MANCHESTER 1824

The University of Manchester

Discovering Sub-diseases with Model-based Machine Learning

Iain Buchan
Director, Farr Institute @ Health eResearch Centre
Director, Centre for Health Informatics, University of Manchester

23rd September 2015
Machine Learning for Personalized Medicine Summer School, Manchester

Endotype Discovery

Aim to identify **subgroups** ("endo-phenotypes" or "endotypes") of disease risk or treatment outcome explained by a **distinctive** underlying **mechanism**

Foundation of **Stratified Medicine,** seeking better-targeted interventions

M.C. Escher

Order and Chaos, 1950

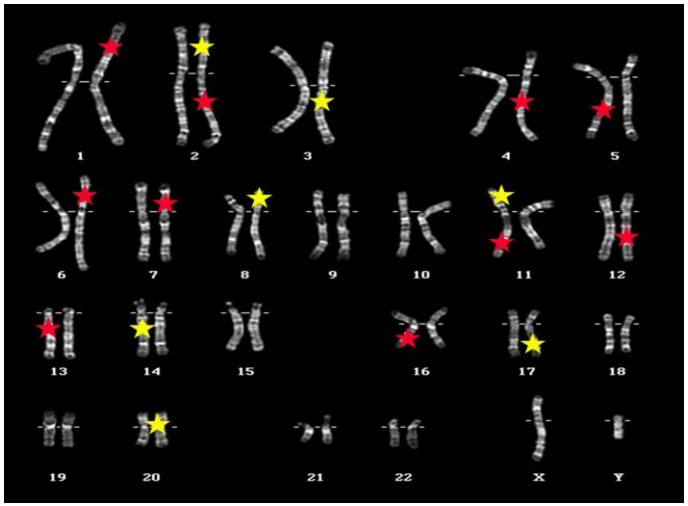
The Farr Institute of Health Informatics Research

Asthma: Low GWAS Yield

- Legacy of non-replicated genetic epidemiology, typical of most common chronic disorders

☆ Linkage in 1 study only

★ Linkage in >1 study



The Farr Institute of Health Informatics Research

Asthma: Gene ∪ Environment

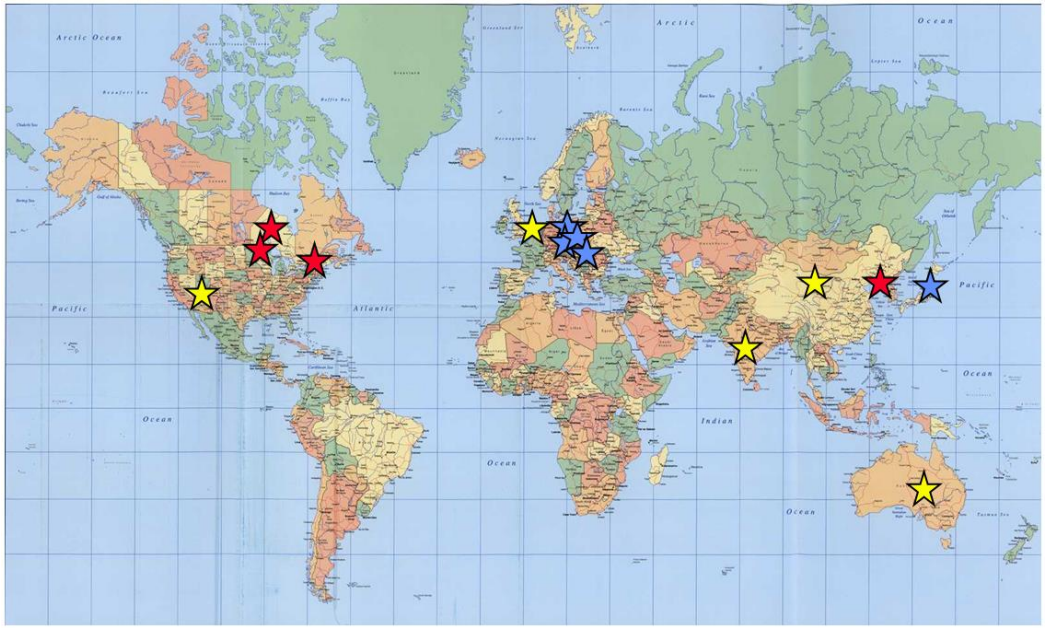
- Important gene-environment interaction information may be averaged out in narrow or aggregated studies

CD14 Endotoxin Receptor

★ C allele associated

★ T allele associated

★ No association



The Farr Institute of Health Informatics Research

Asthma: Heterogeneity

- Seems to be a collection of **several diseases**,
  each with distinctive pathophysiology,
  and environmental ∪ genetic associates ("asthma endotypes")

- Usually starts early in life,
  and may **progress**, **remit** or **relapse** over time

- Single cohort study **hypothesis-driven** epidemiology
  lacks temporal and environmental **complexity**
  needed to smoke out endotypes

Scaling-up

- **Multi-cohort**
  - Different windows on calendar **time**
  - Different windows on human **development**
  - Variety of **populations**
  - Variety of **environments**

- **Multi-disciplinary** and **multi-perspective**
  - Biostatistics (**deductive**) and machine-learning (**inductive**)
  - Tapestry of **reasoning** about mechanisms

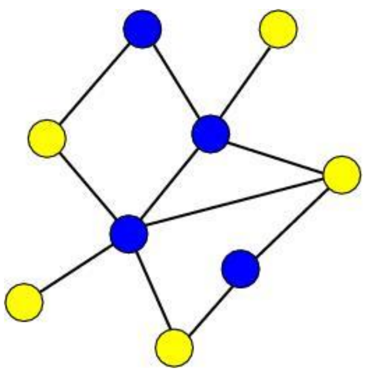
- **Hypothesis forming** *and* **following**

