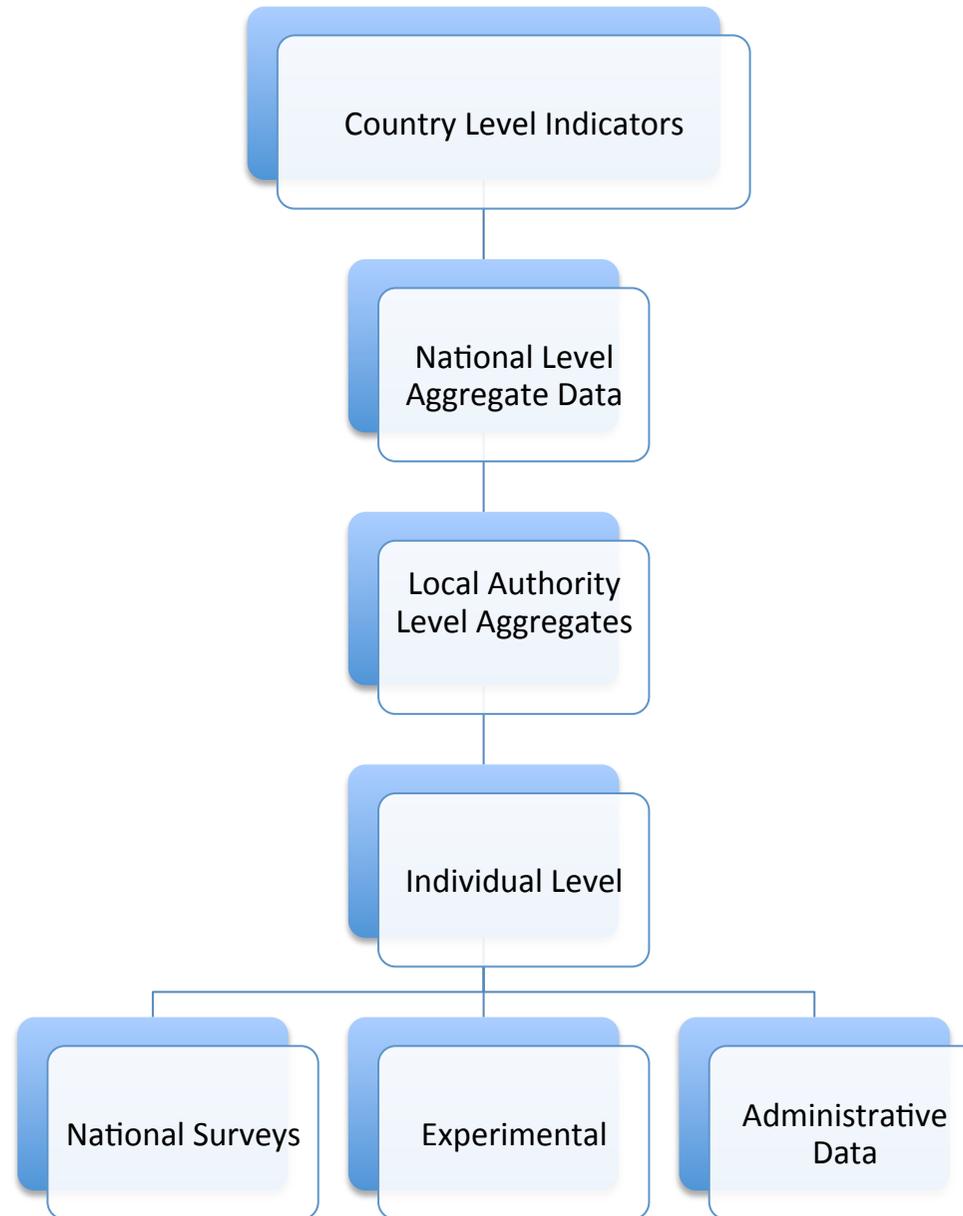


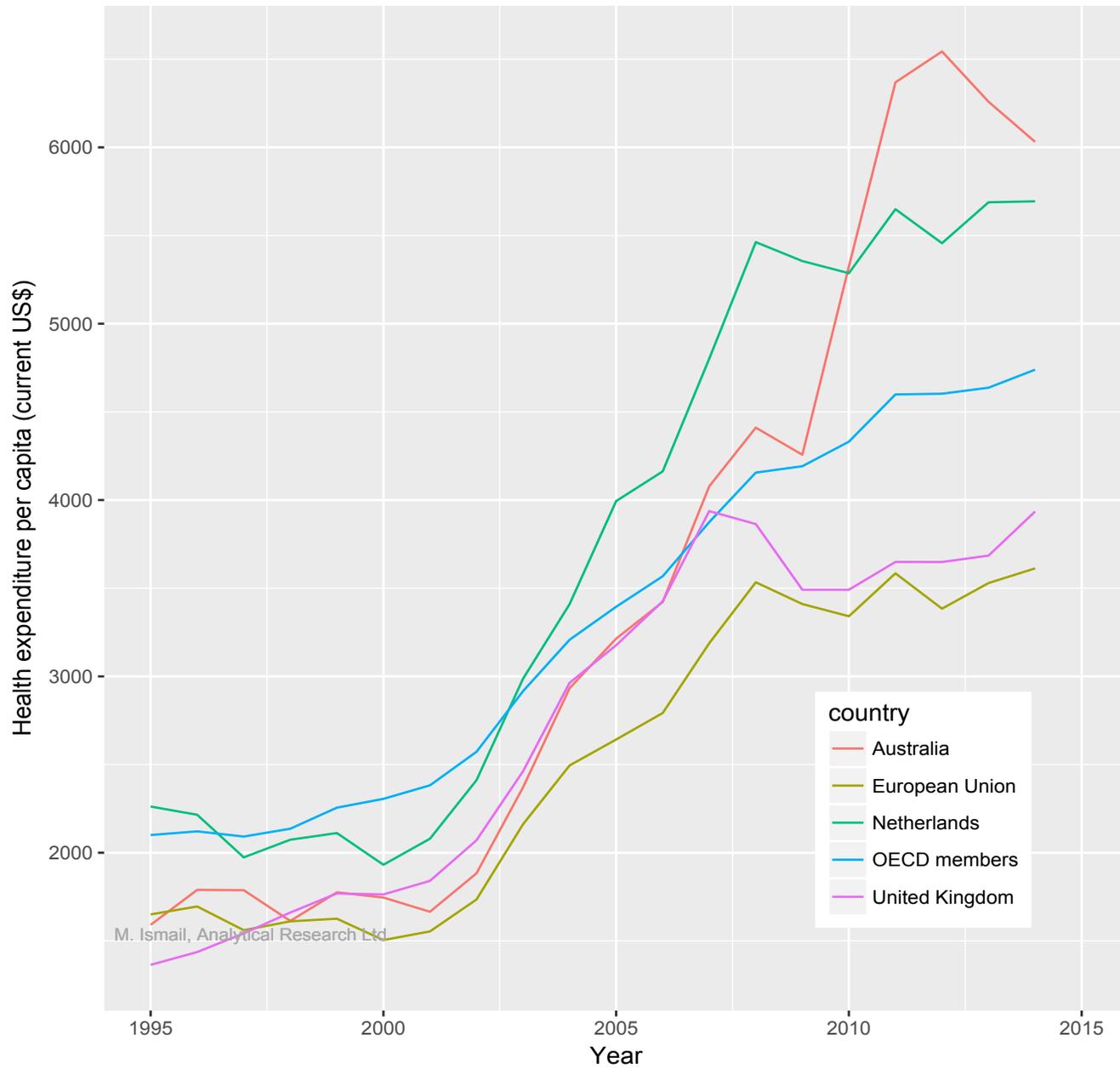
Designing Accessible and Useable Data for Researchers

