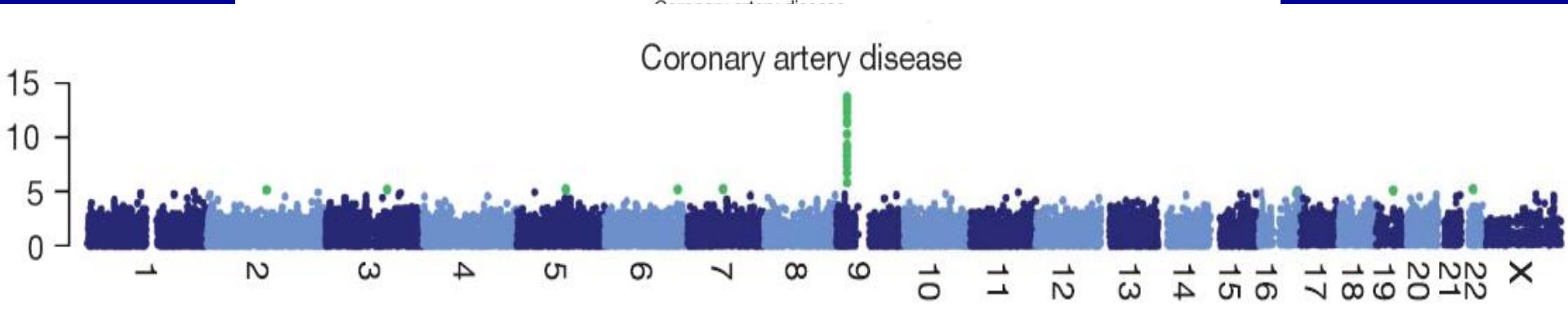
Value and challenges of combining large cohorts

Rory Collins
UK Biobank Principal Investigator

BHF Professor of Medicine & Epidemiology
Nuffield Department of Population Health
Richard Doll Building
University of Oxford, UK

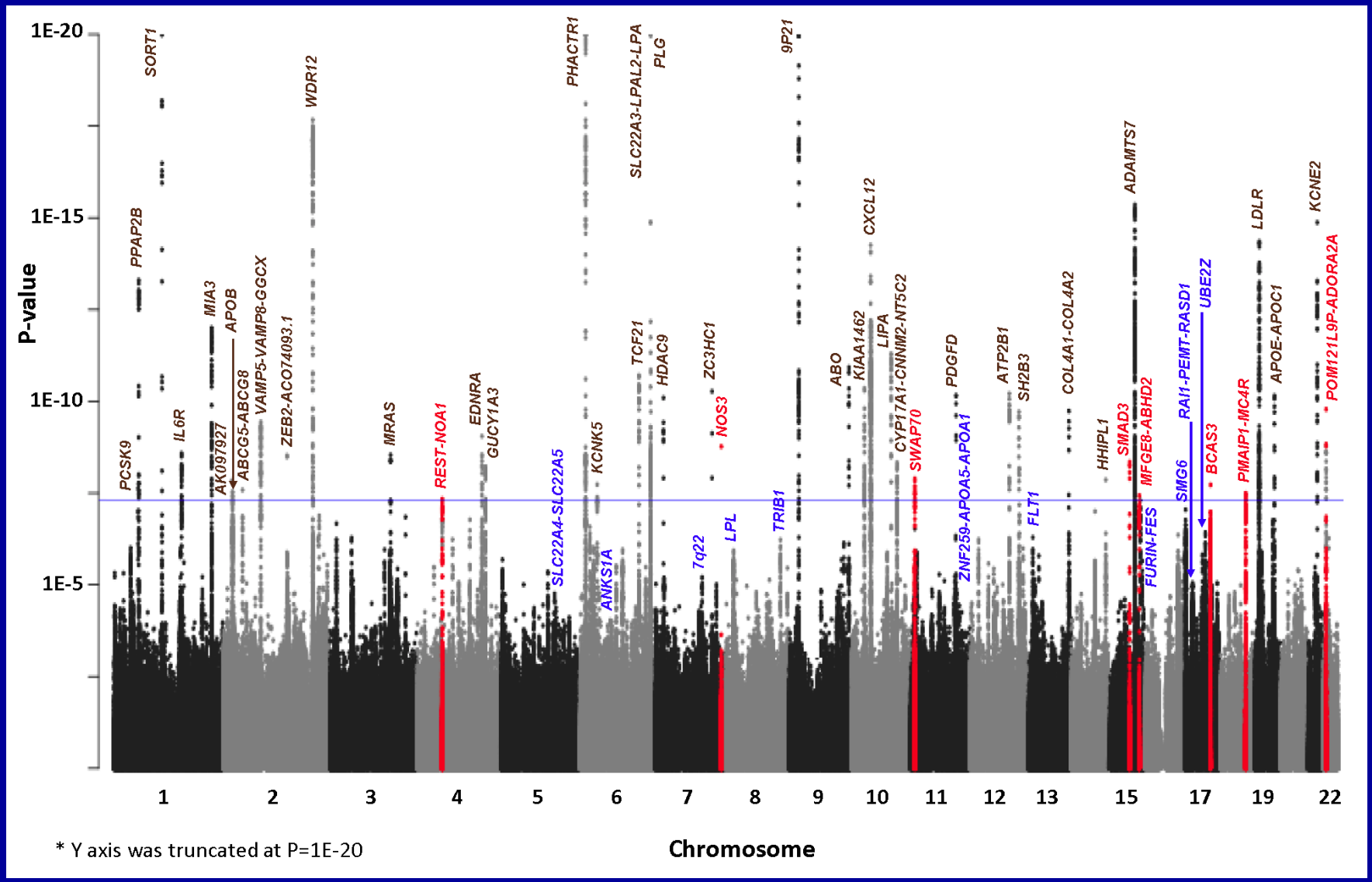
rory.collins@ndph.ox.ac.uk

Wellcome Trust Case Control Consortium: 2,000 cases of 7 conditions and 3,000 controls (Nature 2007)



CARDIoGRAMplusC4D (61K CHD cases;124K controls): 56 GWAS significant genetic loci

(Nature Genetics 2015)