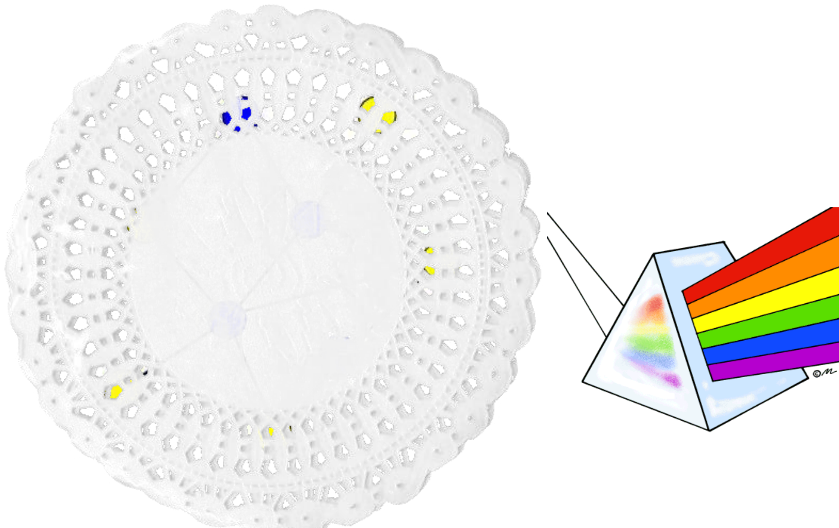
Health Data: No Mining Please

Problem Space

Observation Space

Data Space





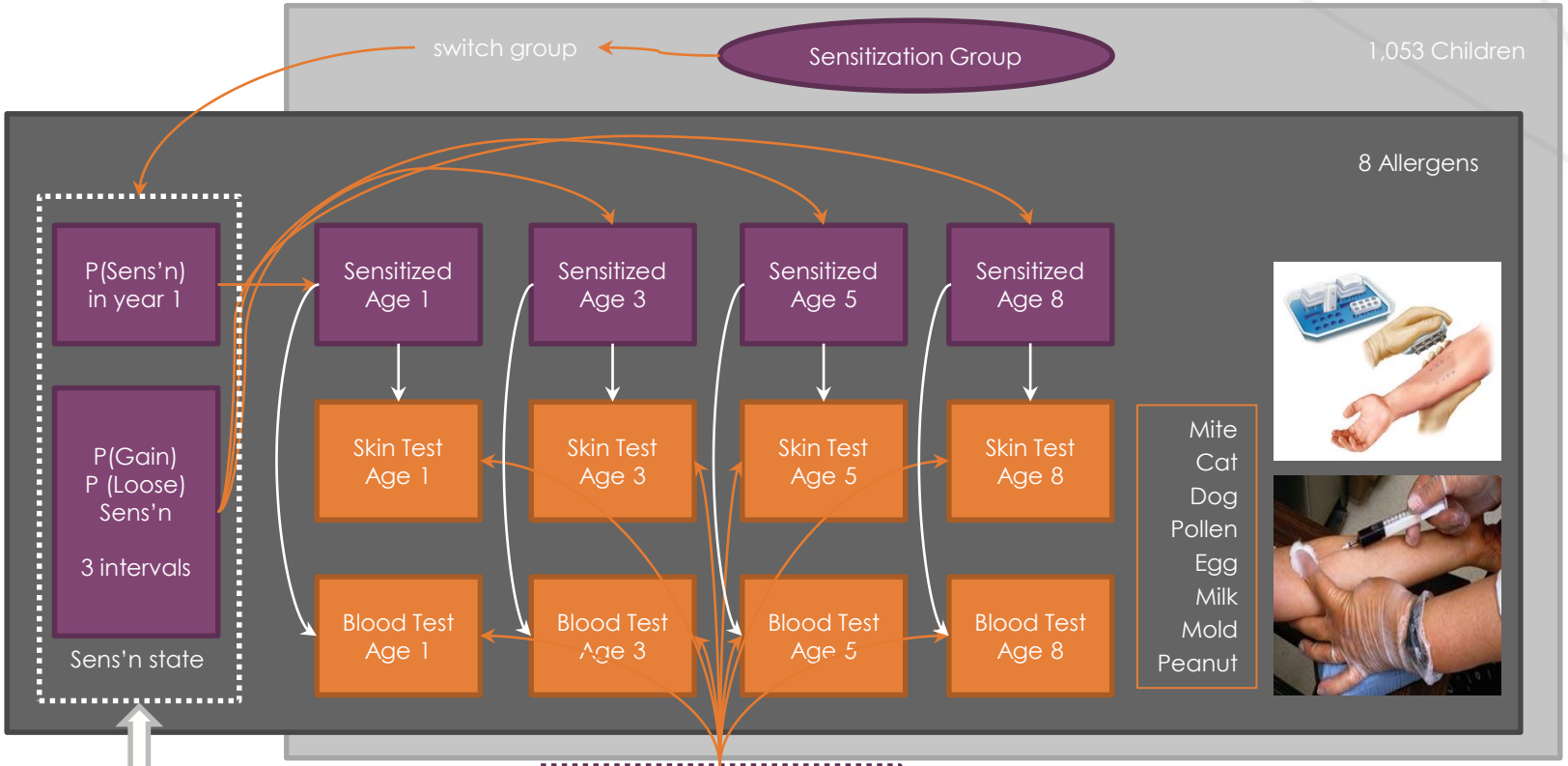
Countries	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Total
Bangladesh	8,221	8,678	8,645	6,046	7,215	7,171	5,492	4,616	8,061	11,487	75,632
Canada	15,799	11,595	10,184	8,864	16,210	21,933	19,519	11,350	15,567	21,878	152,899
China*	54,902	53,229	49,135	43,835	60,111	76,918	77,208	51,059	64,068	82,868	613,333
Colombia	14,153	12,945	11,761	9,966	14,498	16,730	18,845	14,720	18,678	25,571	157,867
Cuba	26,209	33,422	17,228	14,132	20,831	27,703	28,272	9,262	20,488	36,261	233,808
Dominican R.	31,807	22,592	17,391	17,864	17,536	21,313	22,604	26,159	30,492	27,504	235,262
Ecuador	8,305	8,062	6,839	8,904	7,685	9,706	10,602	7,066	8,611	11,608	87,388
El Salvador	17,888	17,958	14,584	14,606	22,578	31,272	31,168	28,231	29,795	21,359	229,439
Germany	6,748	5,619	5,442	5,201	7,638	9,886	8,961	5,064	7,099	9,264	70,922
Guatamala	8,753	7,778	7,752	7,308	9,970	13,567	16,229	14,386	17,999	16,825	120,567
Guyana	9,467	7,228	3,957	3,300	5,746	9,303	9,962	6,809	6,329	9,318	71,419
Haiti	18,368	15,091	13,444	16,532	22,364	27,120	20,268	12,293	13,998	14,529	174,007
Honduras	5,870	7,615	6,445	4,809	5,939	6,615	6,461	4,645	5,505	7,012	60,916
India	44,817	38,048	36,458	30,237	42,046	70,290	71,105	50,228	70,116	84,681	538,026
Iran	11,020	9,846	7,809	7,203	8,519	10,497	13,029	7,230	10,434	13,887	99,274
Ireland	1,731	999	944	812	1,315	1,522	1,425	983	1,531	2,088	13,350
Jamaica	19,073	17,804	15,062	14,733	16,000	15,393	14,898	13,347	14,414	18,346	159,070
Japan	6,011	5,004	5,063	4,217	7,094	9,619	9,301	5,971	7,694	8,768	68,742
Korea	17,991	14,109	13,595	12,840	15,830	20,742	21,021	12,382	19,766	26,561	174,837
Mexico	163,489	146,828	131,534	147,573	173,919	206,426	219,380	115,585	175,364	161,445	1,641,543
Nicaragua	6,903	6,672	3,517	13,389	24,029	19,896	10,850	4,094	4,000	3,305	96,655
Nigeria	10,216	6,699	7,758	6,769	7,853	8,291	8,129	7,872	9,374	10,598	83,559
Pakistan	12,512	12,959	13,093	13,496	14,535	16,448	13,743	9,415	12,086	14,926	133,213
Peru	12,817	10,820	10,137	8,438	9,613	11,131	11,999	9,409	11,781	15,352	111,497
Philippines	53,435	47,235	32,838	31,026	42,474	53,154	51,308	45,250	57,827	60,748	475,295
Poland	15,768	12,038	8,466	8,798	10,114	11,818	12,746	10,510	14,250	15,452	119,960
Soviet Union*	40,737	32,688	30,155	18,985	26,443	56,286	56,936	34,892	42,705	66,007	405,834
Trinidad &Tob.	7,344	6,362	4,809	4,283	6,660	6,665	5,771	4,138	5,384	6,568	57,984
U.K	13,584	10,686	8,994	7,690	13,385	18,436	16,181	9,527	14,915	19,800	133,198
Vietnam	42,052	38,494	17,313	20,393	26,747	35,531	33,627	22,087	31,514	32,784	300,542
Yugoslavia*	11,853	10,404	8,007	8,987	17,023	34,620	41,272	8,699	17,024	23,205	181,094
Other	184,998	150,791	119,832	125,332	167,887	178,316	175,420	136,263	179,273	242,388	1,660,480
Total U.S.	**902,841**	**790,098**	**648,191**	**646,568**	**849,807**	**1,064,318**	**1,063,732**	**703,542**	**946,142**	**1,122,373**	**8,737,812**

$y = b1x1 + b2x2 + b3x3 + c$

...Health is measured with **error** and **missingness**:
Endo-phenotypes are resolved as if the researcher
was looking through a prism and doyley at the problem

The Farr Institute of
Health Informatics Research

Hypothesising with Data



infer.net

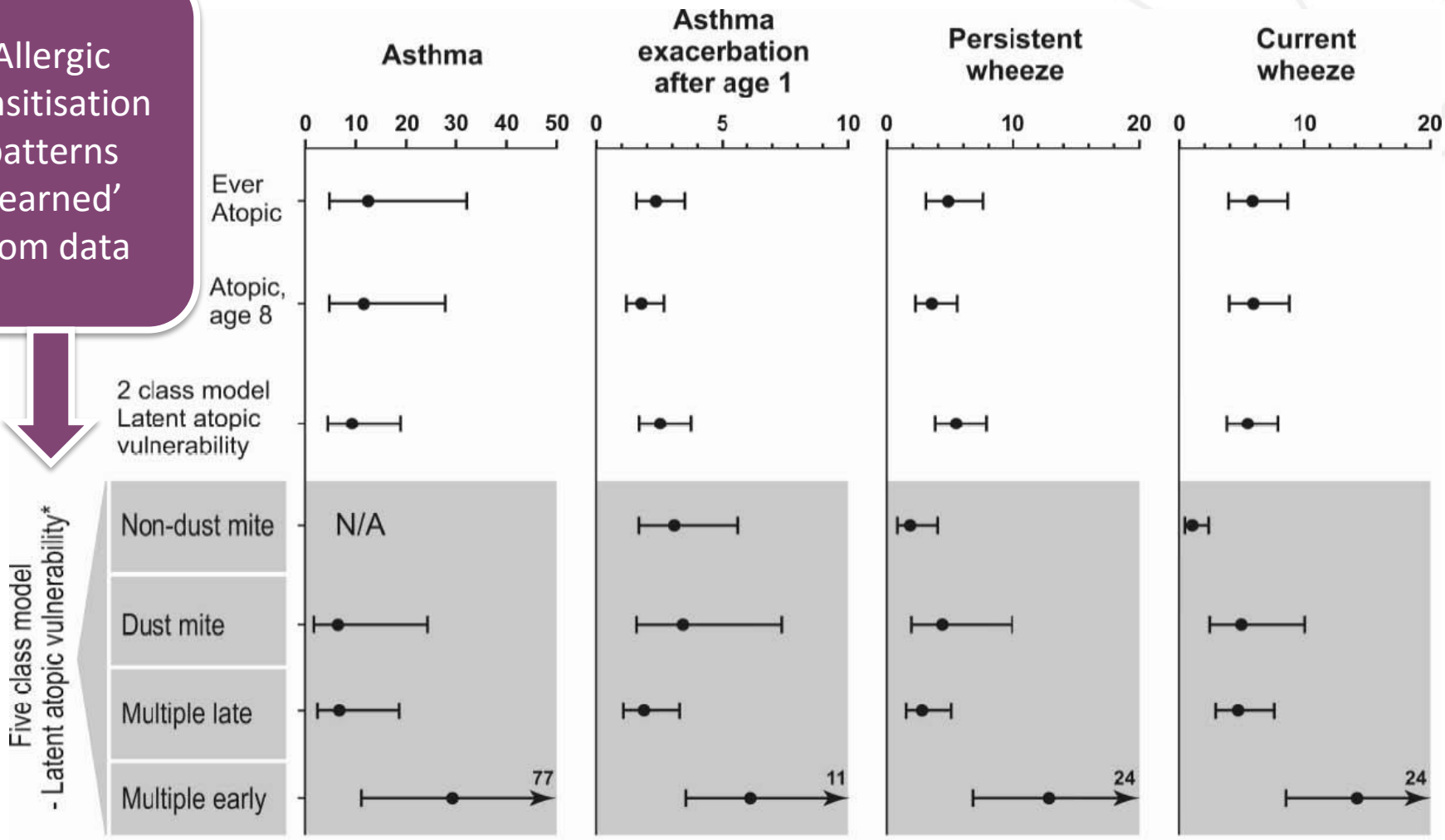
Machine-learning software & partial statistical models