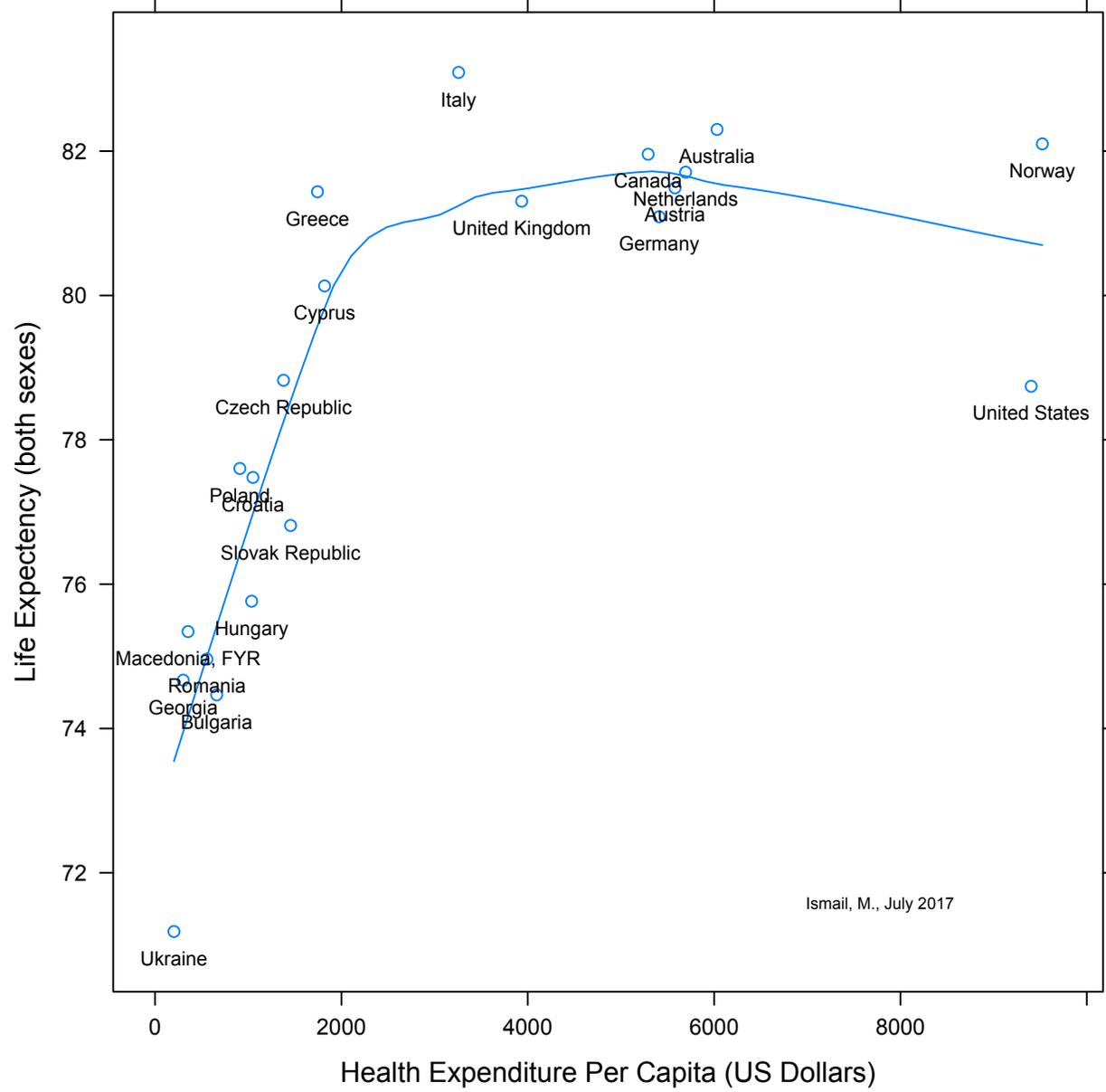
Mohamed Ismail
Analytical Research Ltd, UK

Introduction

- This presentation aims to:
 - summaries the different types of data for social policy use.
 - highlight how to make data available for researchers.
- It uses case studies of practical applications to illustrate the potential and limitations of different data types.
- It discusses the benefits of early considerations for data collection and design, in particular, administrative data.

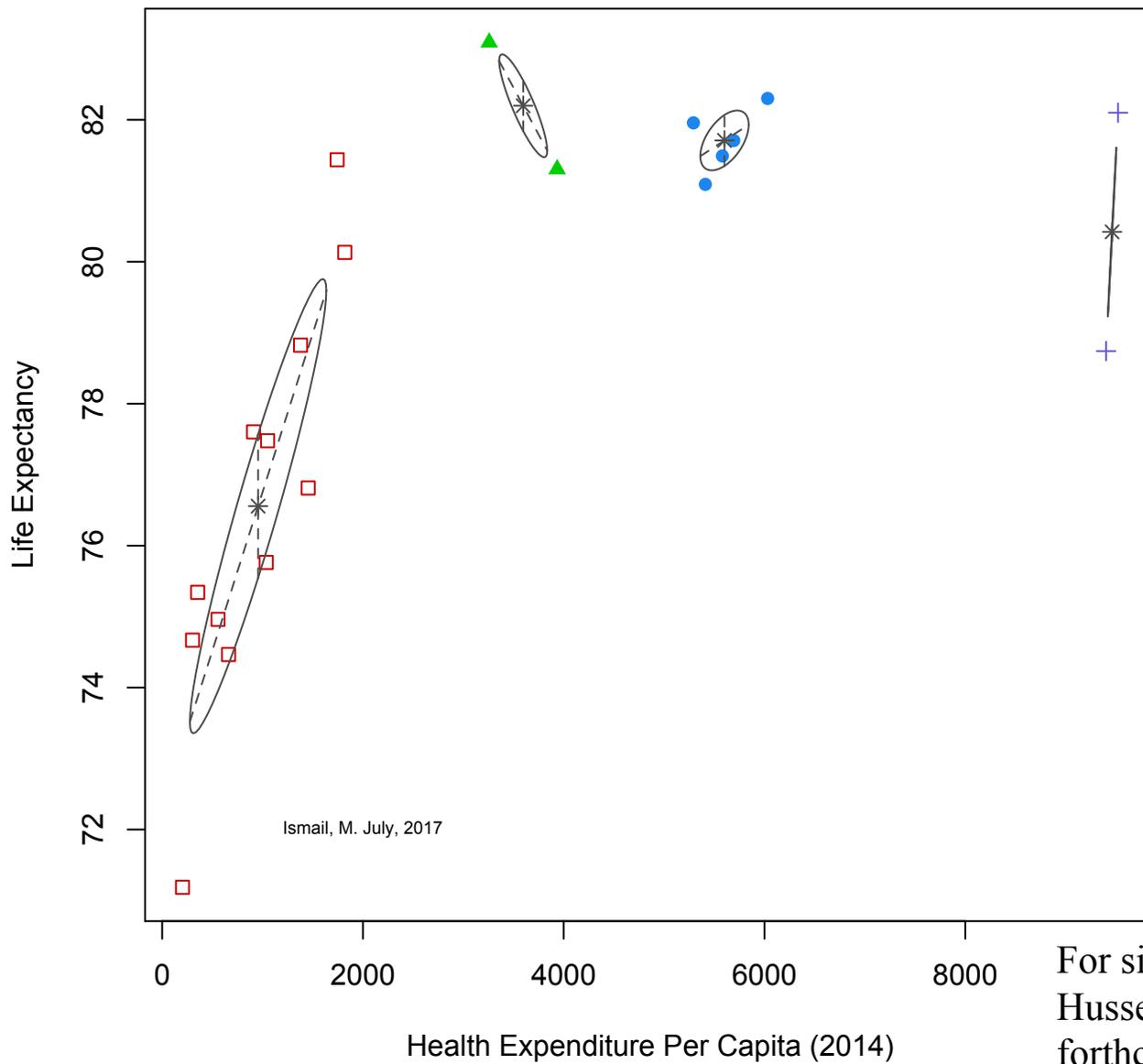






Ismail, M., July 2017

Model-Based Clustering According to Health Expenditure Per Capita and Life Expectancy



Ismail, M. July, 2017

For similar analysis, see Hussein and Ismail, 2016, forthcoming.

National Level Aggregate Data



The screenshot shows the top portion of the StatBank Norway website. At the top left is the logo for Statistisk sentralbyrå (Statistics Norway). To the right of the logo are links for 'A A A', 'NORSK', 'COOKIES AND PRIVACY', 'KEYWORDS A-Z', and 'CONTACT US'. Below these is a search bar with a magnifying glass icon and a 'SEARCH' button. A navigation bar below the search bar contains links: '> STATISTICS', '> RESEARCH', '> DATA COLLECTION', '> ABOUT STATISTICS NORWAY', and '> MY PAGE'.

[Home](#) > StatBank Norway

StatBank Norway

StatBank contains detailed tables with time series. You can create your own selections and save these in different file formats. We also offer an API for Statistikkbanken.

StatBank is updated at 05:00 and 11:30 (CET), during which time the tables are temporarily unavailable for up to five minutes. Figures which are being revised emerge as '0' or '-' between 05:00 and 08:00.