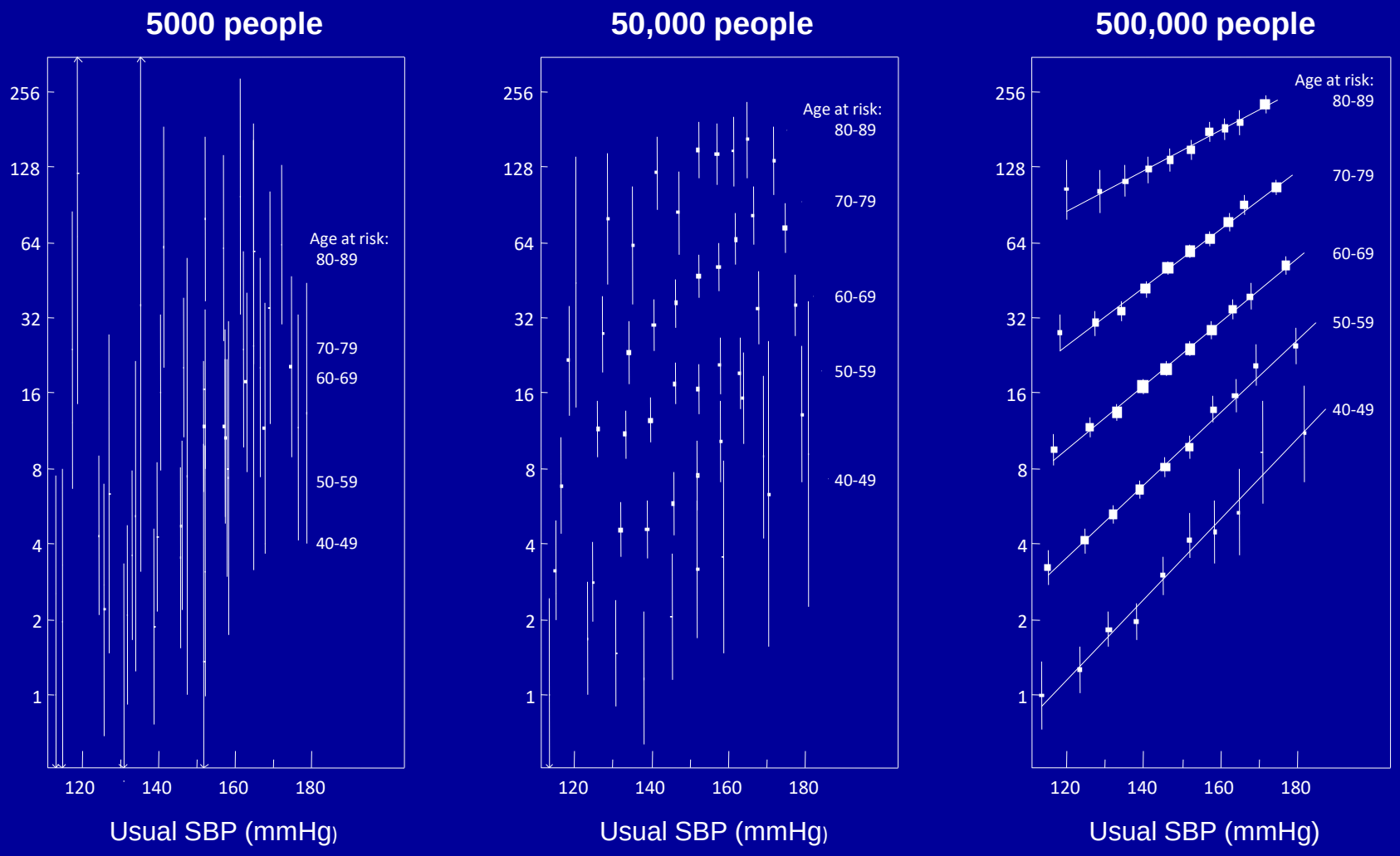
* Y axis was truncated at P=1E-20

Advantages of PROSPECTIVE cohorts for studying the causes of different diseases

- Risk factors can be measured before disease develops, helping to avoid “reverse causality” (important for gene-environment interactions)
- Appropriate controls can be selected from the same population as the disease cases, so that confounding by other factors is less extreme
- Effects of an exposure (e.g. smoking) on many different diseases (e.g. lung disease, cancer, vascular disease, dementia) can be assessed

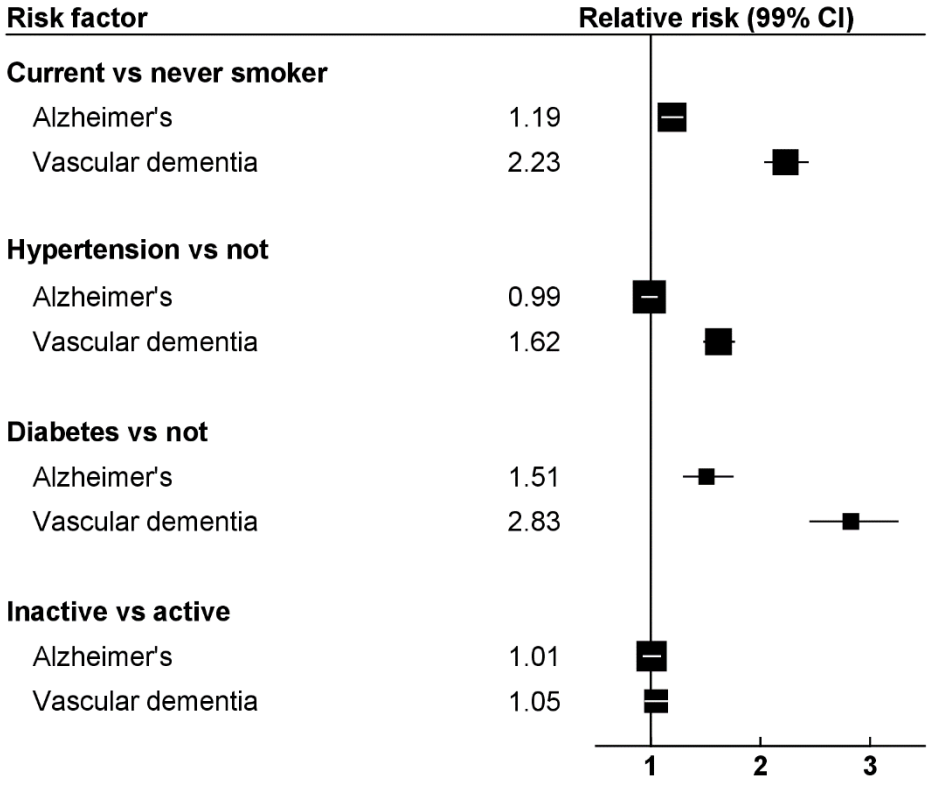
But prospective cohorts need to be LARGE since only a proportion of the participants develop any particular disease during prolonged follow-up

Prospective studies need to be LARGE: CHD versus SBP for 5K *vs* 50K *vs* 500K people



Value of very long-term follow-up in large prospective cohorts: Million Women Study