MAAS

Farr The Farr Institute of Health Informatics Research

New Asthma Risk Factor Found

Allergic sensitisation patterns 'learned' from data

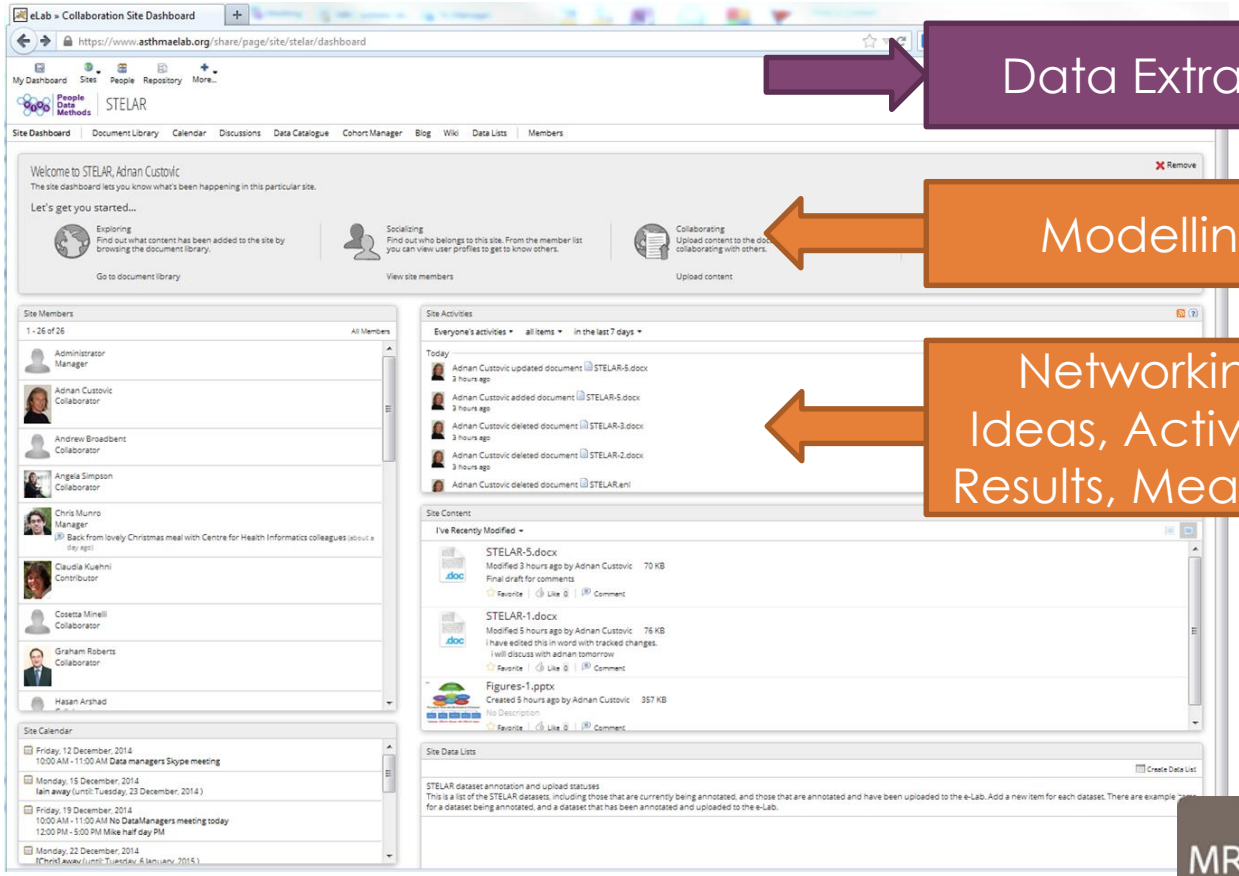


Cross Cohort Team Research

Data & Harmonized Metadata from Cohorts

MRC STELAR Consortium: www.asthmaelab.org

- MAAS
- SEATON
- ASHFORD
- ALSPAC
- IOW



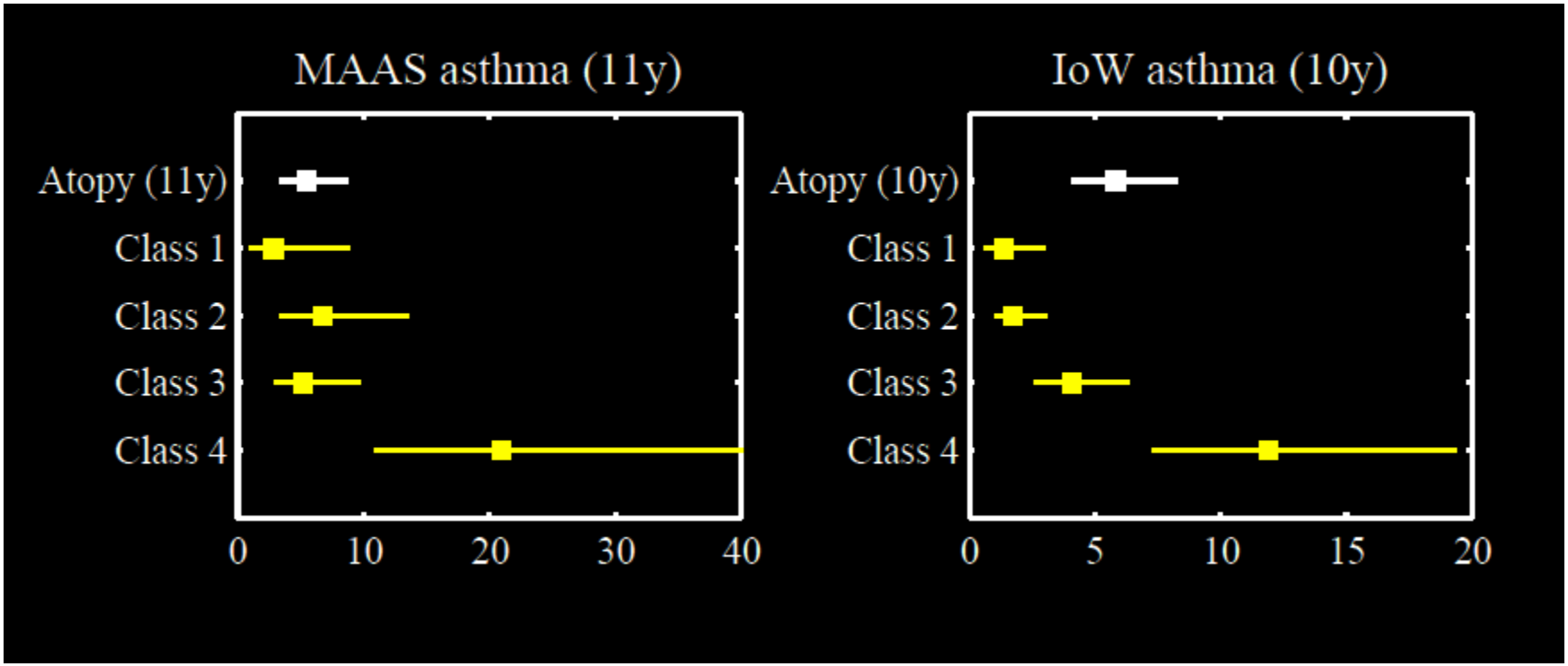
Data Extracts

Modelling

Networking: Ideas, Activities, Results, Meanings

MRC Medical Research Council

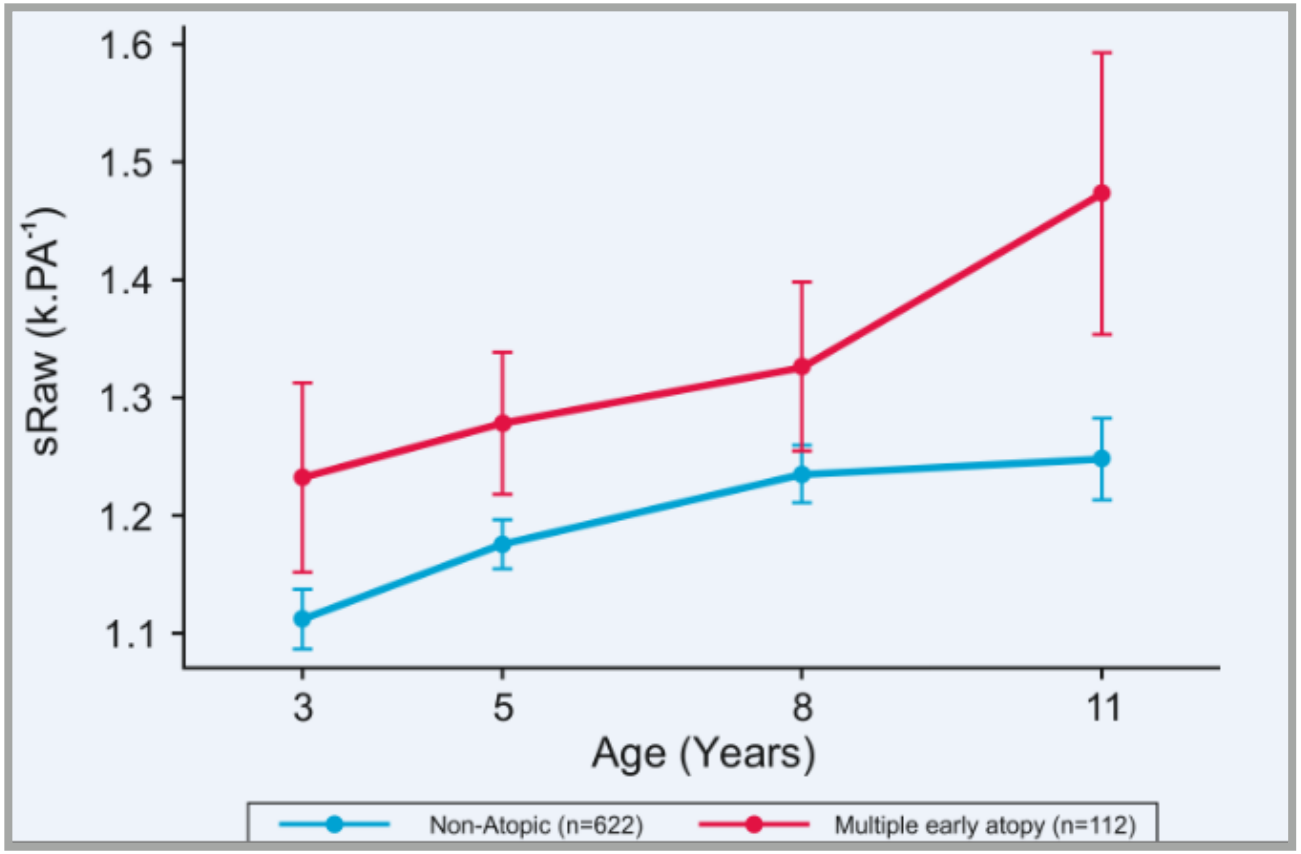
‘Learned’ Atopy Classes Portable



Lazic et al, Allergy 2013; 68(6): 764-70

The Farr Institute of Health Informatics Research

Risk Factor Development



Risk group 'induced' from data shows different natural history of airways resistance