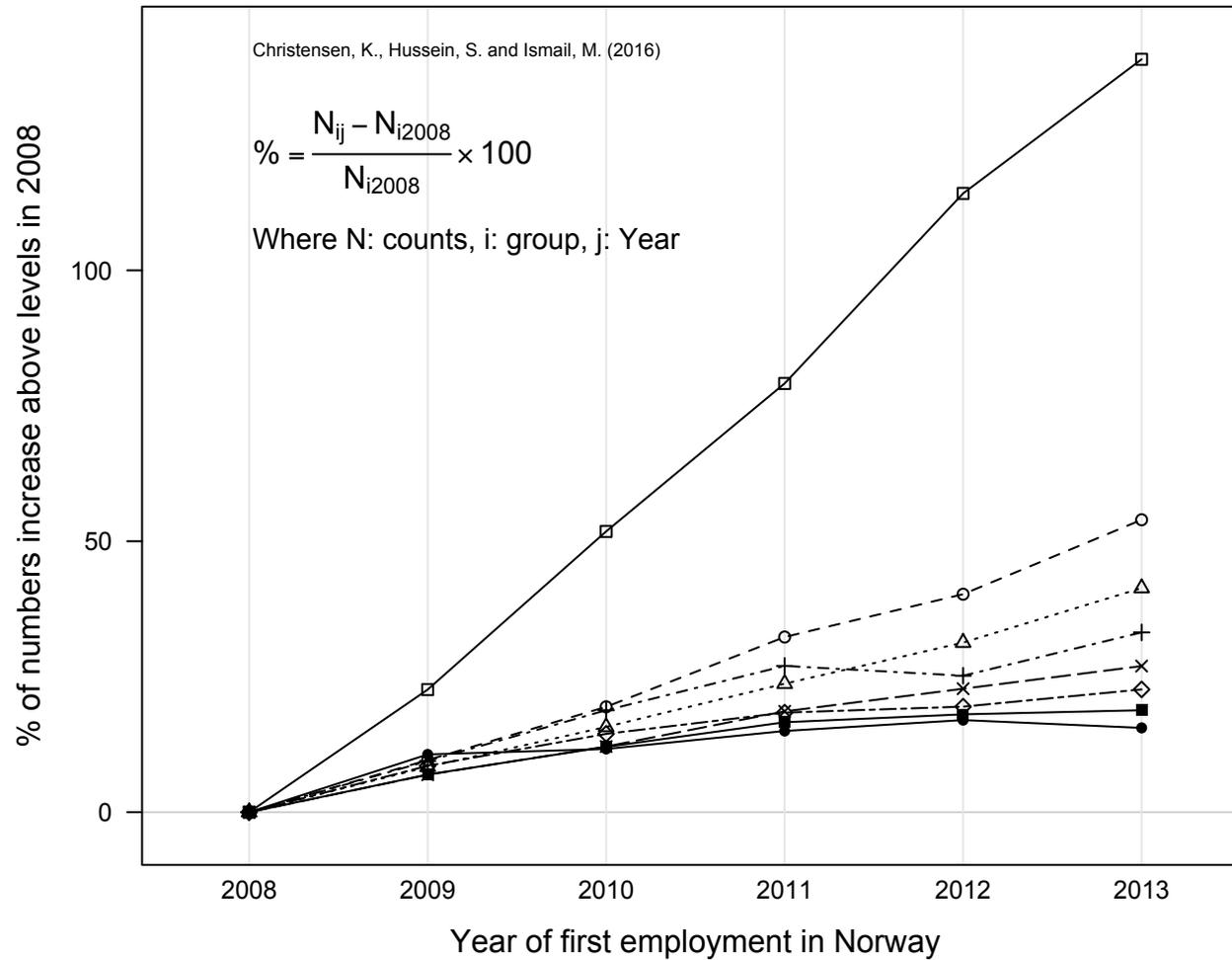
Choose topic and table

- [Open data with API](#)
- [How to use StatBank Norway](#)
- [Changes to tables in Statbank](#)

- + Agriculture, forestry, hunting and fishing
- + Banking and financial markets
- + Construction, housing and property
- + Culture and recreation
- + Education
- + Elections
- + Energy and manufacturing
- + Establishments, enterprises and accounts
- + External economy

Nationality/Sending Country

- | | | | | | |
|---|--------------------------------|---------|---|---------------------------|---------|
| □ | EU countries in Eastern Europe | — | × | South and Central America | - - - - |
| ○ | Africa | - - - - | ◇ | Western Europe | - - - - |
| △ | Asia | ⋯ | ■ | The Nordic countries | — |
| + | Eastern Europe else | - - - - | ● | North America and Oceania | - - - - |



The Nordic countries: Denmark, Finland, Faroe Islands, Greenland, Iceland, Sweden.
 Western Europe: Andorra, Austria, Belgium, Cyprus, France, Germany, Gibraltar, Greece, Ireland, Italy, Liechtenstein, Luxembourg, Malta, Monaco, Netherland, Portugal, San Marino, Spain, Switzerland, United Kingdom, Vatican City.
 EU countries in Eastern Europe: Bulgaria, Estonia, Croatia, Czech Republic, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia.
 Eastern Europe else: Albania, Belarus, Bosnia and Herzegovina, Kosovo, Macedonia, Moldova, Montenegro, Russia, Serbia, Ukraine.

Case Study: Do Personal Budgets Increase the Risk of Abuse?

- Personal Budgets (PBs) have moved to the mainstream in adult social care in England.
- Impact on safeguarding is not clear.
- Abuse of Vulnerable Adults (AVA) data.
- Adult Social Care Combined Return (ASC-CAR) data.
- Summary data provided by English local councils at the local council level.
- Individual level data from three purposively selected councils (2,209 individual referral records).
- See Ismail, M. et al. (2017).

Challenges of Using AVA

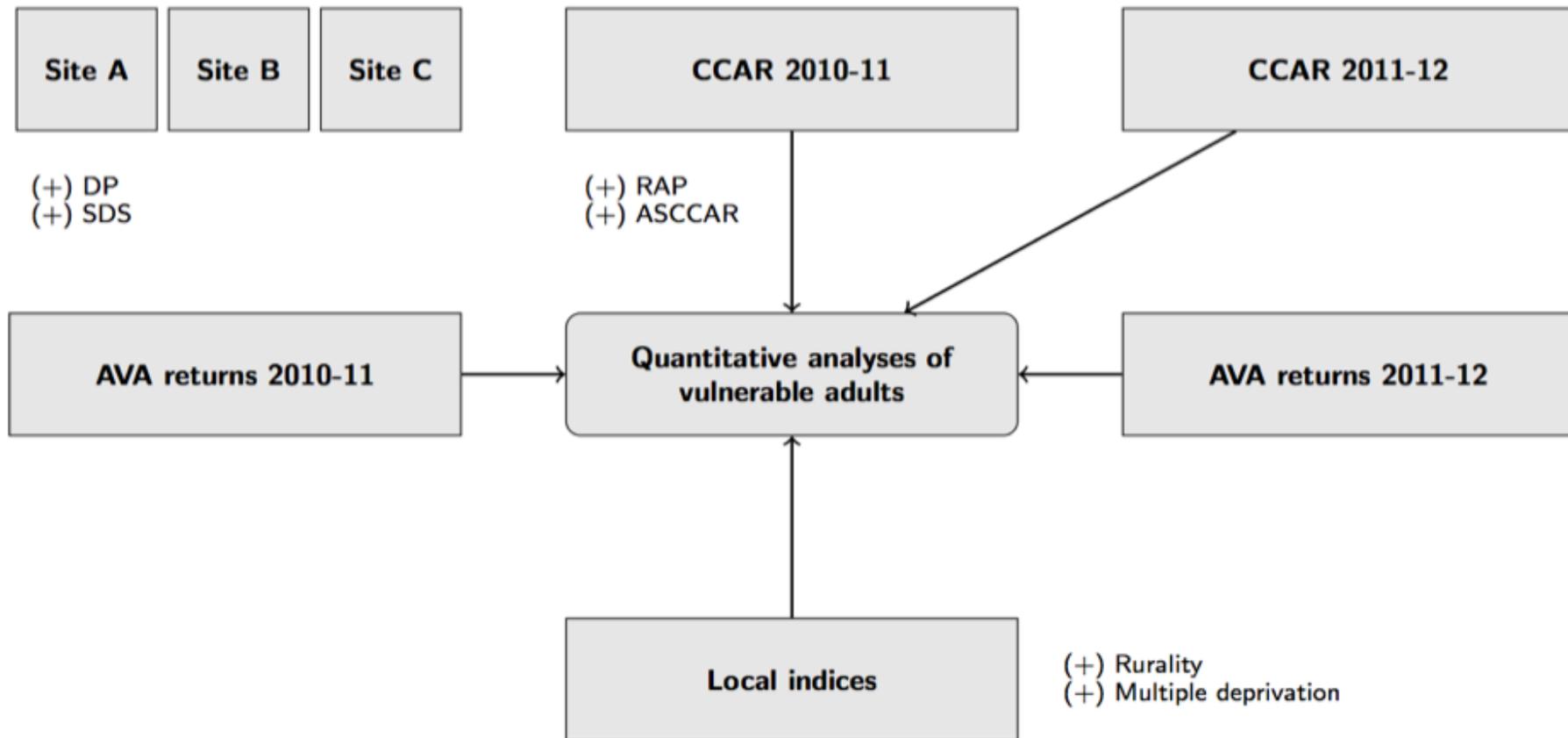
- Around 2000 variables.
- Mechanical variable names, making it very hard to communicate within a team.
- Potential for masking and hiding relationships and interaction between factors.
- Data quality issues, with counts rounded to the nearest 5.
- Technically challenging for uses other than calculating basic frequencies.
- Can only infer relationships at the local authority level.