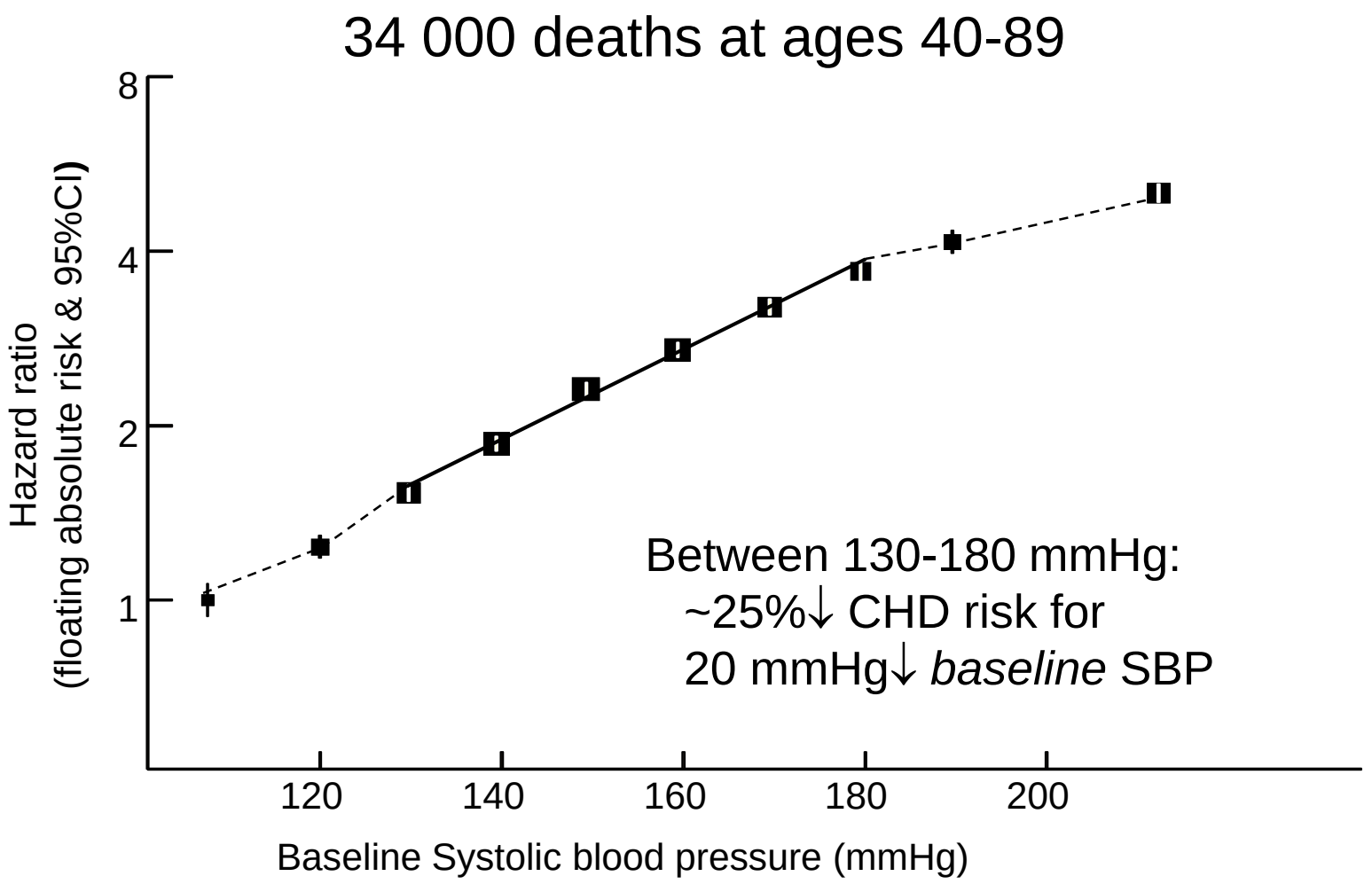
- 1.3 million women (av. age 56y) recruited in 1996-2001; 24,000 with incident dementia after 20 years of follow-up
- Symptoms of dementia cause many to change behaviour long before diagnosis, distorting short-term associations
- Identification of causal factors for dementia without bias requires exclusion of first 10 years of follow-up
- Clear associations with dementia risk can then emerge (especially for vascular dementia)



Importance of correcting for “regression dilution” bias to avoid under-estimation of disease associations

- Associations with measures made at “baseline” tend to underestimate the real associations of disease risk with long-term “usual” levels of such risk factors

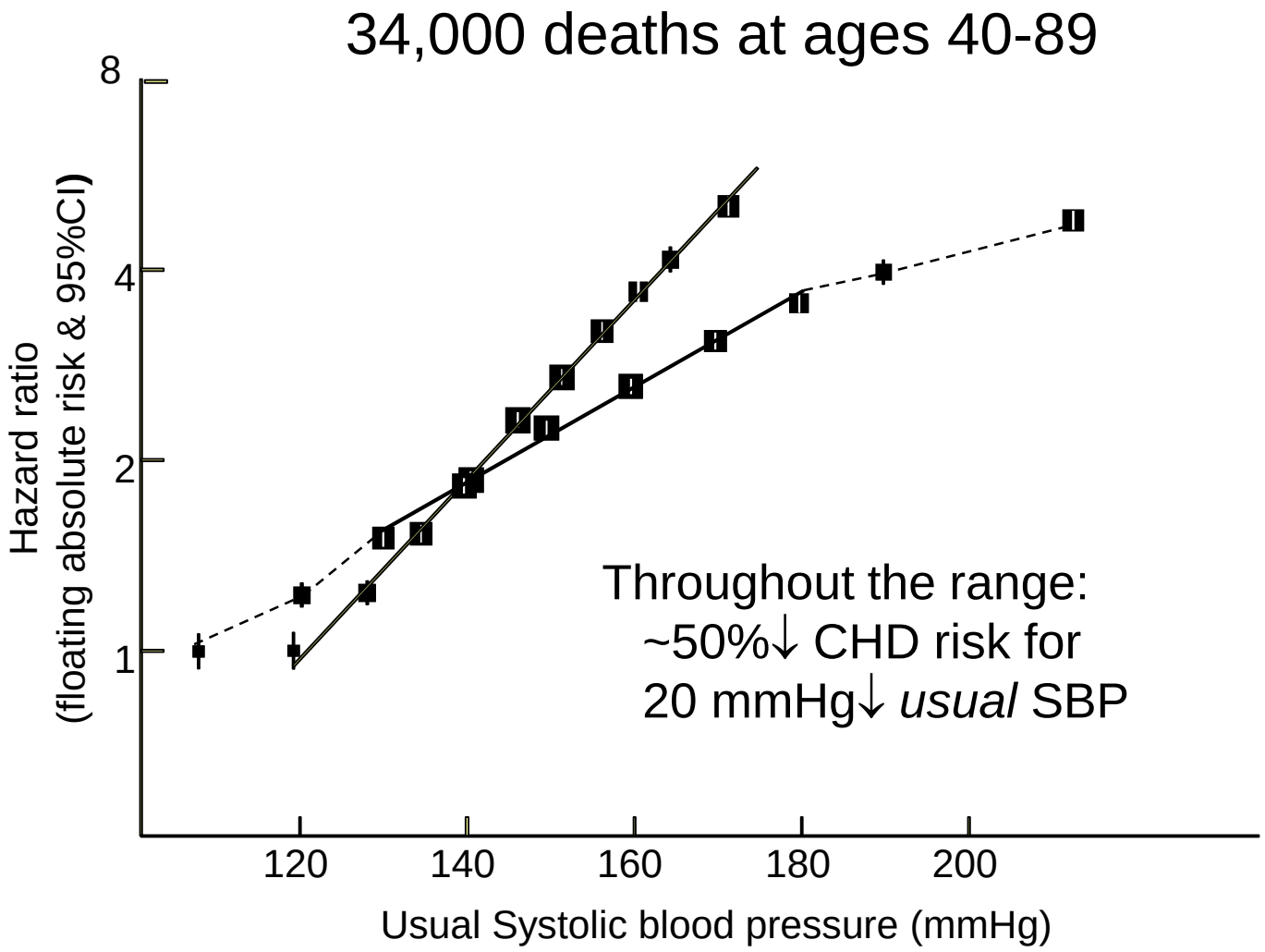
Prospective Studies Collaboration (PSC): CHD mortality versus baseline SBP