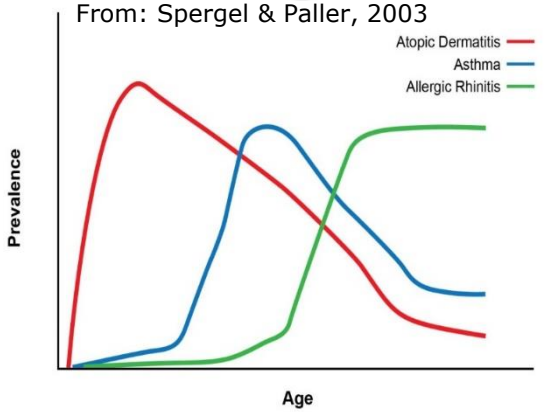
Belgrave et al, Am J Respir Crit Care Med 2014;189(9):1101-9

The Farr Institute of Health Informatics Research

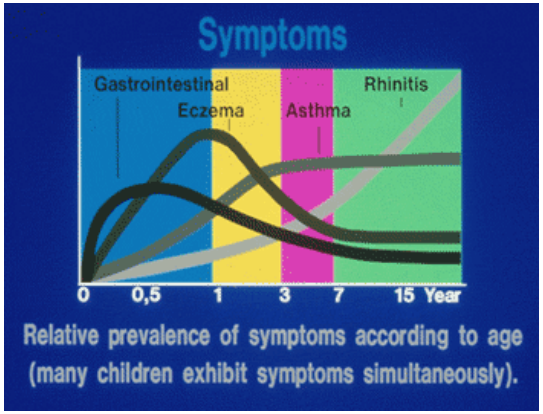
Assumed Biology: Atopic March

- Progression of **allergy**
  Eczema → Asthma → Rhinitis
- Inferred from **population** summary →
- Assumed **causal** link between eczema – asthma & rhinitis
- Clinical response:
  target children with eczema to reduce progression to asthma

From: Spergel & Paller, 2003

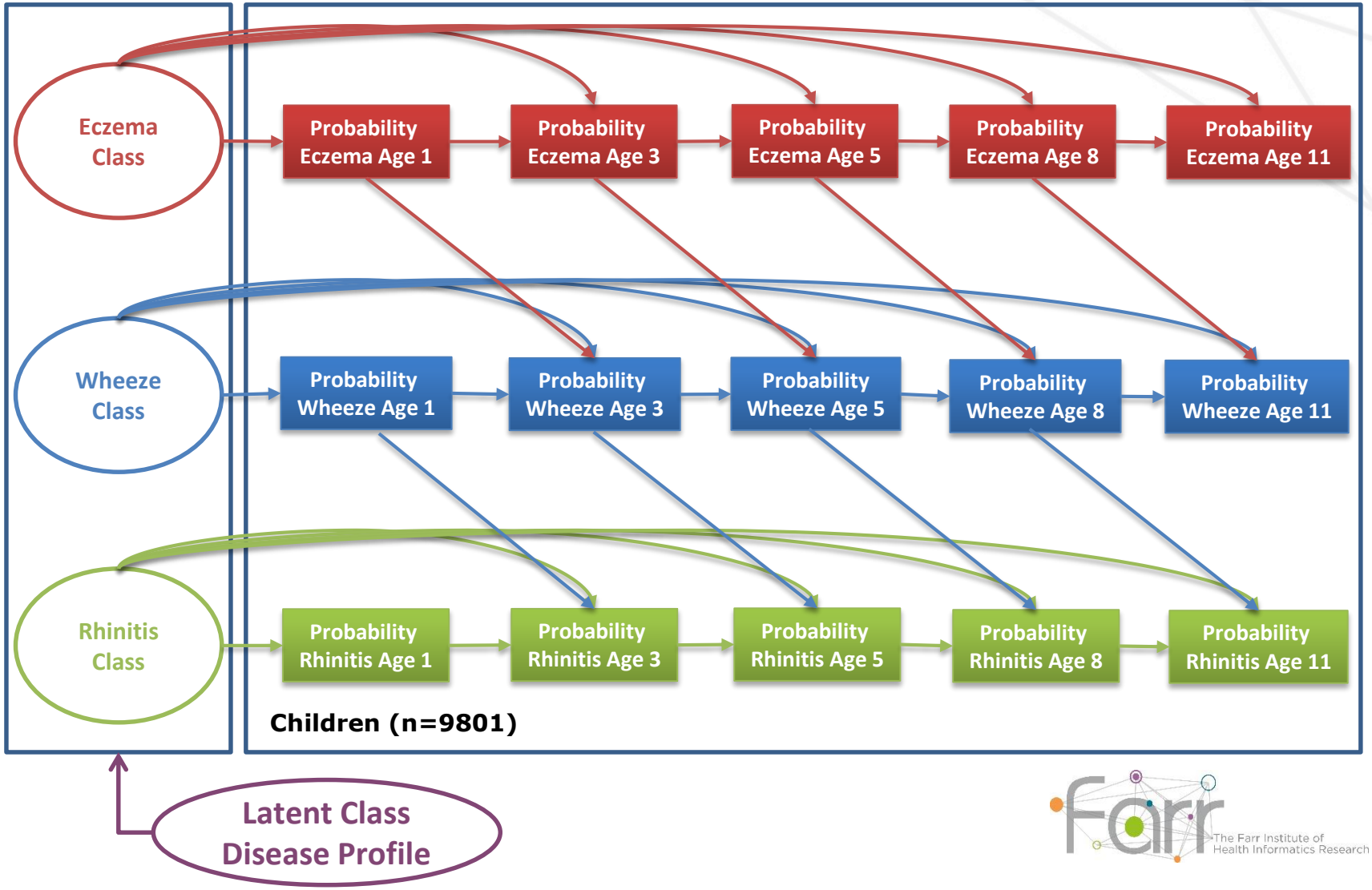


From: World Allergy Organization, 2014



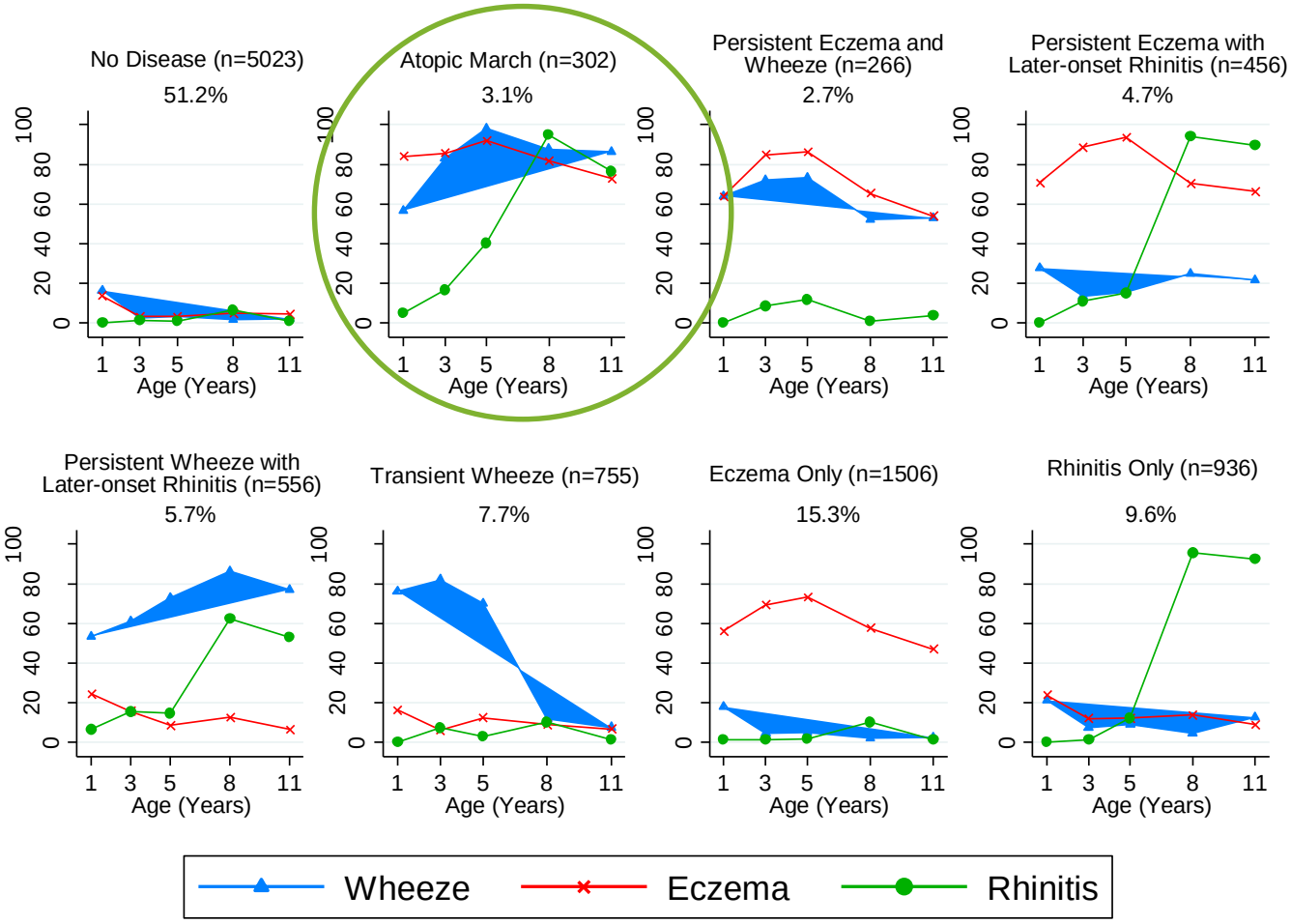
The Farr Institute of Health Informatics Research