Local Authority Level Aggregates

Year	ItemBaseID	LocationID	Council Code	102	104	106	107	108	109	110	111	112	113	114	116	117
			Council Name	Cumbria	Northumber	Gateshead	Newcastle up	North Tynesi	South Tynesi	Sunderland	Hartlepool U	Middlesbrou	Redcar and C	Stockton-on-	Durham	Darlington U
1112FY	5000101	AVA_tbl1_pg1_row1_col1		55	95	35	0	45	25	35	10	25	25	30	55	45
1112FY	5000102	AVA_tbl1_pg1_row1_col2		50	60	20	0	45	15	40	10	25	15	15	40	25
1112FY	5000103	AVA_tbl1_pg1_row1_col3		105	155	60	0	90	35	75	20	45	40	45	95	65
1112FY	5000104	AVA_tbl1_pg1_row1_col4		20	40	25	50	5	5	5	0	10	15	15	30	20
1112FY	5000105	AVA_tbl1_pg1_row1_col5		25	15	20	70	20	5	5	10	5	10	5	25	10
1112FY	5000106	AVA_tbl1_pg1_row1_col6		45	55	45	120	25	10	10	15	15	25	20	55	30
1112FY	5000107	AVA_tbl1_pg1_row1_col7		0	5	5	10	0	5	0	0	0	5	0	5	0
1112FY	5000108	AVA_tbl1_pg1_row1_col8		0	0	5	20	0	0	0	5	0	5	0	0	0
1112FY	5000109	AVA_tbl1_pg1_row1_col9		0	5	5	30	0	5	0	5	0	5	0	5	5
1112FY	5000110	AVA_tbl1_pg1_row1_col10		20	30	25	45	10	5	5	0	5	20	15	25	10
1112FY	5000111	AVA_tbl1_pg1_row1_col11		15	15	15	60	20	10	0	10	5	15	10	25	10
1112FY	5000112	AVA_tbl1_pg1_row1_col12		35	45	40	105	30	15	5	10	10	35	25	50	20
1112FY	5000113	AVA_tbl1_pg1_row2_col1	..		5	5	0	5	0	0	5	0	0	0	0	5
1112FY	5000114	AVA_tbl1_pg1_row2_col2	..		5	0	0	0	0	0	0	0	0	0	5	0
1112FY	5000115	AVA_tbl1_pg1_row2_col3	..		10	5	0	5	5	0	5	0	0	0	5	5
1112FY	5000116	AVA_tbl1_pg1_row2_col4	..		0	5	5	0	0	0	0	0	0	0	0	5
1112FY	5000117	AVA_tbl1_pg1_row2_col5	..		0	0	5	0	0	0	0	0	0	0	0	0
1112FY	5000118	AVA_tbl1_pg1_row2_col6	..		5	5	10	5	0	0	0	0	0	0	0	5
1112FY	5000119	AVA_tbl1_pg1_row2_col7	..		0	0	0	0	0	0	0	0	0	0	0	0
1112FY	5000120	AVA_tbl1_pg1_row2_col8	..		0	0	5	0	0	0	0	0	0	0	0	0
1112FY	5000121	AVA_tbl1_pg1_row2_col9	..		0	0	5	0	0	0	0	0	0	0	0	0
1112FY	5000122	AVA_tbl1_pg1_row2_col10	..		5	0	0	0	0	0	0	0	0	0	0	0
1112FY	5000123	AVA_tbl1_pg1_row2_col11	..		0	0	5	0	0	0	0	0	0	0	0	0
1112FY	5000124	AVA_tbl1_pg1_row2_col12	..		5	5	5	5	0	0	0	0	0	0	0	5
1112FY	5000125	AVA_tbl1_pg1_row3_col1		15	235	30	0	50	25	30	5	40	20	35	35	45
1112FY	5000126	AVA_tbl1_pg1_row3_col2		10	170	30	0	30	25	20	15	25	40	20	35	40
1112FY	5000127	AVA_tbl1_pg1_row3_col3		25	405	60	0	80	50	50	20	65	55	55	70	85
1112FY	5000128	AVA_tbl1_pg1_row3_col4		5	25	20	85	5	5	5	5	10	10	15	40	5
1112FY	5000129	AVA_tbl1_pg1_row3_col5		0	15	20	100	5	10	5	5	10	25	10	20	5
1112FY	5000130	AVA_tbl1_pg1_row3_col6		5	45	40	185	10	15	10	10	20	35	25	60	10
1112FY	5000131	AVA_tbl1_pg1_row3_col7		0	0	0	20	0	5	0	0	5	0	0	0	0
1112FY	5000132	AVA_tbl1_pg1_row3_col8		0	0	5	15	0	0	0	0	0	5	0	0	0
1112FY	5000133	AVA_tbl1_pg1_row3_col9		0	0	5	35	0	5	0	0	5	5	0	5	0
1112FY	5000134	AVA_tbl1_pg1_row3_col10		5	30	25	80	5	10	5	5	10	5	15	35	5
1112FY	5000135	AVA_tbl1_pg1_row3_col11		5	15	20	95	5	10	0	5	5	15	10	20	5
1112FY	5000136	AVA_tbl1_pg1_row3_col12		10	45	40	175	10	20	5	10	15	25	25	55	10
1112FY	5000137	AVA_tbl1_pg1_row4_col1	..		5	0	0	0	0	0	0	5	0	0	0	0
1112FY	5000138	AVA_tbl1_pg1_row4_col2	..		0	0	0	0	0	0	5	0	0	0	5	0
1112FY	5000139	AVA_tbl1_pg1_row4_col3	..		5	0	0	0	0	0	5	5	0	0	10	0
1112FY	5000140	AVA_tbl1_pg1_row4_col4	..		0	0	0	0	0	0	0	0	0	0	0	0