Importance of correcting for “regression dilution” bias to avoid under-estimation of disease associations

- Associations with measures made at “baseline” tend to underestimate the real associations of disease risk with long-term “usual” levels of such risk factors
- “Regression dilution” may be caused by measurement errors, by short-term biological variation (e.g. diurnal or seasonal variation) or by longer-term fluctuations (e.g. due to age, activities, diet, disease, treatment)
- Periodic re-assessments of subsets of the participants every few years then allow “time-dependent” regression dilution corrections to be made for multiple factors (as well as “calibration” using more precise assessments)

PSC: CHD mortality versus usual SBP (after adjustment for regression dilution)

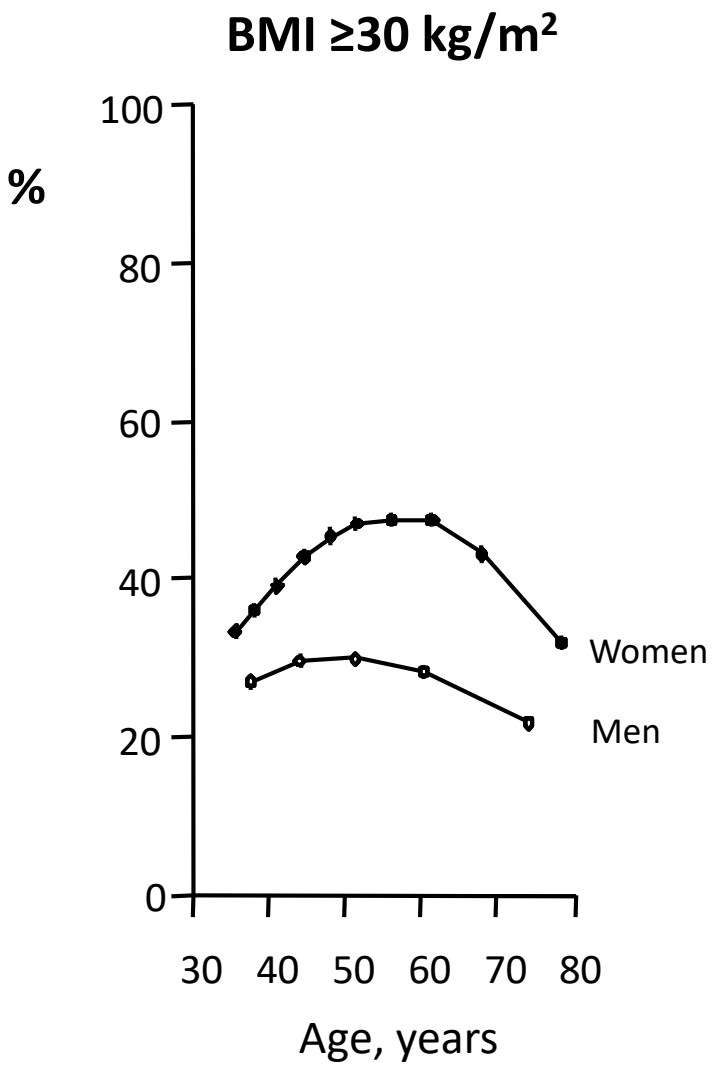


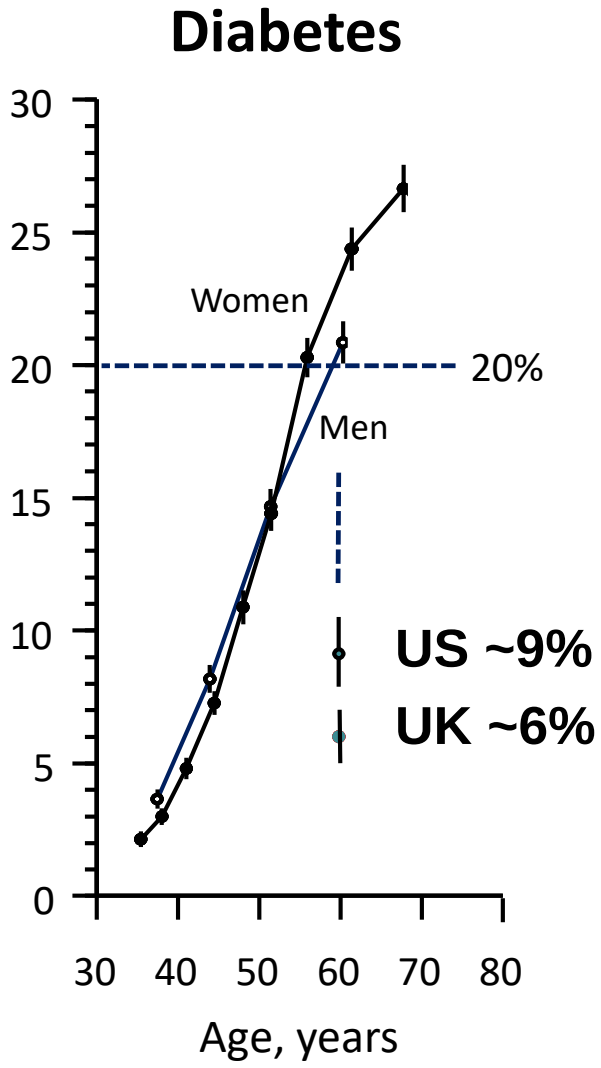
Value of establishing prospective cohorts in different populations

- Complementary studies yield opportunities for replication and for combined analyses
- Genetic heterogeneity between populations may help to assess health-related variants
- Diseases that are rare in one population may be common in another population
- Ability to study a wider range of exposures (e.g. adiposity in China vs UK vs Mexico)