Individual-level Longitudinal Analysis



The Farr Institute of Health Informatics Research

Myth Bust by Learning from Data



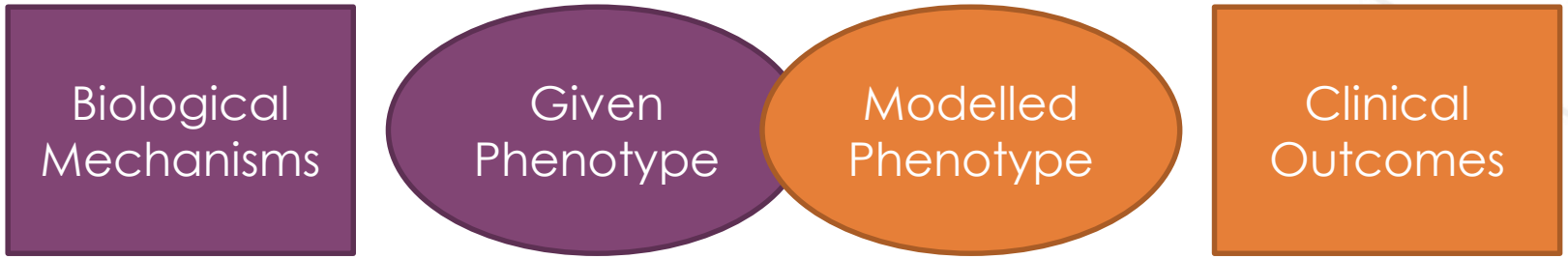
MRC Medical Research Council

MRC STELAR consortium working at scale across MAAS and ALSPACS cohorts

From: Belgrave et al. Developmental Profiles of Eczema, Wheeze, and Rhinitis: Two Population-Based Birth Cohort Studies. PlosMedicine 2014

Farr The Farr Institute of Health Informatics Research

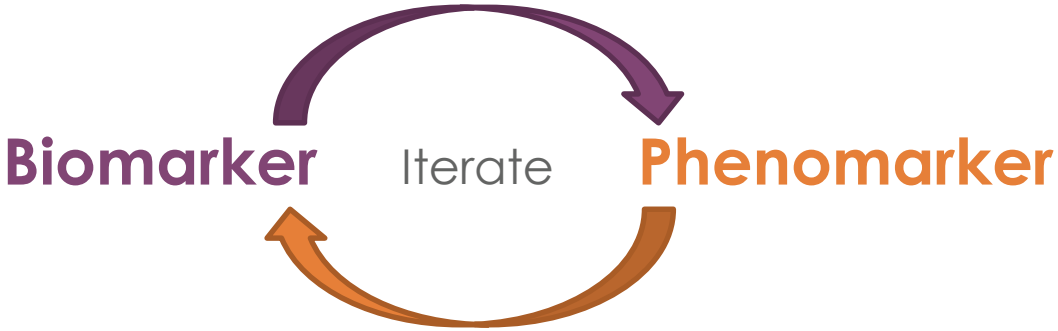
Joined-up Bio-Health Modelling



Bioinformatics

Health Informatics

Integrative Informatics needed



Portable Model
(not a fixed label/rule)

The Farr Institute of Health Informatics Research

Real World Phenomarking

?how to generate real world evidence on gliflozins

Total evidence-base predicts < 30% outcomes

Mr Jones…

His GP…

Lifestyle factors: diet, exercise

His nephrologist…

BP control

His diabetologist…

Glucose control

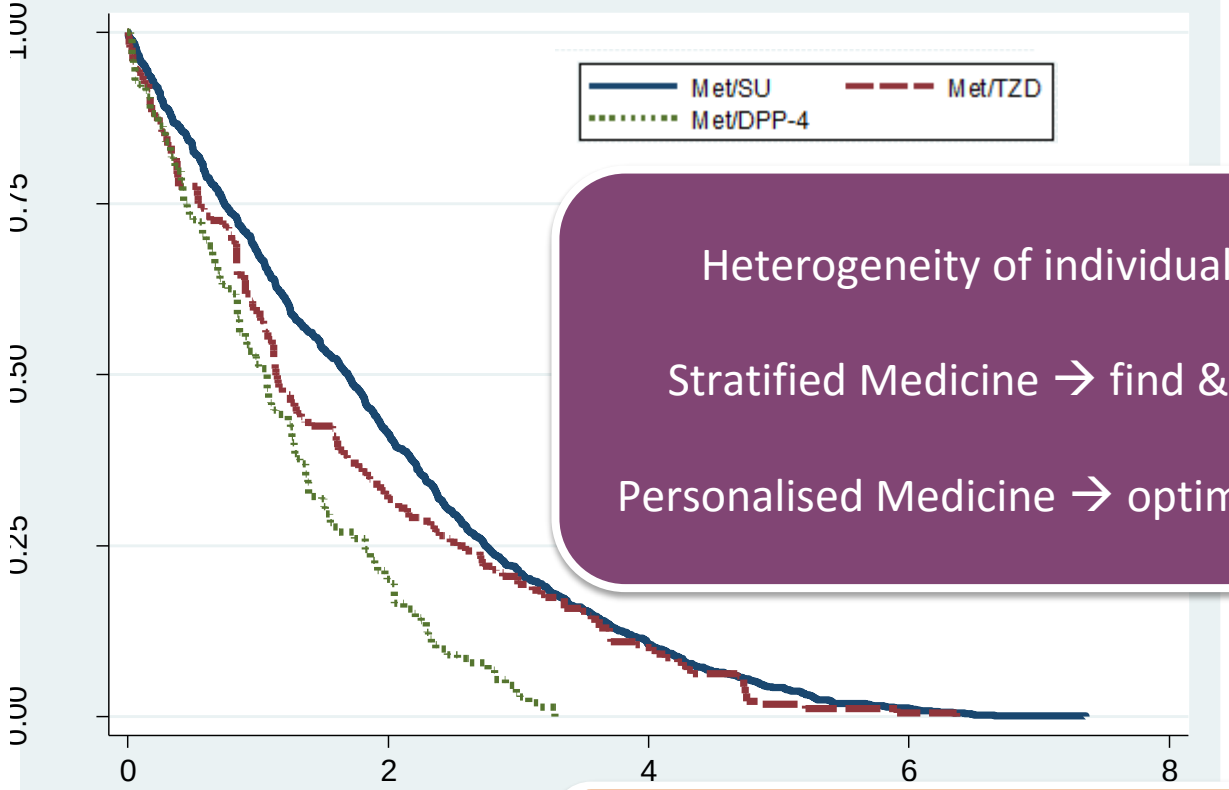
3 x evidence pipelines
1 x complex patient

↑ Weight → ↑ BP

Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. Ann. Fam. 2009;7:357–363.

Farr The Farr Institute of Health Informatics Research

Dual Therapy for Diabetes



Heterogeneity of individual treatment response

Stratified Medicine → find & treat more subgroups

Personalised Medicine → optimise individual's response

Time to microvascular event from diagnosis of diabetes
Inverse probability weighted marginal structural model
Average causal effects of dual therapies

Data source: CPRD
Analysis: A Wright, D Ashcroft, R Emsley

Data: Does Size Matter?