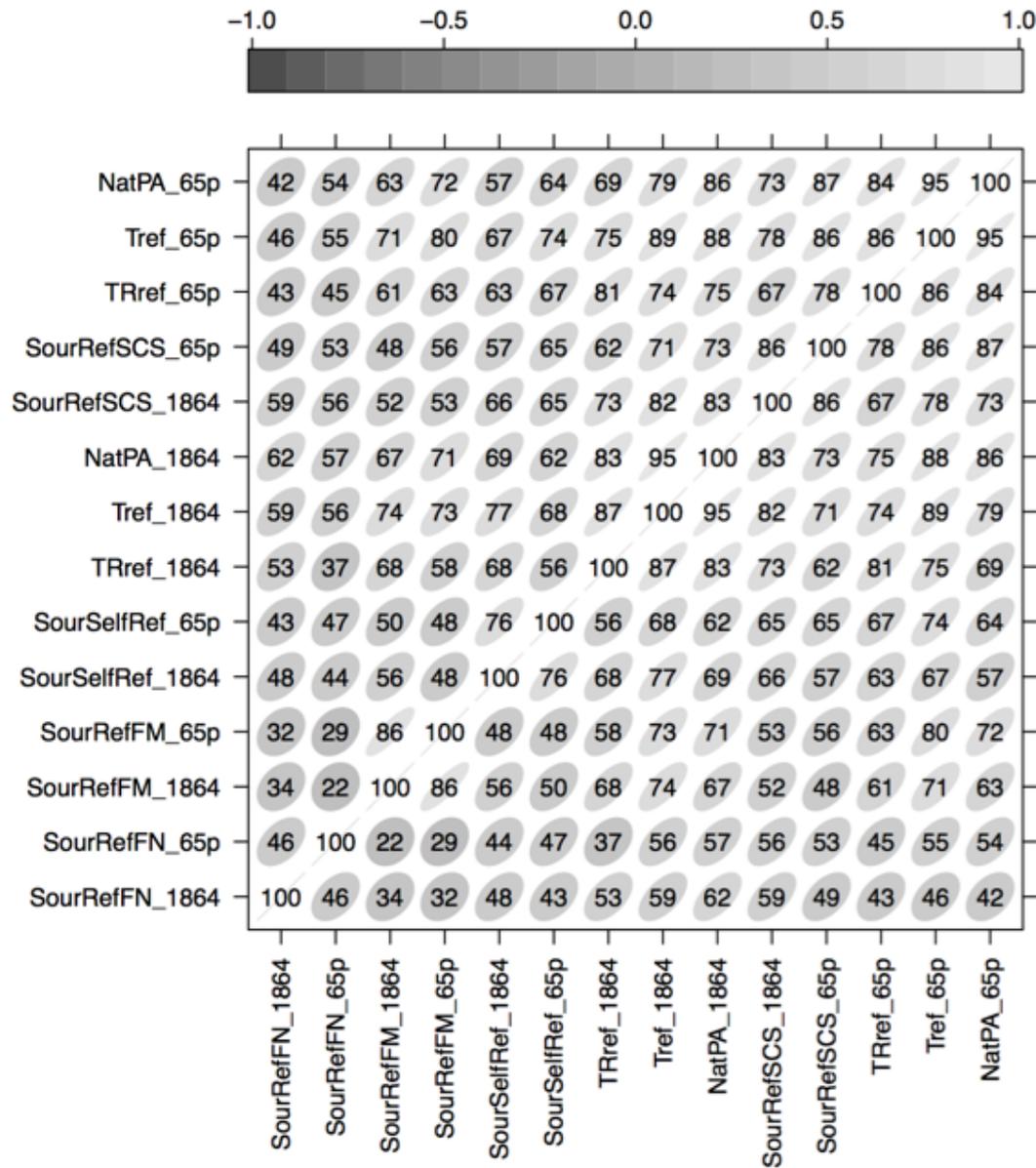
AVA RAW DATA STORAGE FORMAT

Data Sets and Scope



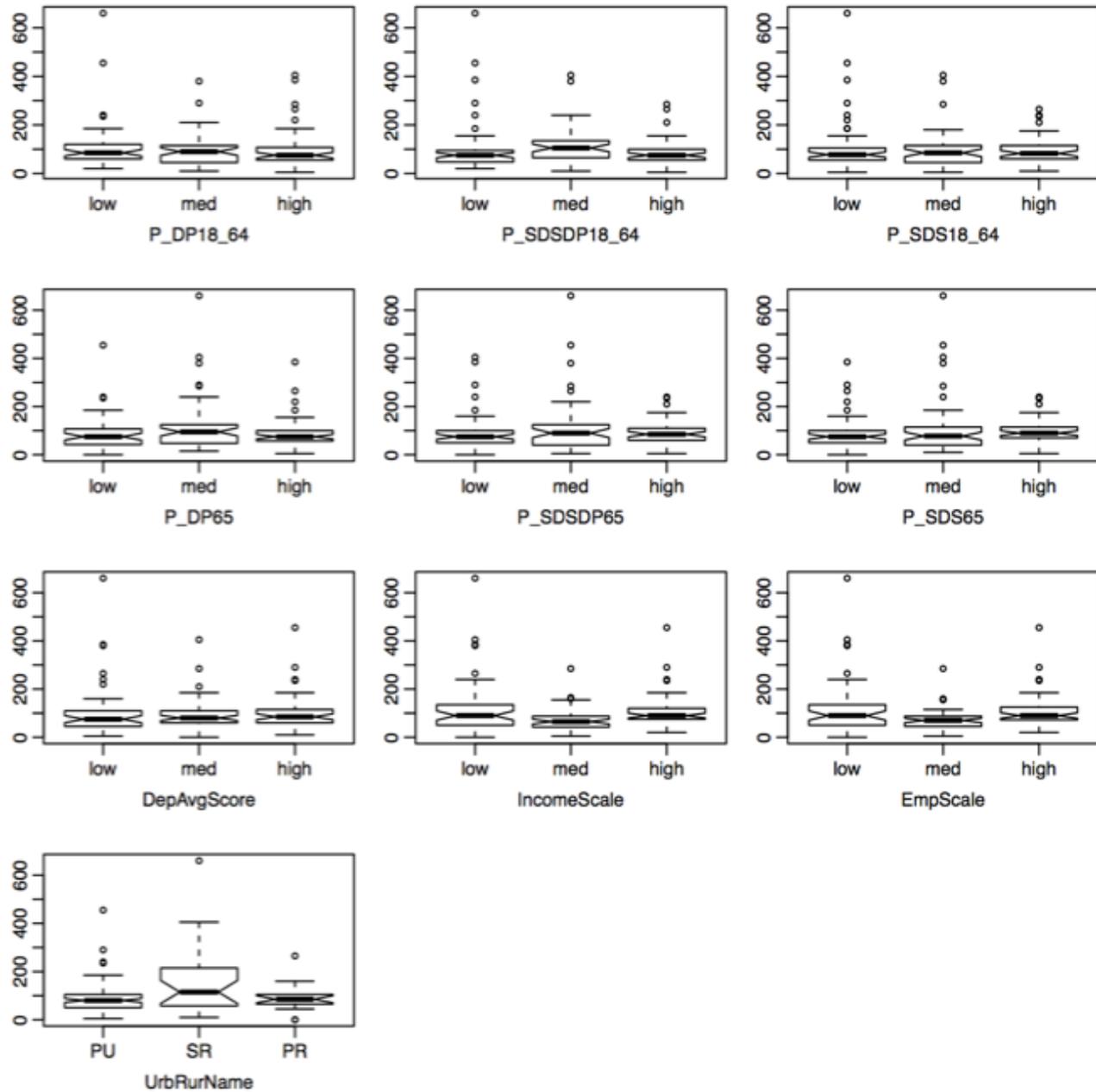
Approaching the Challenges

- Use 'proxy' variables to investigate uptake level of the different elements of personal budgets.
- Link to other local area characteristics such as deprivation level and rurality.
- Use visualisation techniques to select relevant variables.
- Complement analysis of national datasets with that of anonymous individual referral records.
- Data dictionary for intrateam communication.



An example of a correlation Matrix as a means for variable reduction, selecting 'unique' variables for inclusion

Employ data visualisation techniques to examine a huge volume of outputs



Individual Level Data

- They denote information about individuals, for example: age, address, education, etc.
- They are either contributed by the individuals themselves in surveys, censuses, etc. or are collected from registers.
- Usually anonymized.
- They allow for the most detailed analysis.

Individual Level: Administrative Data

- Data which are the result of the operation of administrative systems.
- Collected by government departments and agencies.
- Used for the purposes of registration, transactions and record keeping.
- Should be reliable due to their officiality.
- Not made by researchers.
- Not for research purpose.

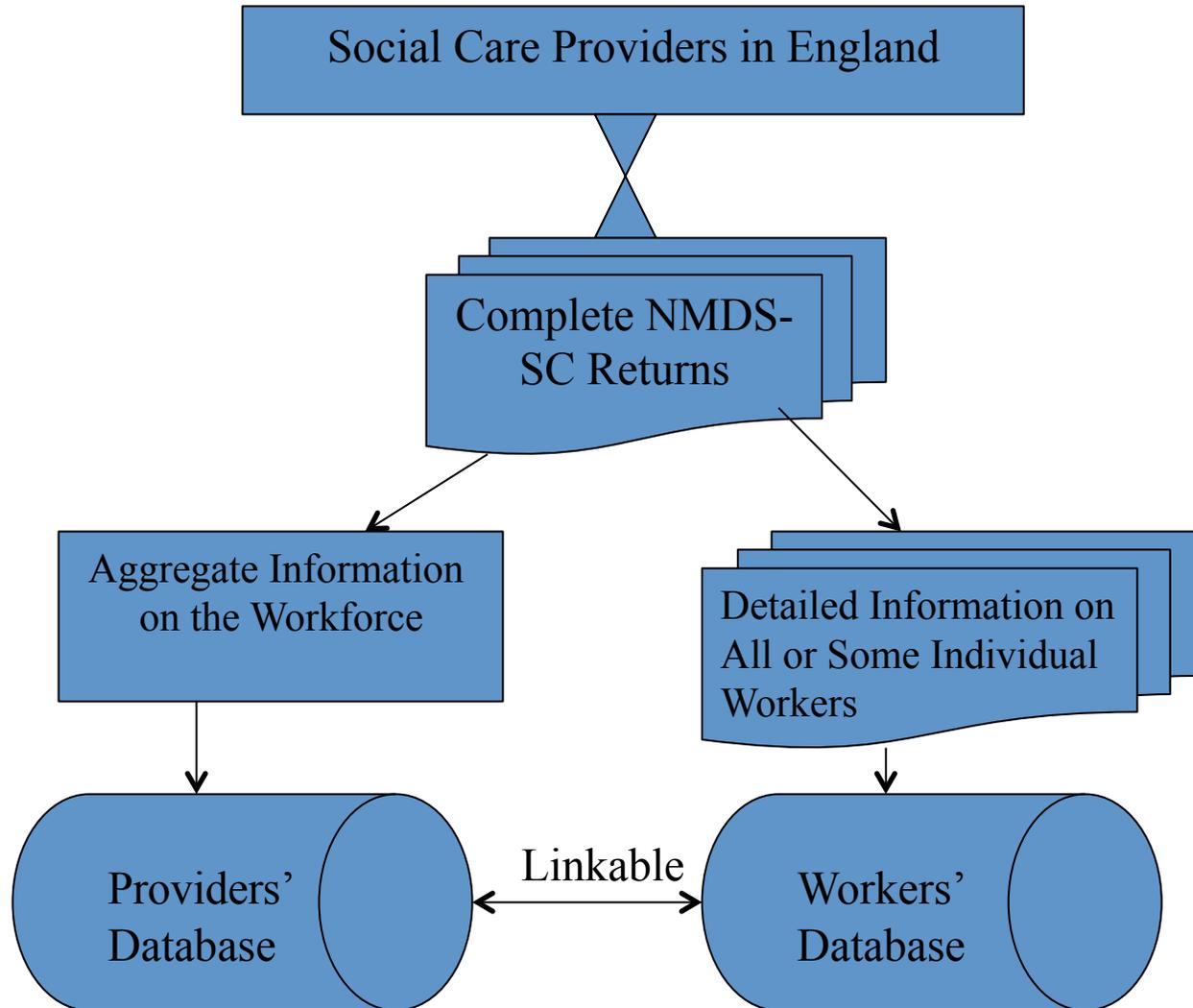
Advantages of Using Administrative Data in Research

- Regularly updated.
- Can provide historical information.
- Collected in a consistent way.
- Subject to quality checks.
- Near 100% coverage of population.
- Reliable.
- Potential for dataset linkage.
- No collection cost.

Disadvantages of Using Administrative Data in Research

- Proxy indicators may be required.
- May lack contextual/background information.
- Changes in definitions can be problematic.
- Missing or erroneous data.
- Omission of variables not deemed important by the administrator
- Metadata issues (may be lacking or of poor quality).
- Data protection issues.
- Access for researchers is dependent on the support of data providers.

Administrative Data Example: NMDS-SC



Snapshot Analysis Example: Male Workers in the Female-dominated Long-term Care Sector

Table 3. Results of fixed-effects of final mixed-effect model examining association between prevalence of men within organizations with different factors at three levels of hierarchy, NMDS-SC March 2011.

Results of final model	Lower 95% CI	β	Upper 95% CI	Standard error	<i>t</i> -Value (<i>F</i> -value)	<i>p</i> -Value
<i>Fixed effects</i>						
Organization size (ref.: micro)					(122.04)	<0.001
Small	-5.23	-3.38	-1.54	0.94	-8.67	<0.001
Medium	-6.45	-5.55	-4.10	1.00	-8.93	<0.001
Large	-8.75	-4.95	-1.16	2.14	-3.9	<0.001
Sector (ref.: LA)					(30.35)	<0.001
Private	2.45	3.38	4.32	0.48	7.10	<0.001
Voluntary	3.94	4.96	5.97	0.52	9.57	<0.001
Type of service (ref.: adult residential)					(69.52)	<0.001
Adult day	2.03	3.45	4.86	0.72	4.76	<0.001
Adult domiciliary	-5.64	-4.89	-4.13	0.39	-12.66	<0.001
Adult community	4.43	5.58	6.74	0.59	9.48	<0.001
Children	7.75	9.46	11.17	0.87	10.86	<0.001
Health/other	5.33	6.51	7.7	0.61	10.75	<0.001
Rurality (ref.: predominantly rural)					(6.07)	0.003
Predominantly urban	-0.03	1.25	2.54	0.65	1.92	0.060
Significant rural	-1.83	-0.48	0.87	0.68	-0.70	0.490
Service users group						
Adults with learning disabilities					(451.4)	<0.001
Yes	5.6	6.19	6.78	0.3	20.44	<0.001
Adults detailed mental health needs					(8.17)	0.004
Yes	-1.27	3.39	8.06	2.38	1.43	0.150
Adults with autistic spectrum disorder					(27.93)	<0.001
Yes	4.09	6.49	8.9	1.23	5.28	<0.001

see Hussein, S., Ismail M. & Manthorpe J. (2016)

Snapshot Analysis Example 2: Random-effects

Table 4. Results of random-effects of final mixed-effect model examining association between prevalence of men within organizations with different factors at three levels of hierarchy, NMDS-SC March 2011.