Heterogeneity (and not representativeness) is what is needed to assess the effects of different risk factors across the full range of relevant exposure levels

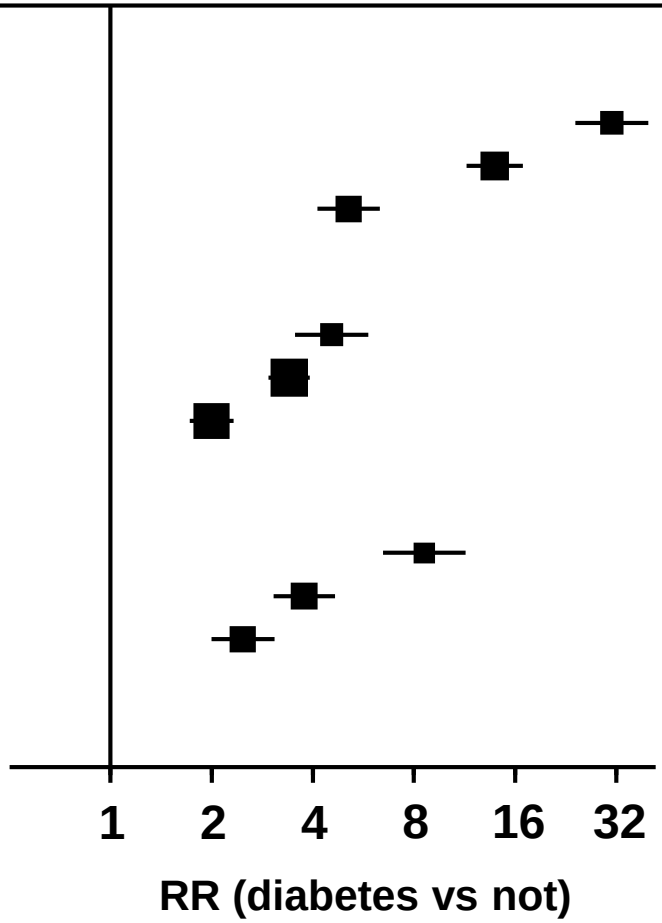
Mexico City Cohort: High prevalence of obesity and of diagnosed diabetes at baseline survey





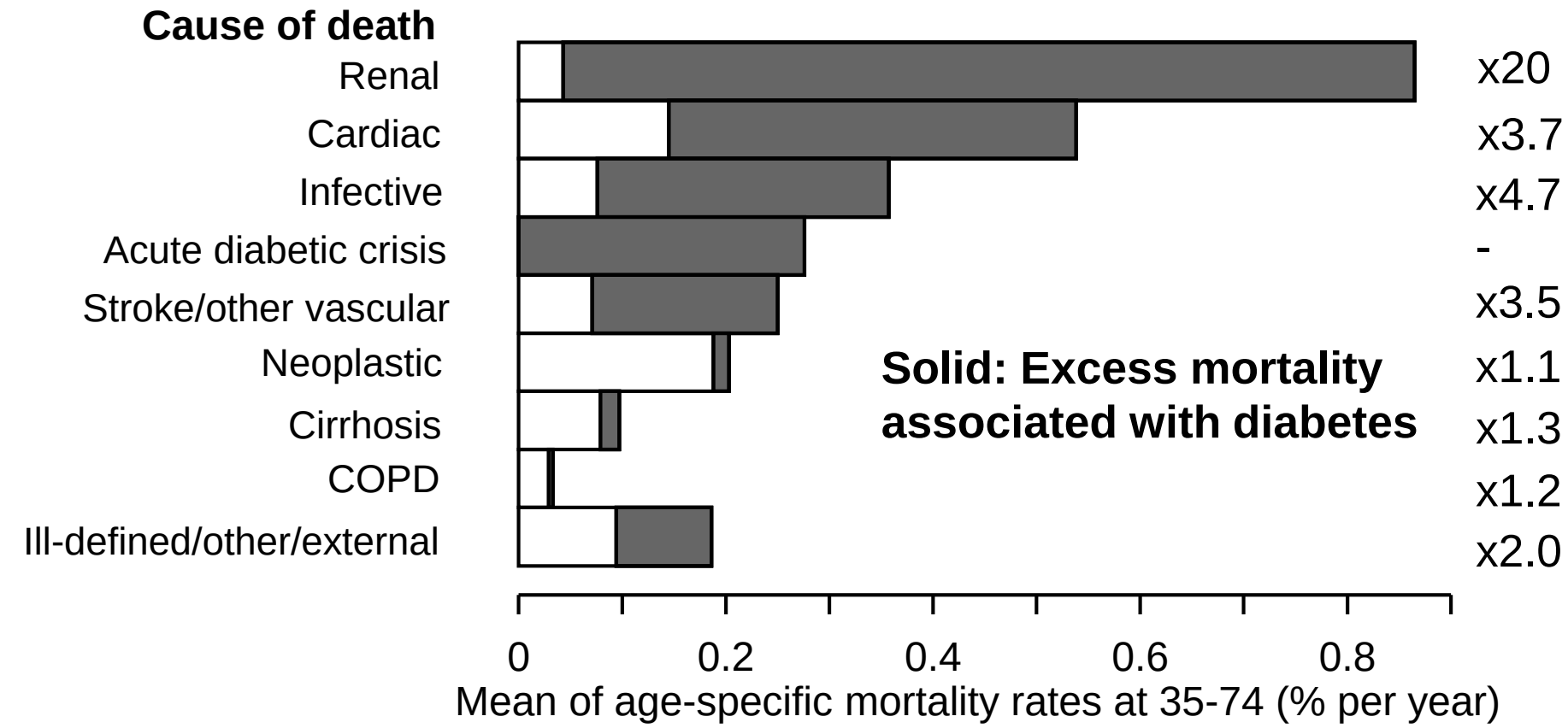
Mexico City Cohort: Relative risks of renal, cardiac and infective mortality associated with DIABETES

Cause of death	Age at risk (years)	No. of deaths	Mortality rate ratio (95% CI) for diabetes vs no diabetes
Renal	35-59	397	31.1 (24.2-39.8)
	60-74	648	13.9 (11.5-16.9)
	75-84	377	5.1 (4.1-6.3)
Cardiac	35-59	360	4.6 (3.5-5.8)
	60-74	823	3.4 (3.0-3.9)
	75-84	830	2.0 (1.7-2.3)
Infection	35-59	242	8.6 (6.5-11.4)
	60-74	369	3.8 (3.1-4.7)
	75-84	384	2.5 (2.0-3.1)



NEJM November 2016

Mexico City Cohort: Main causes of death at ages 35-74 caused by DIABETES



Poor glycaemic control identified as the likely cause of high mortality rates from renal disease and other causes

(leading to a rapid public health intervention in Mexico)