DATA

Vast volume, velocity, variety…

TSUNAMI

METHODS & OUTPUTS

Supra-linear growth in papers & tools

BLIZZARD

EXPERTISE

Similar number of analysts

DROUGHT

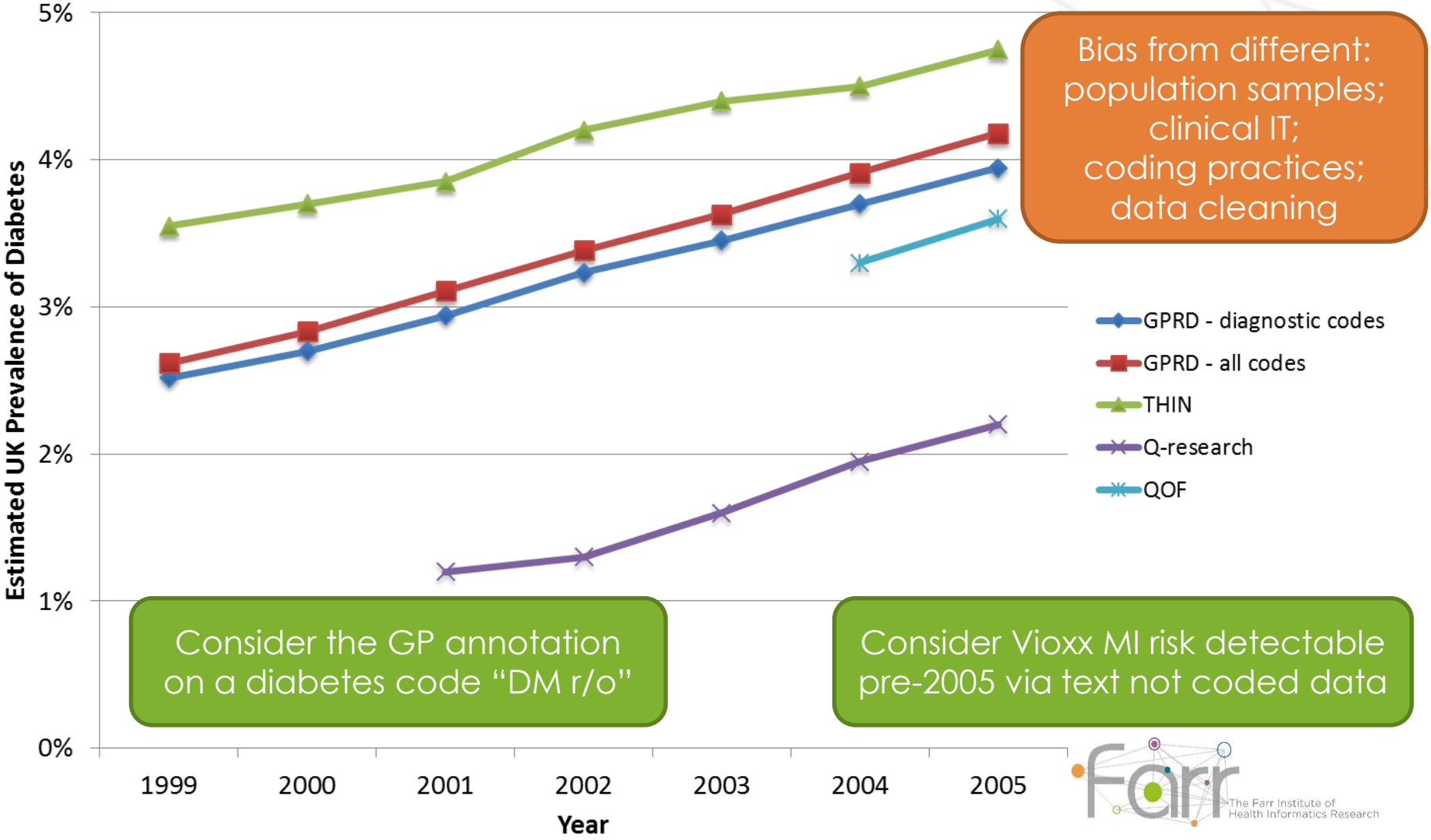
More data * small-scale research = more small-scale research
>n with >heterogeneity can *reduce* 'power'

Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005 Aug;2 (8):e124.

Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf. 2013 Oct;36 Suppl 1:S5-14.

Farr
The Farr Institute of Health Informatics Research

Beyond In-licensed Data



Bias from different:
population samples;
clinical IT;
coding practices;
data cleaning

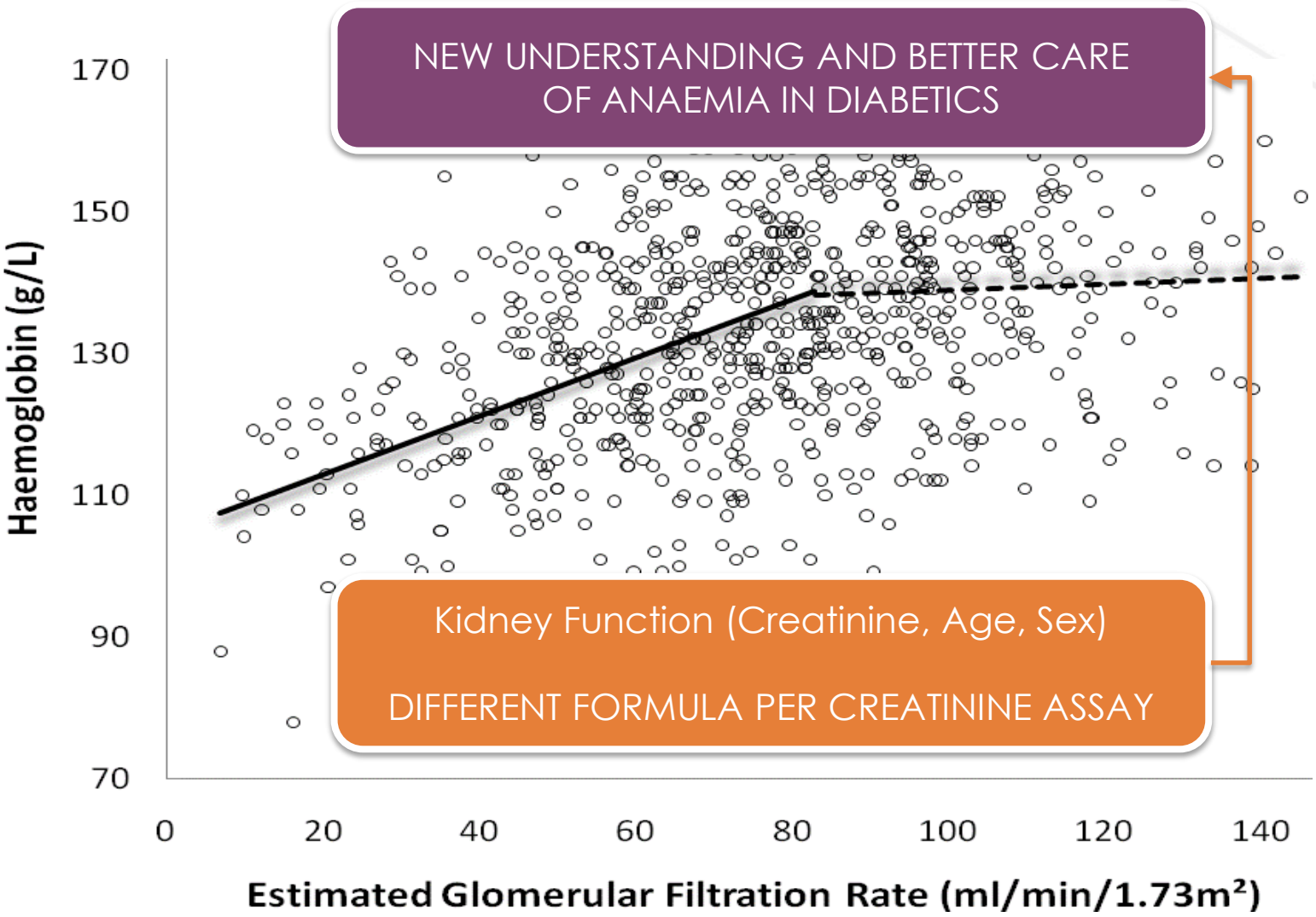
Consider the GP annotation on a diabetes code "DM r/o"

Consider Vioxx MI risk detectable pre-2005 via text not coded data

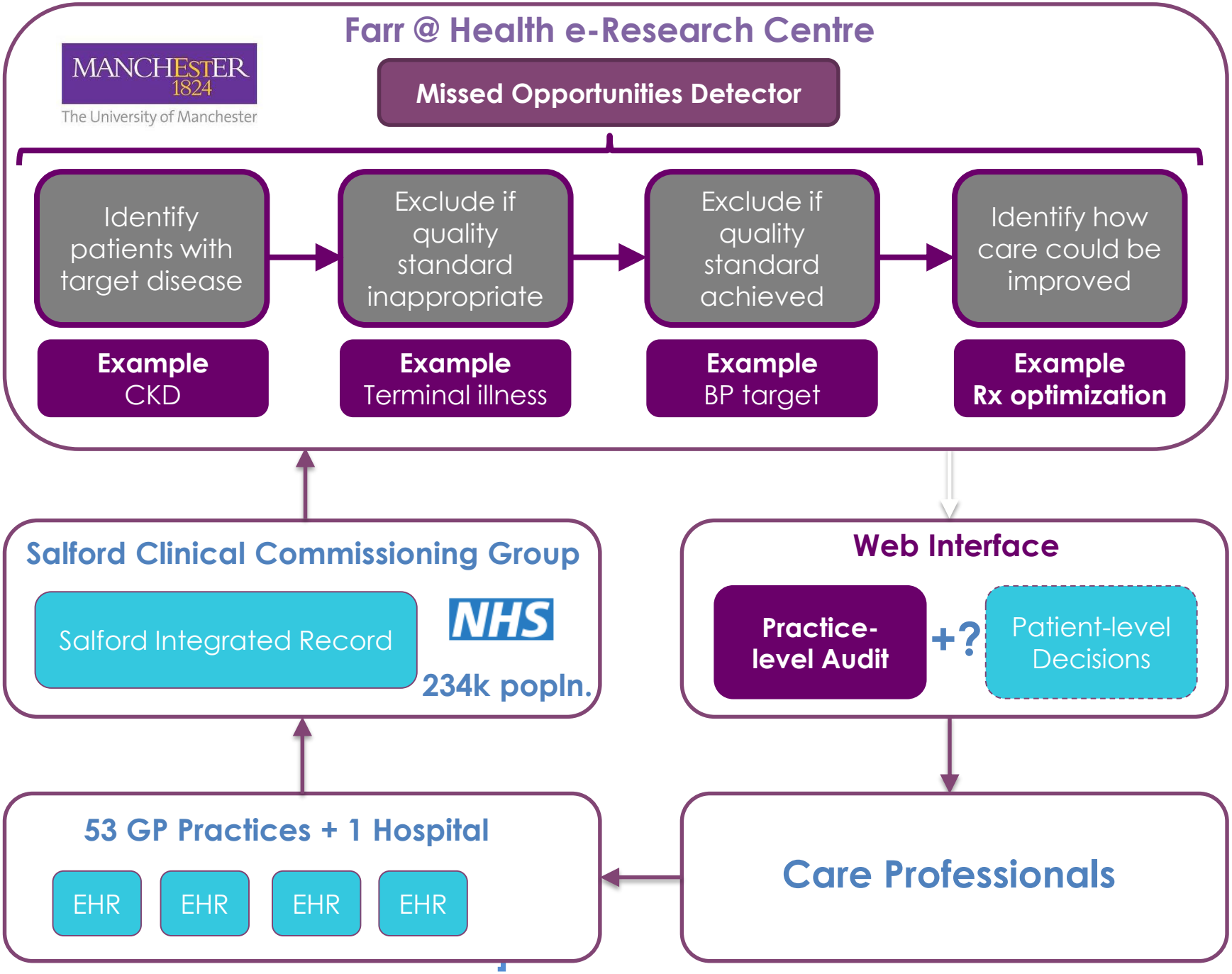
The Farr Institute of Health Informatics Research