Random effects	Lower 95% CI	Estimated value	Upper 95% CI	Estimated variance	Overall variance (%)
Region	1.07	1.88	3.33	3.55	0.97
Local council	2.37	2.74	3.17	7.50	2.06
Organization (within group)	18.62	18.80	18.99	353.59	96.97
Total variance				364.64	

Longitudinal Analysis of Individual-level Records

- Example: using NMDS-SC providers' database for investigating workforce stability over time.
- Longitudinal changes in care workers' turnover and vacancy rates over time .
- From January 2008 to January 2010.
- Changes in reasons for leaving the sector, identified by employers.
- Differentiating between those with improved (reduced) turnover rates and those with worse (increased) turnover rates.

Challenges in Using Administrative Data for Longitudinal Analysis

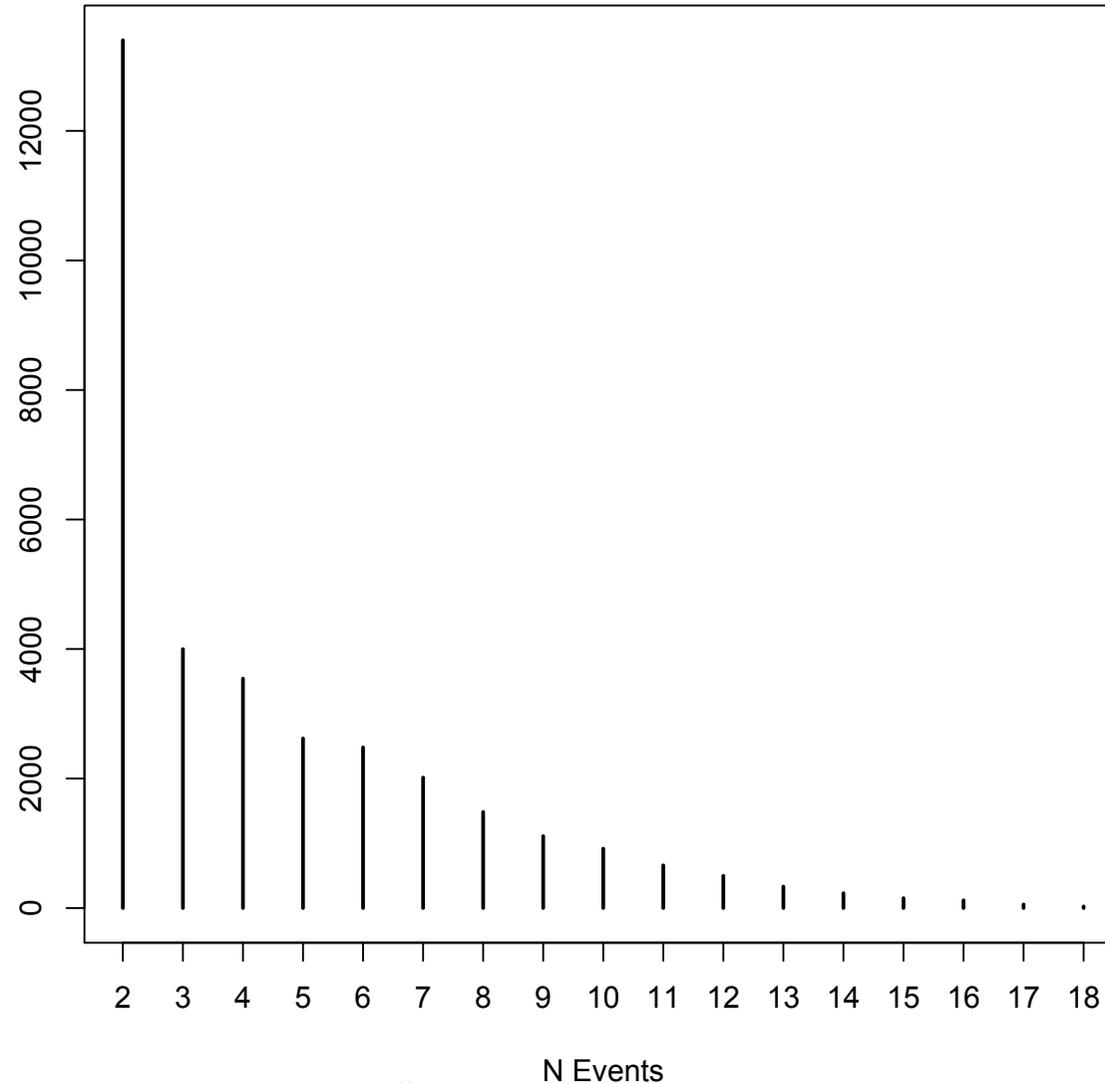
- No sampling framework.
- No regular intervals for data collection.
- Irregularities in data completion by different providers.
- Additions/alterations of variables and fields.
- Cumulative nature and consequences on data size and structure.
- Archiving.

Approaching the Challenges

- Mapping.
- Meta-data analysis.
- Use of graphical tools.
- Analysis based data extraction.

Meta-data Analysis: Providers with Different Number of Events

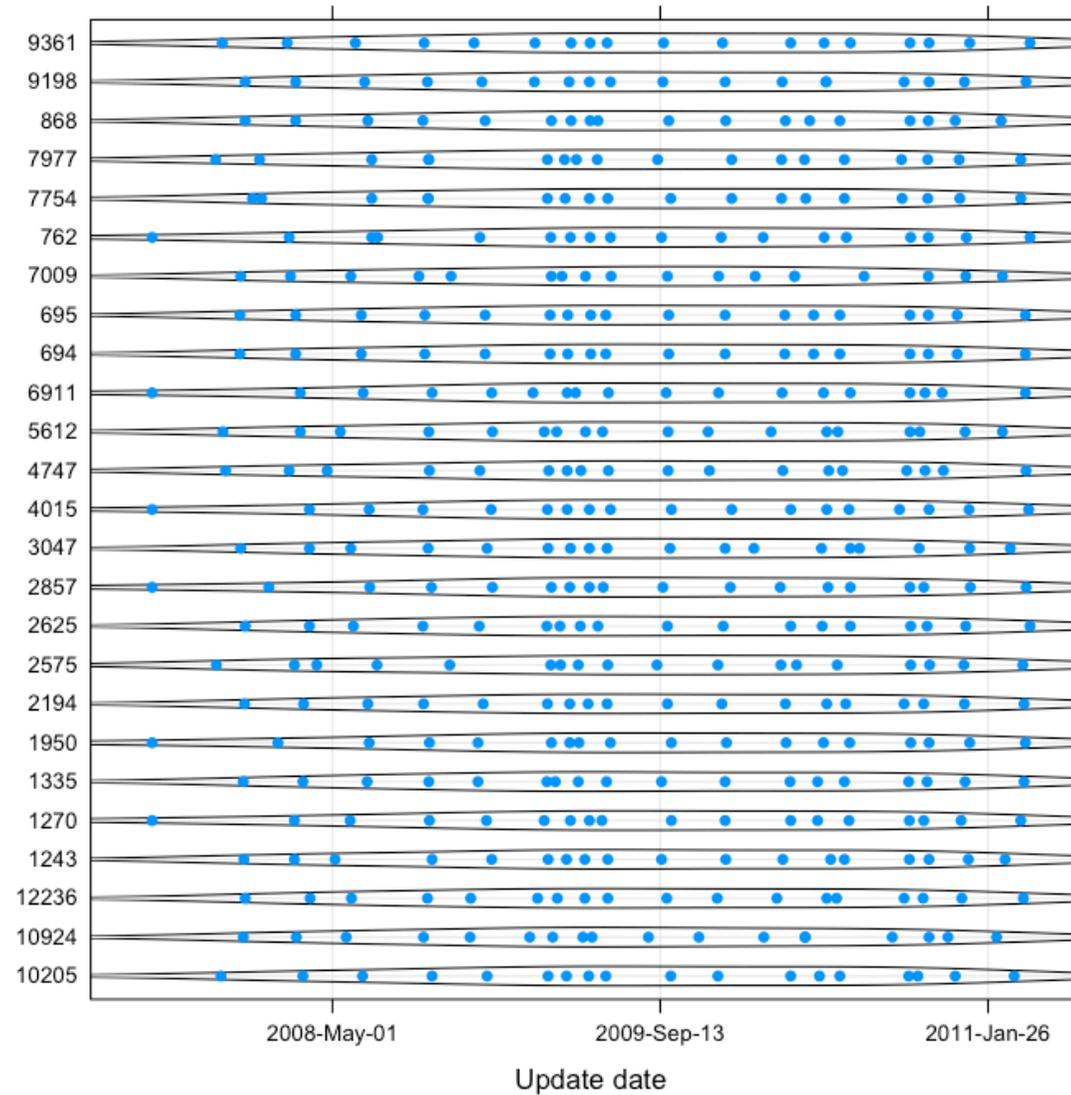
Updates	N provider	N this point down
2	13401	33643
3	3999	20242
4	3543	16243
5	2619	12700
6	2481	10081
7	2015	7600
8	1484	5585
9	1111	4101
10	918	2990
11	660	2072
12	498	1412
13	332	914
14	229	582
15	153	353
16	119	200
17	56	81
18	25	