Issues with access to cohorts established to support an extensive range of uses by different researchers

- Insufficient specificity: Lack of detailed characterisation of disease outcomes may either limit utility or result in delays before such information can be made available

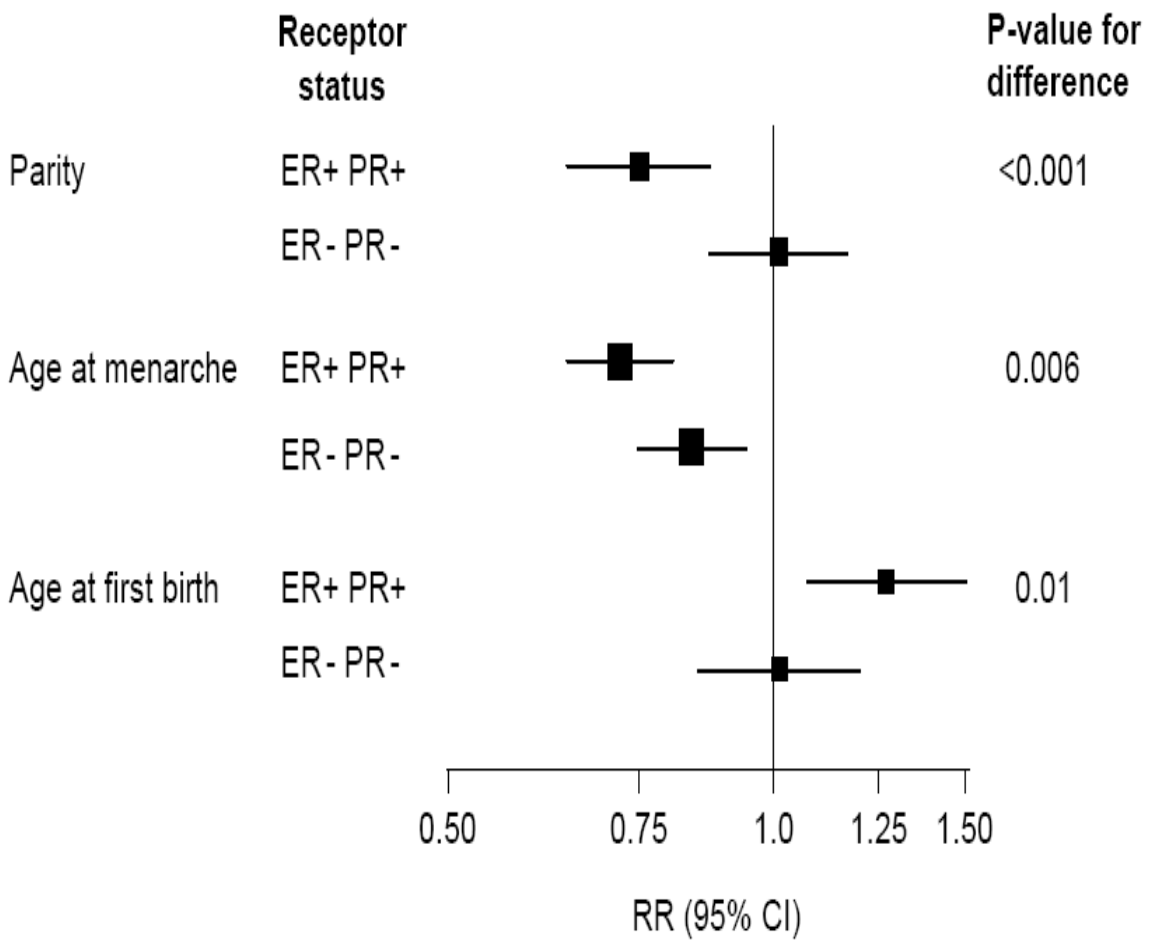
Strategies for scalable phenotyping in large population-based biobanks

- Phenotyping of both the participants and their disease outcomes is important:
  - Retrospective studies tend to focus on phenotyping disease outcomes (as that is how the participants are identified)
  - Prospective studies tend to focus on phenotyping participants (as outcomes are typically detected by linkage to eHRs)
- Prospective studies with 10,000s of many different types of outcome need phenotyping methods that are scalable
- However, approaches for large-scale health outcome phenotyping using eHRs are only now being developed (as in the China Kadoorie Biobank and UK Biobank)

Importance of greater emphasis on “phenotyping” incident health outcomes in prospective cohorts

- Enhancement of power to detect associations between risk factors and disease outcomes
  (false positive diagnoses have main adverse impact)
- Increased specificity of disease classification allows the detection of specific associations
  (i.e. risk factor may only be linked to disease sub-type)

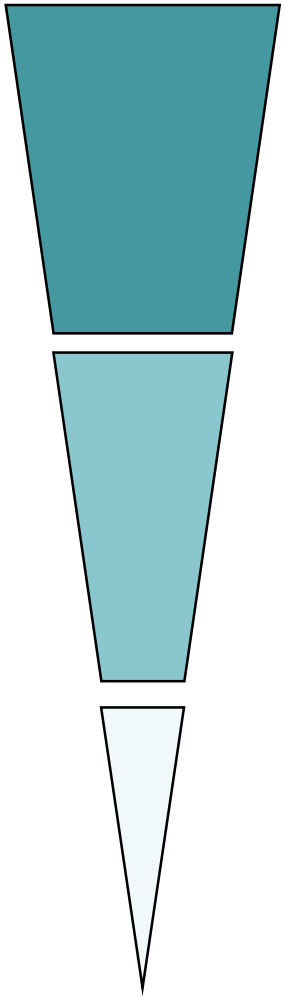
Reproductive factors and breast cancer risk by hormone receptor status of the tumour



Importance of greater emphasis on "phenotyping" incident health outcomes in prospective cohorts

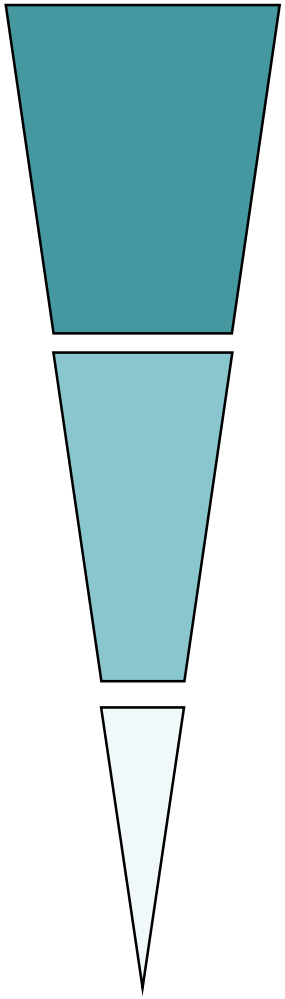
- Enhancement of power to detect associations between risk factors and disease outcomes
  (false positive diagnoses have main adverse impact)
- Increased specificity of disease classification allows the detection of specific associations
  (i.e. risk factor may only be linked to disease sub-type)
- "Future-proofing" of the outcome data so that more detailed phenotyping is possible in future
  (i.e. retain data/samples to allow refined sub-typing)