From E. Kontopantelis & T. Doran

Key Local Metadata

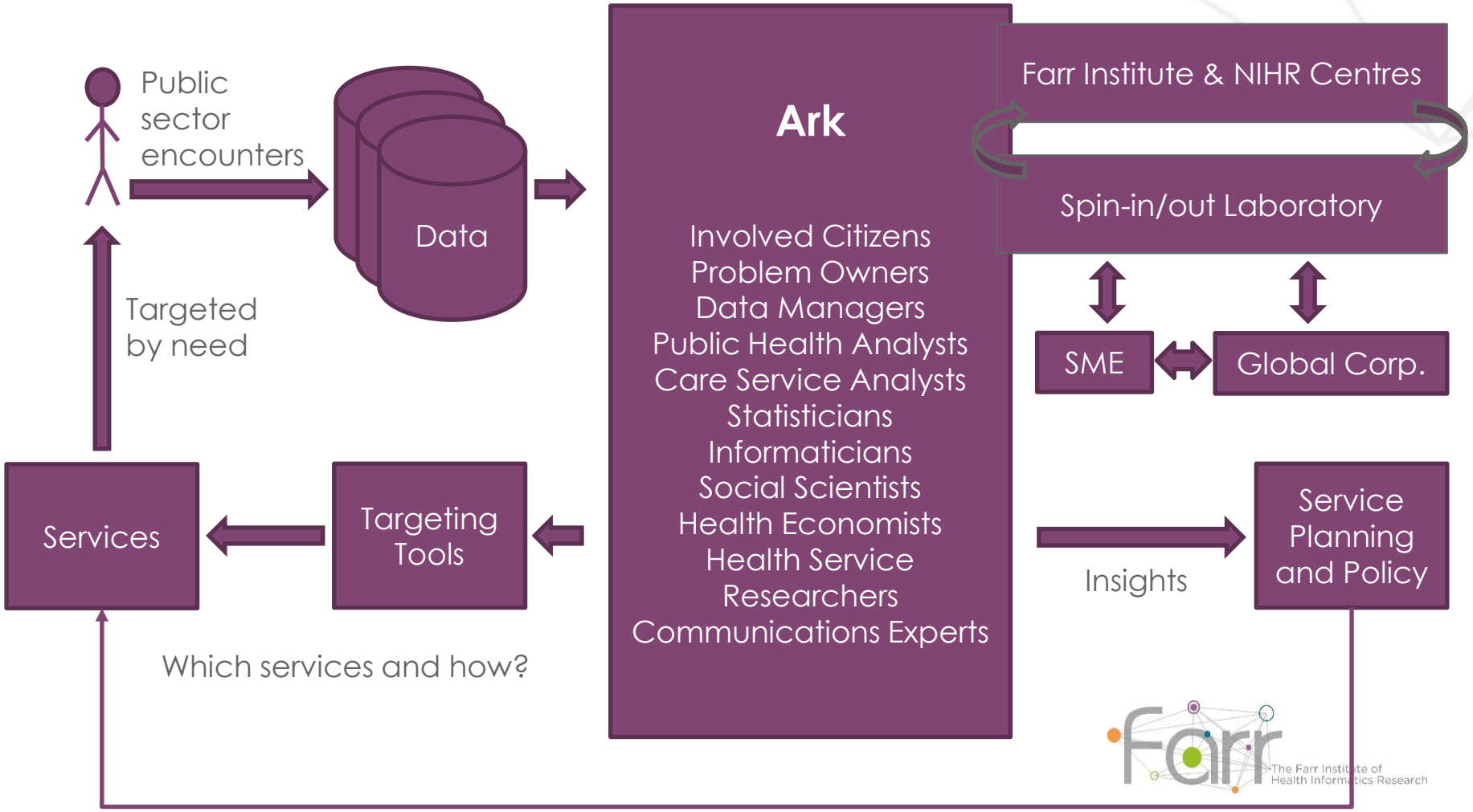


Data source:
Salford eHR

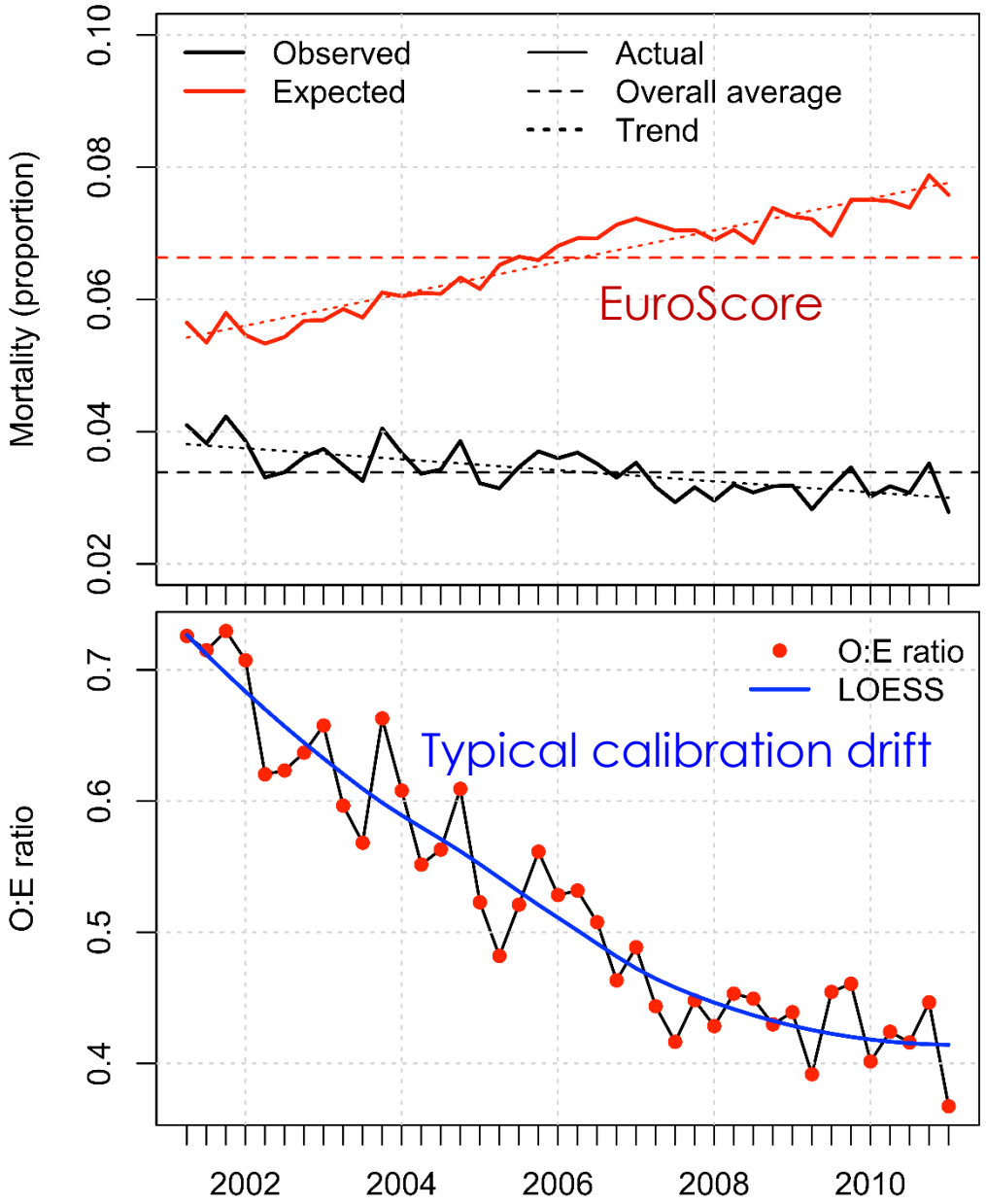

Farr @ Health e-Research Centre
MANCHESTER 1824
The University of Manchester
Missed Opportunities Detector
Identify patients with target disease
Exclude if quality standard inappropriate
Exclude if quality standard achieved
Identify how care could be improved
Example CKD
Example Terminal illness
Example BP target
Example Rx optimization
Salford Clinical Commissioning Group
Salford Integrated Record
NHS
234k popln.
Web Interface
Practice-level Audit
+?
Patient-level Decisions
53 GP Practices + 1 Hospital
EHR
EHR
EHR
EHR
Care Professionals