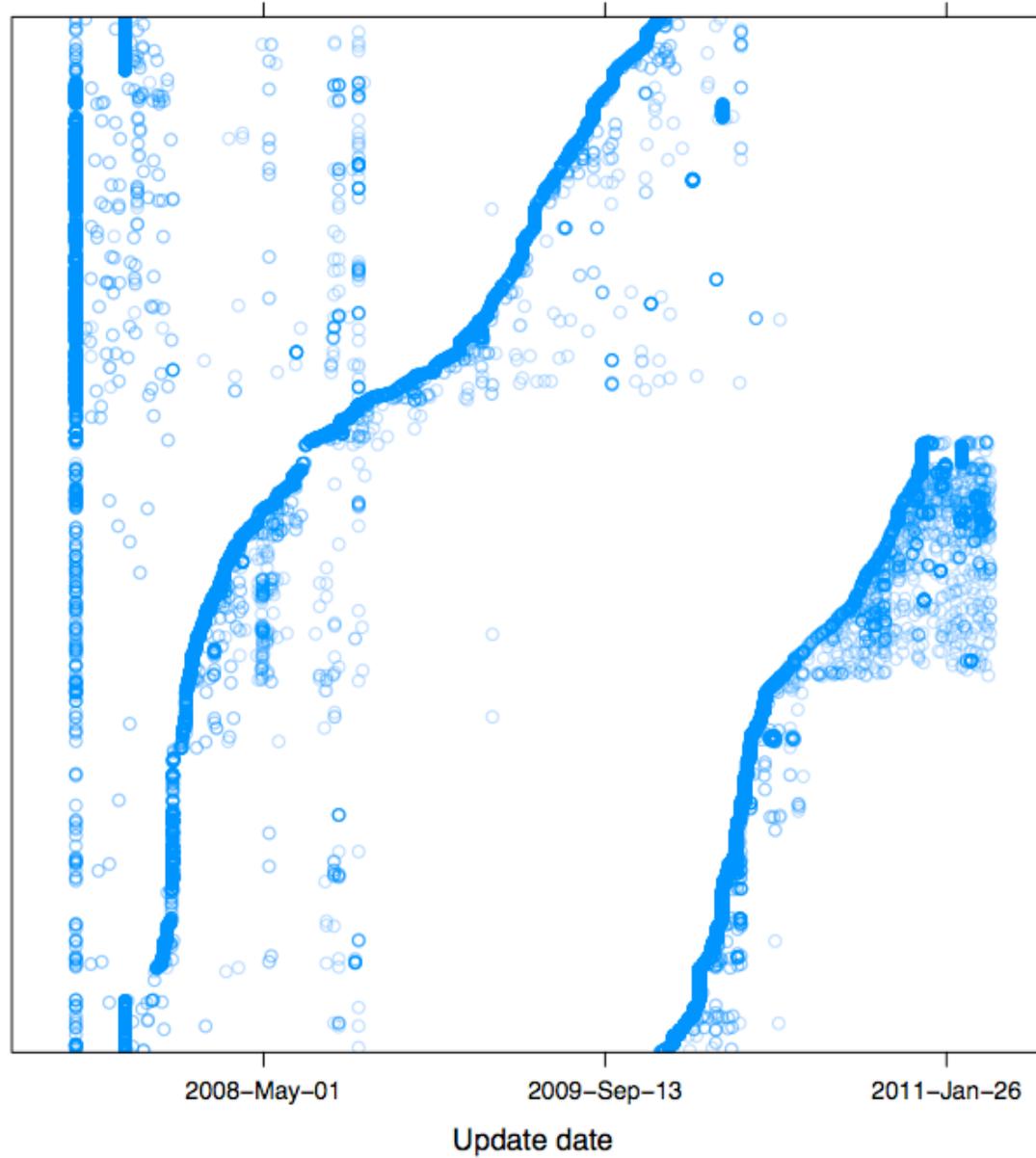
Aug, 2017

FSSI, Melbourne

Specific Example: Providers with 18 Updates



Specific Example 2: Providers with 2 Updates



Providers' Level Longitudinal Mapping

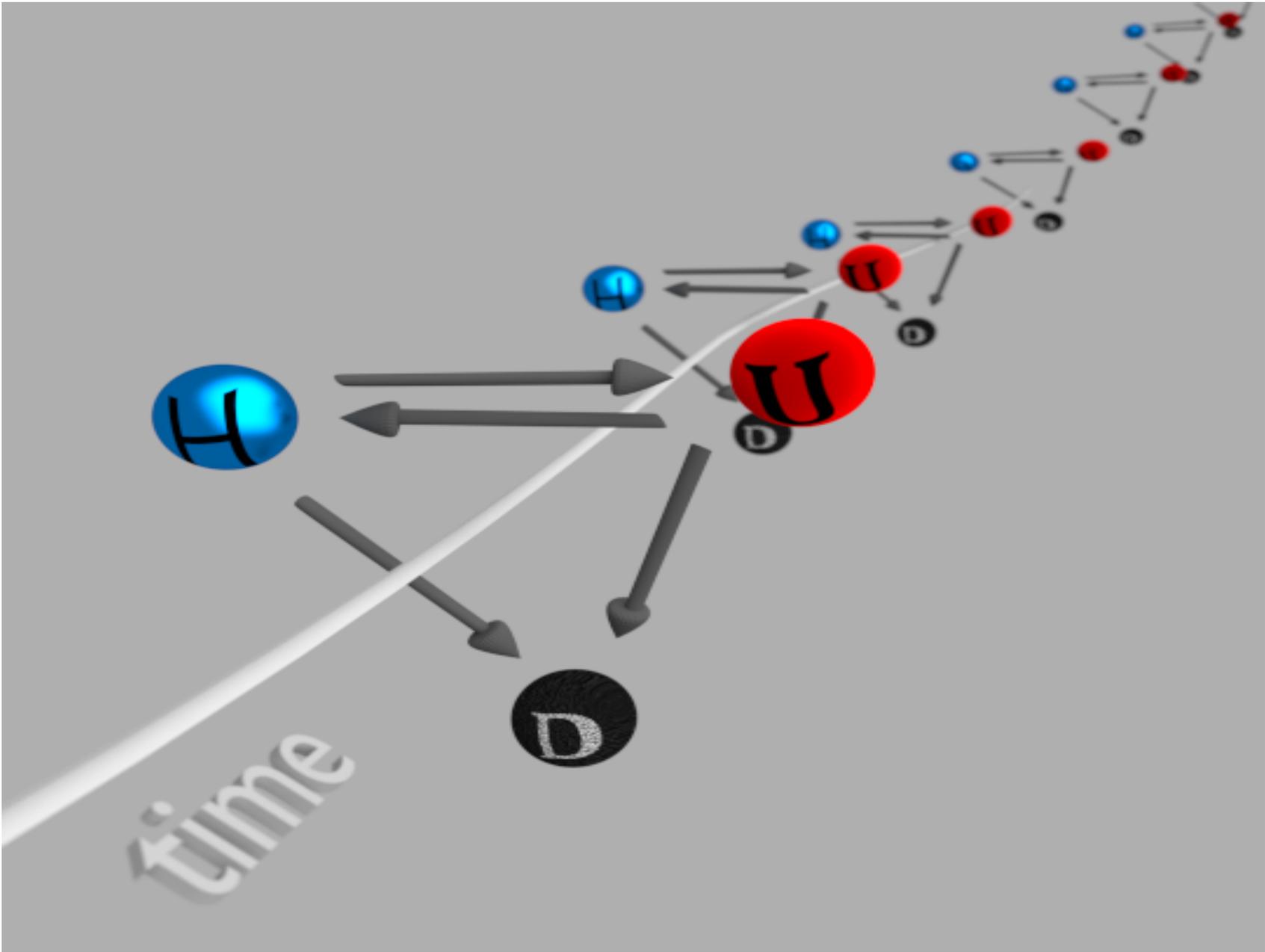
- From December 2007 to March 2011.
- Linked 18 separate databases on the providers' level.
- Each has from 13,095 to 25,266 records.
- 421,671 valid records included in the construction.
- Number of updates ranged from 0 to 18 per provider.
- Continuous process, more records added every 3 months
 - For details, see Hussein, S., Ismail, M. and Manthorpe, J. (2016) Changes in turnover and vacancy rates of care workers in England from 2008 to 2010.