Scalable approach to ascertaining health outcomes, and subsequent confirmation and classification



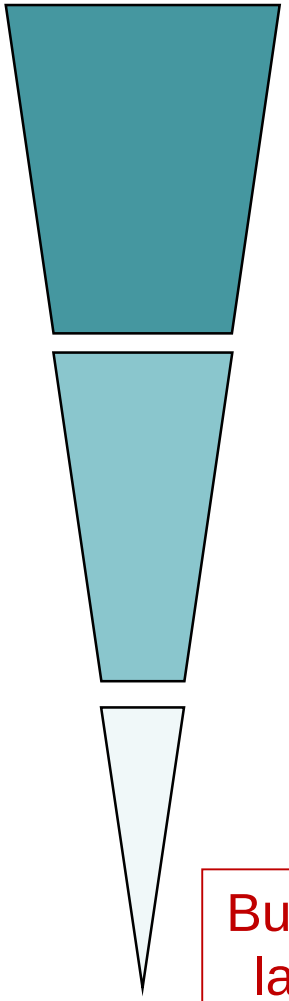
APPROACH	CHARACTERISTICS	POSSIBLE DATA SOURCES
ASCERTAINMENT of suspected cases	Cost-effective Feasible Scalable	Death registers Cancer registers Hospital episodes Primary care records Web-based questionnaires

Scalable approach to ascertaining health outcomes, and subsequent confirmation and classification



APPROACH	CHARACTERISTICS	POSSIBLE DATA SOURCES
ASCERTAINMENT of suspected cases	Cost-effective Feasible Scalable	Death registers Cancer registers Hospital episodes Primary care records Web-based questionnaires
CONFIRMATION of “case-ness”	As above, but greater cost/lower feasibility	Cross-referencing e-records Disease registers

Scalable approach to ascertaining health outcomes, and subsequent confirmation and classification



APPROACH	CHARACTERISTICS	POSSIBLE DATA SOURCES
ASCERTAINMENT of suspected cases	Cost-effective Feasible Scalable	Death registers Cancer registers Hospital episodes Primary care records Web-based questionnaires
CONFIRMATION of “case-ness”	As above, but greater cost/lower feasibility	Cross-referencing e-records Disease registers
CLASSIFICATION of disease cases	More involved and costly per case	Review of clinical records Tumour collections/assays (e.g. sequence data) Specialised databases (e.g. imaging)

But there are no established processes for large-scale health outcome phenotyping

Issues with access to cohorts established to support an extensive range of uses by different researchers

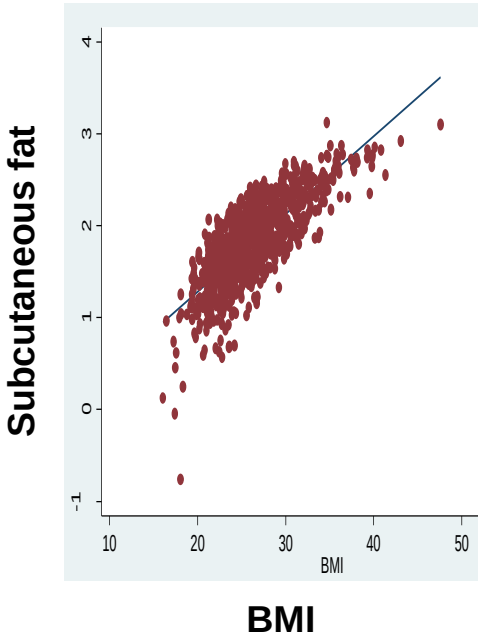
- Insufficient specificity: Lack of detailed characterisation of disease outcomes may either limit utility or result in delays before such information can be made available
- Data inaccessibility: Researchers may not be able to handle complex data (e.g. from imaging) and, instead, need it to be converted into accessible “information”

Similar age, gender, BMI & % body fat, but different amounts of INTERNAL FAT

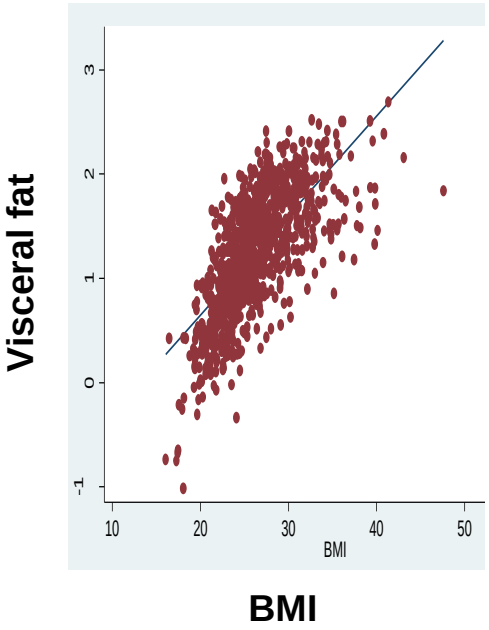
5.86 litres of internal Fat

1.65 litres of internal fat

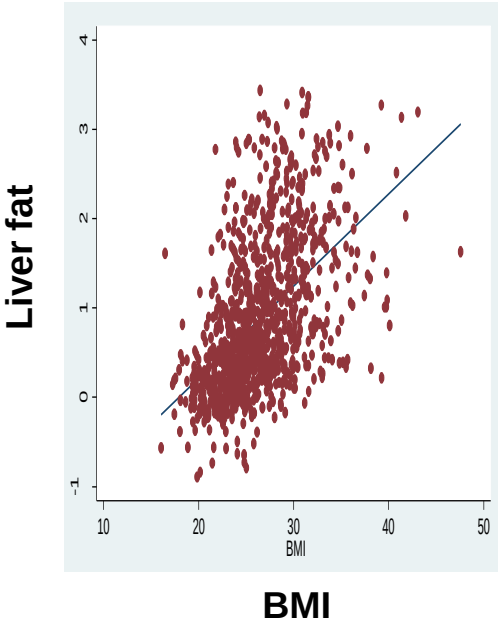
Abdominal MRI body fat measures versus BMI (based on automated MR analyses by AMRA)



Subcutaneous adipose tissue (r=0.77)