From B. Brown

Connected Health Cities



Clinical Outcome Prediction



Production line of clinical prediction models is broken

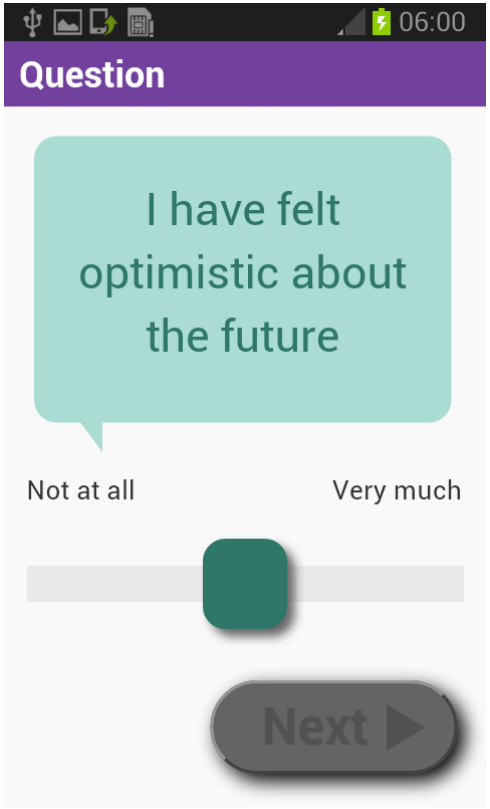
EU Directive 2007/47:
The law now sees algorithms as medical devices

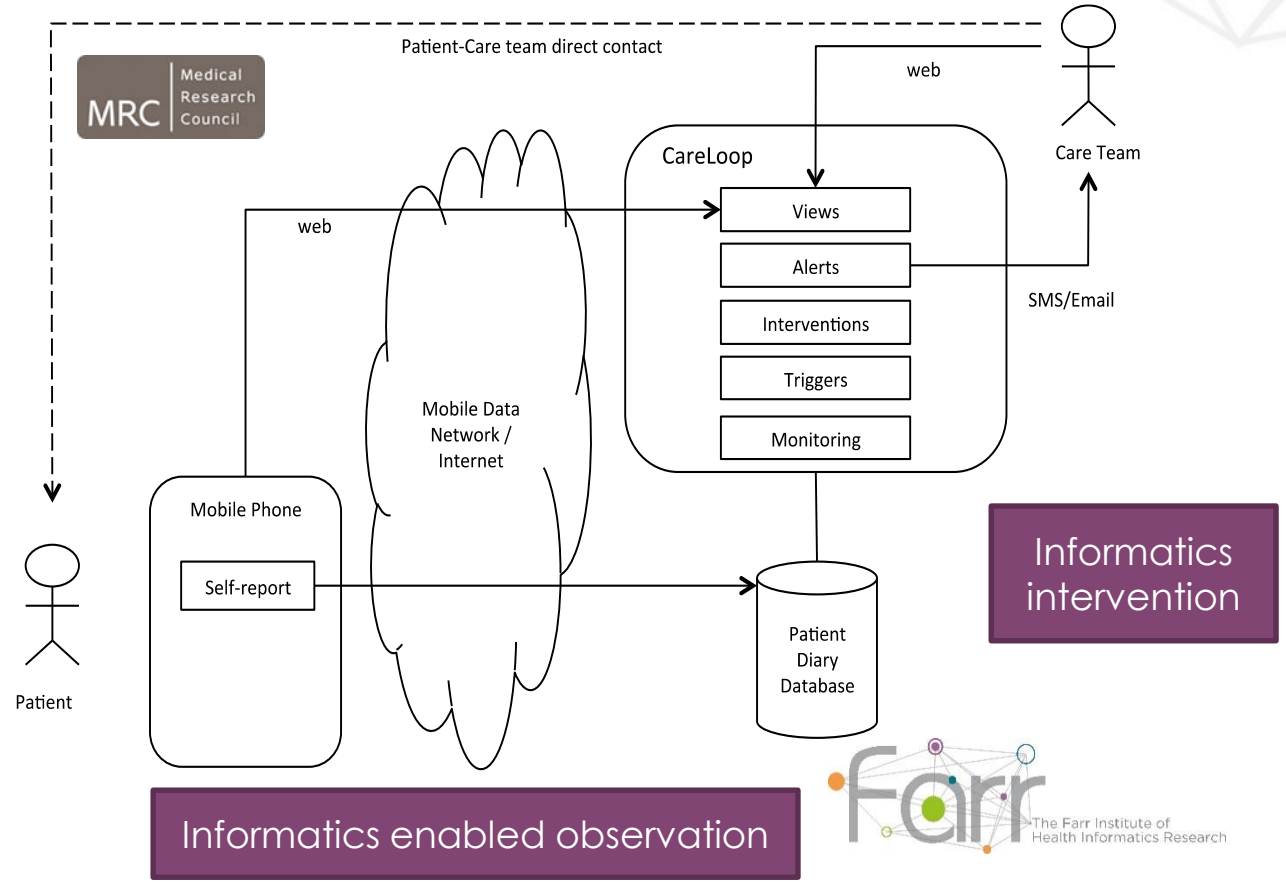
The Farr Institute of Health Informatics Research

From G. Hickey & B. Bridgewater

Health System Extension: Mobile

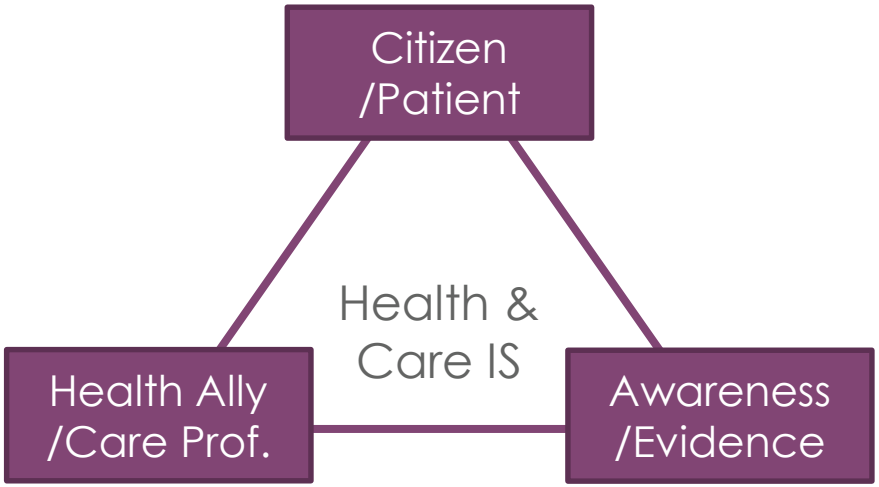
Aim: To Reduce Relapse in Schizophrenia via Smartphone
Drug + behaviour (information * psychological endotype) = outcome





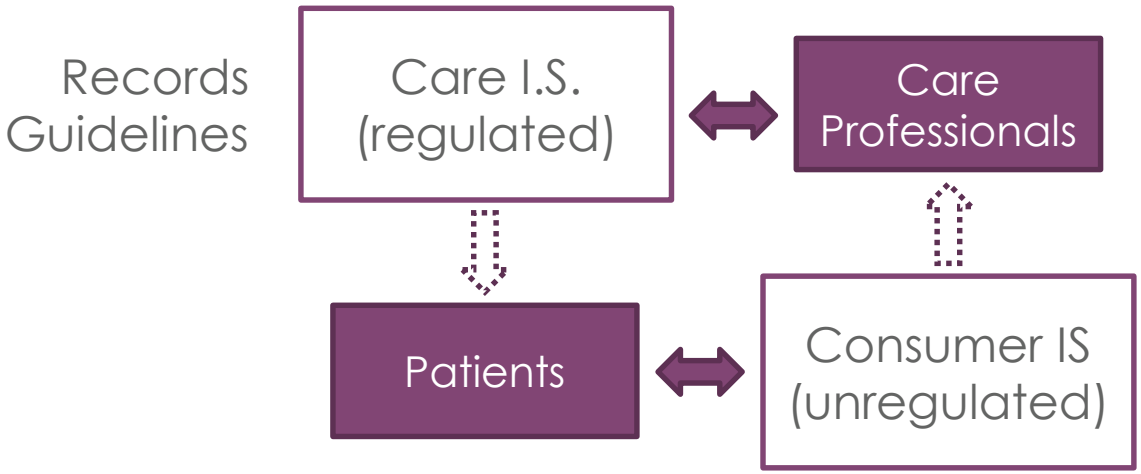
From J. Ainsworth & S. Lewis

Digital Health Triangle



Digital Health Economy (20?)**

Ubiquitous technologies
Unifying models
Usable interfaces (avatar etc.)
Stratification ↔ personalisation
Actionable micro-evidence
Service ↔ research



Two Worlds of Healthcare (2015)

Apps & sensors
Healthcare information
Social networks

Depth, Trust and Scale

PUBLIC TRUST

PATIENT TRACTION

Learning Health System
e.g. "Connected Health City"

- Patient / Mobile Data
- Shared Care Record
- Molecular characterisation
- Core Outcomes
- Dynamic Prediction Models
- Clinical Performance Networks
- Routine Randomisation
- Learning from Other Systems

INTEGRATIVE INFORMATICS

From care + data
→ research
→ translation

To a federation of research in care: pulled & shaped by local communities, with academic and industry partners

Farr

Sub-disease Research Tips

1. Target
   - Plausible diagnostic **aggregation**
   - Unexplained **variation** in clinical outcomes

2. Data
   - Multiple populations/settings (**heterogeneity** & **replication**)
   - Useful **temporal** structure

3. Analytics
   - Multi-perspective ML **pre-model framing** – don't rush in!
   - Heuristic **phenomarker-biomarker** resolution