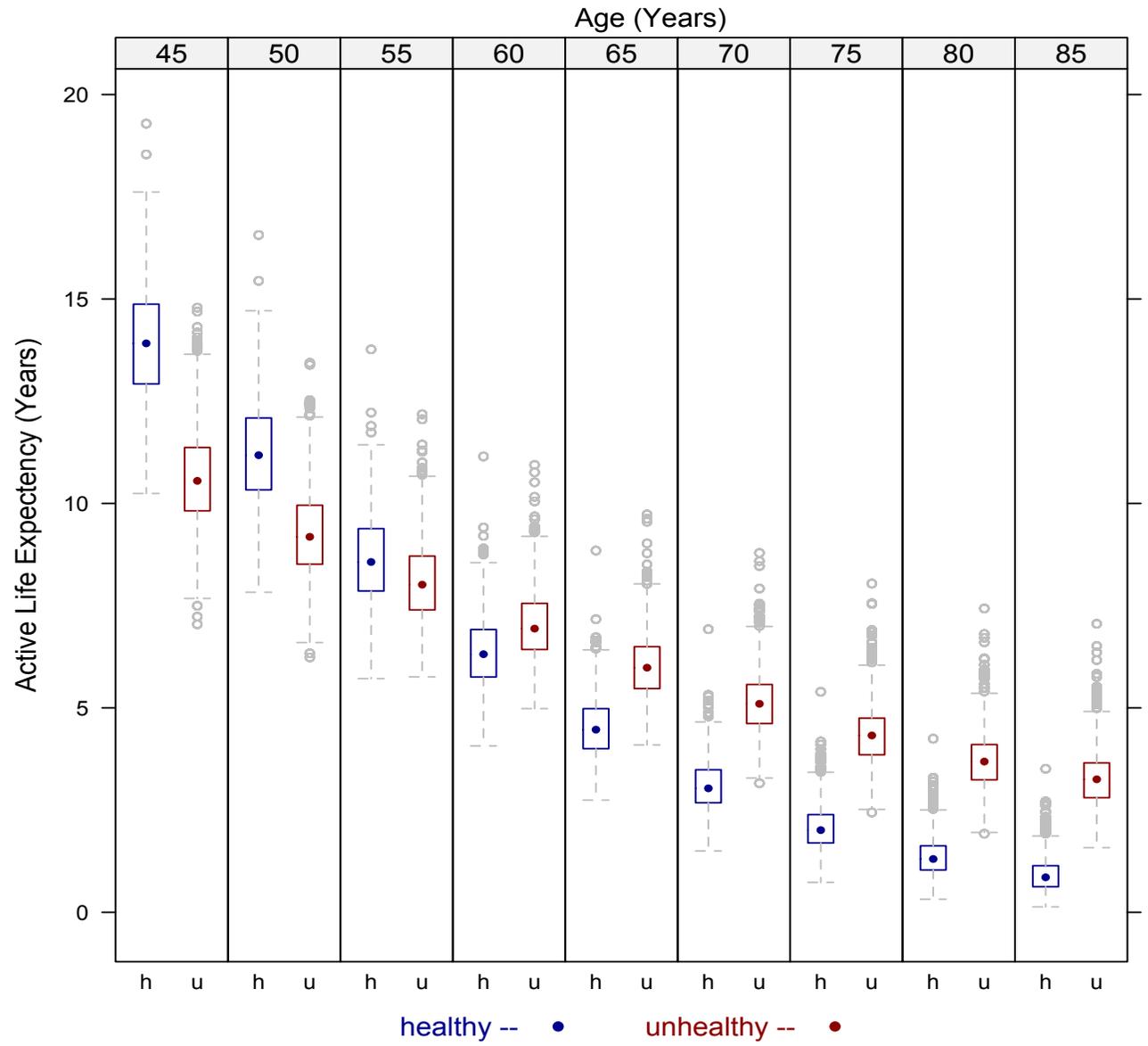
Workers' Level Longitudinal Analysis

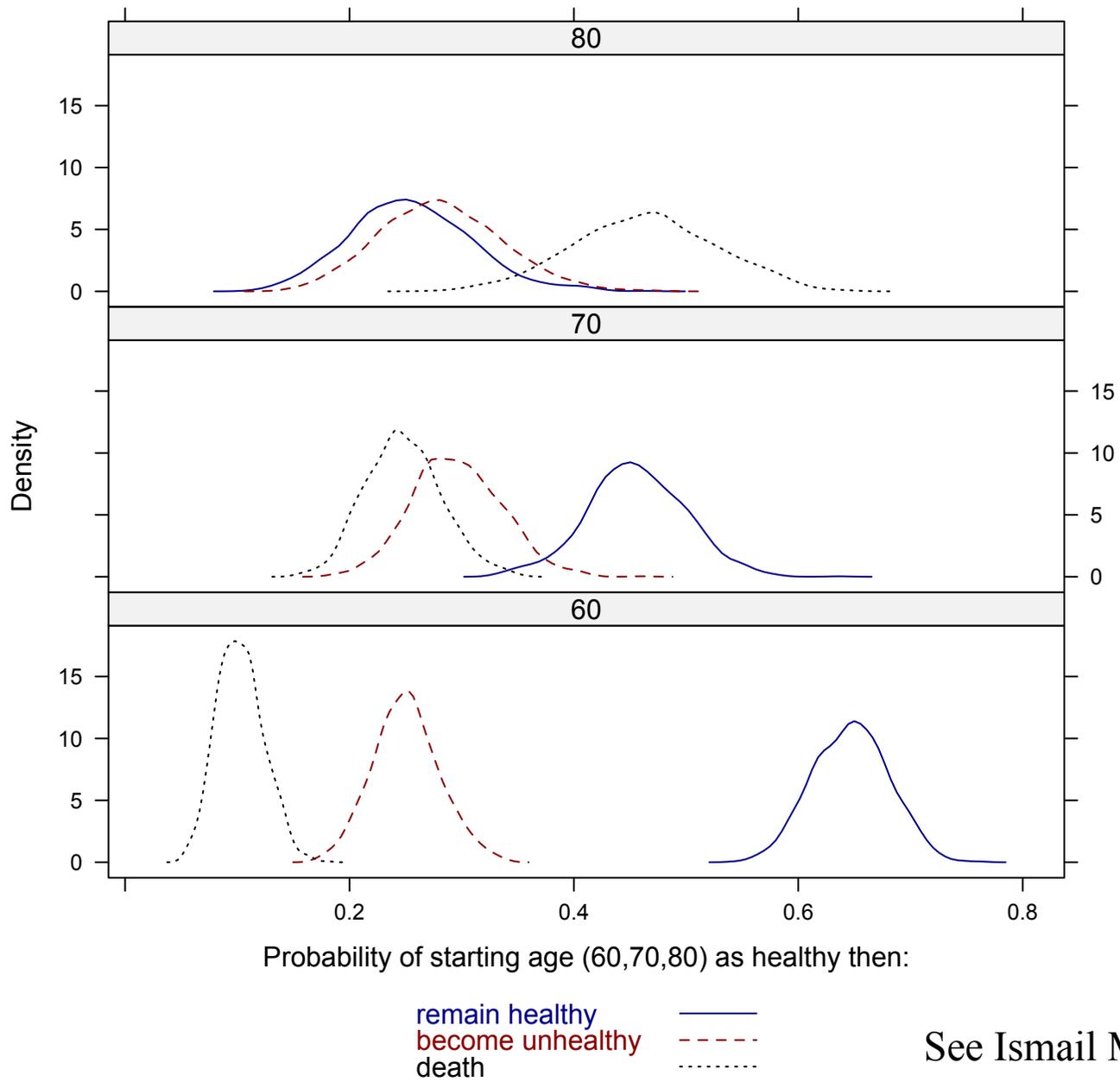
- A much larger database
 - same period of time: over 11M records
- Providers not required to complete information for 'all' workers
 - structural/design missing data
 - true missing data
- Linkage issues
 - more data fields required for identification and linkage
- Considerably large number of variables and fields
 - careful planning; analysis-tailored data retrieval
- Changes in database
 - amendments, new variables etc.
 - programming intensive and demanding models (may not be replicable for different databases)
- See Ismail, M. (2013)

Individual Level: Survey Data

- An example using data from a US National Health Epidemiologic Follow-up.
- Modeling longevity, unhealthy years and burden on different sectors.







See Ismail M., 2014

Conclusion

- Data could be an invaluable tool to inform decision making and to support policy.
- Different levels of data allow different levels of analysis.
- Data have many advantages, both for practice and academic research.
- Each data type has advantages and disadvantages.
- Careful design and implementation can leverage data usage in a cost effective effective way.
- When new schemes are being developed, it is the best time to think about data collection.

Conclusion 2: Administrative Data

- More data ≠ better analysis
- Plan ahead; NMDS didn't originally collect workers' nationalities!
- Consult with all relevant stakeholders.
- Invest in quality assurance at entry point.
- Data validation.
- Have formal procedures for data archiving.
- Regular briefings reinforce value.

References

- Hussein, S. and Ismail, M. (Forthcoming) Long-Term Care Policies in the Gulf Region: A Case Study of Oman. *Journal of Aging and Social Policy*
- Ismail, M., Hussein, S., Stevens, M., et al. (2017) Do personal budgets increase the risk of abuse? Evidence from English national data, *Journal of social policy*, 46(2):291-311
- Christensen, K., Hussein, S. and Ismail, M. (2016) Migrant intelligence shaping work destination choice: the case of long-term care work in the United Kingdom and Norway. *European Journal of Aging*. (Open access)
- Hussein, S. and Ismail, M. (2016) Ageing and Elderly Care in the Arab Region: Policy Challenges and Opportunities. *Ageing International*. (Open access)
- Hussein, S., Ismail, M. and Manthorpe, J. (2016) Changes in turnover and vacancy rates of care workers in England from 2008 to 2010: Panel analysis of national workforce data. *Health & Social Care in the Community*, 24(5):547-56.
- Hussein, S., Ismail, M. and Manthorpe, J. (2016) Male workers in the female-dominated long-term care sector: evidence from England. *Journal of Gender Studies*. 25(1): 35-49.
- Ismail, M. (2014) Longevity, Unhealthy Years And Burden On Different Sectors. *Analytical Research, Research & Insights*, Issue 3. www.analyticalresearch.uk
- Ismail, M. (2013) Longitudinal Data Analysis. *Analytical Research, Research & Insights*, Issue 1. www.analyticalresearch.uk

Thank you!

mohamed@analyticalresearch.co.uk
www.analyticalresearch.uk