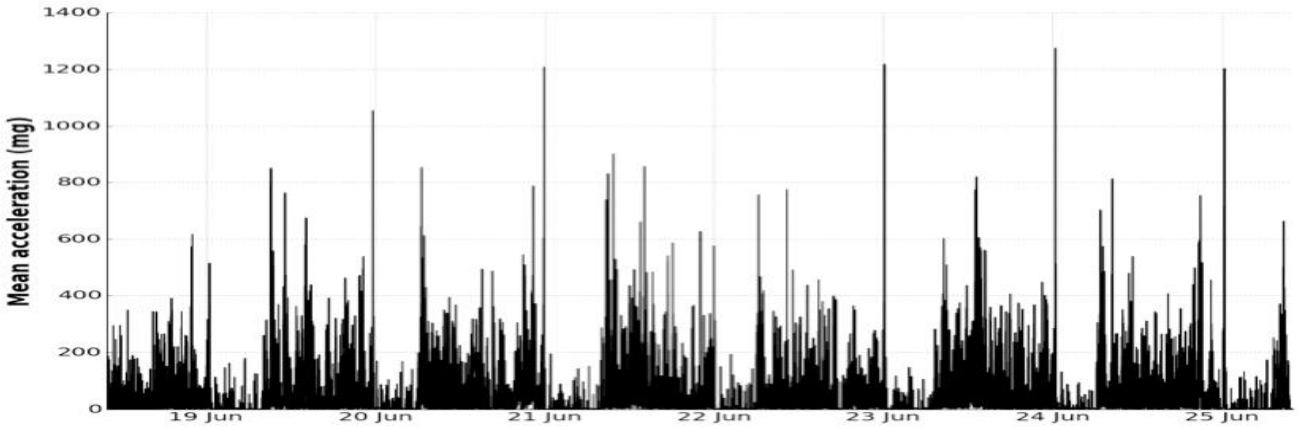
Visceral adipose tissue (r=0.69)

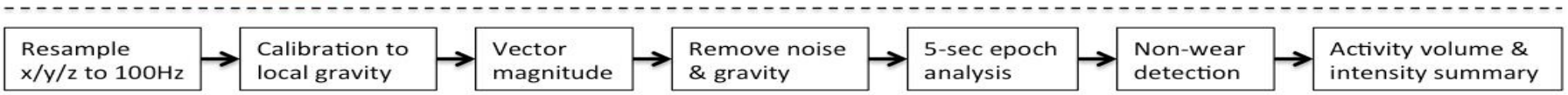


Liver fat (r=0.51)

UK Biobank activity monitor project

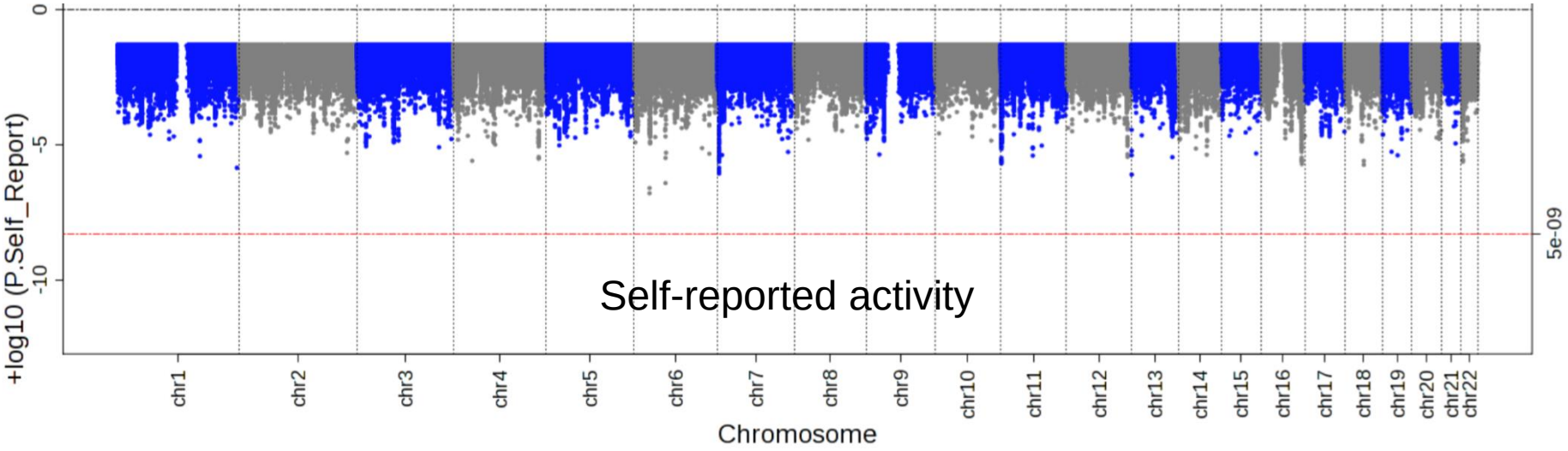
- 103,712 participants
- 7 days data per participant
- 100Hz tri-axial acceleration data
- 180 million movement readings per participant





Doherty 2017; PLoS ONE 12(2): e0169649

Accelerometer assessment of activity versus self-reported activity enhances biological understanding



Doherty et al. 2017 bioRxiv

Issues with access to cohorts established to support an extensive range of uses by different researchers

- Insufficient specificity: Lack of detailed characterisation of disease outcomes may either limit utility or result in delays before such information can be made available
- Data inaccessibility: Researchers may not be able to handle complex data (e.g. from imaging) and, instead, need it to be converted into accessible “information”
- Depletable sample: A “resource” needs to be able to provide appropriate samples for a wide range of uses or, preferably, the results of a wide range of assays

Advantages of “whole cohort” sample assays: reduce depletion and increase accessibility

- Uncontrolled assays may deplete the available sample rapidly, preventing subsequent studies
- Assays conducted in separate subsets of a cohort (e.g. “nested” case-control strategies) may yield assay data that are not comparable
- By contrast, assays conducted in the whole cohort support many different comparisons, as well as minimising depletion, improving quality control and being cost-effective

But, such “whole cohort” strategies are costly and typically restricted to “standard” assays (although industry investment can help)

Strategies for improving data utility (accessibility and interoperability)

- Improve how researchers can visualise and navigate data resources (e.g. search by disease area or data type, or support more visual exploration of the data)
- Not sustainable to move increasingly large datasets to researchers, and instead analysis by researchers within the data resource environment needs to be supported
- Open source data platforms (e.g. Biospheres, AllofUs), as well as commercial offerings, are being developed; cloud computing is becoming a cost-effective enabler
- Alignment to data standards encourages interoperability and facilitates analyses across several different cohorts

Strategies for improved biobanking (with open access for researchers)

- Establish several large-scale prospective studies to assess the full effects of genes, environment and lifestyle on many different health outcomes
- Enhance baseline phenotyping of large subsets (including remote assessment and imaging), with repeat assessments in subsets during follow-up in order to allow “regression dilution” correction
- Establish centralised systems for health follow-up (including direct assessment of participants), and develop scalable outcome phenotyping strategies
- Enhance “visualization” of increasingly complex databases and develop analysis platforms that facilitate access by many different researchers