BREADTH VS DEPTH OF PHENOTYPING

Julie E. Buring, ScD
Professor of Medicine
Brigham and Women’s Hospital
Harvard Medical School

Importance of Phenotyping

- No question about importance of needing adequate phenotyping data.
- Genetic variants alone don't account for all chronic disease; important to evaluate environmental as well as gene-environment interactions.
- It not identified, could mask the detection of a genetic effect, or lead to inconsistencies between populations with different environments.
- Can suggest approaches for modifying the effects of genes by avoiding or modifying the appropriate environment (prevention, treatment).

Breadth vs Depth of Phenotyping

- Certainly there are trade-offs in terms of logistical issues: depth of information inversely related to sample size and directly related to cost.
- But a resource itself is not simply all “broad” or “deep. That is a judgment relative to the questions being evaluated. Same cohort can have phenotypes broad enough to allow a substantial scientific contribution for some questions, and/or deep enough for others.

Women’s Health Study – as an Example of a Cohort Contributor to the Scientific Commons

- Women’s Health Study (WHS) designed as randomized trial of aspirin and vitamin E in primary prevention on TWO outcomes: cancer and CVD. Jointly funded, NCI and NHLBI.
- 39,874 participants: trial began in 1992, ended in 2004, then followed observationally to present (mean follow-up, 18 years)
- Participants throughout US, follow-up conducted entirely by mail from research office: not distributed, local sites, in-person visits. Staff are cross-trained, can move from study to study during different phases and stages of investigations.

Women's Health Study

- From beginning of study, had second goal: to maximize the potential of the cohort to be a resource when the trial was over, that could be used by many investigators over time for a wide range of health-related outcomes.
- We did not know what questions would arise in future, but knew wanted to be able to use this resource to contribute to the evaluation of whatever questions did arise as important, in a timely and cost-efficient way.

Women's Health Study

- Our basic strengths would be: large sample size of women, geographically distributed throughout the US, with extensive duration of follow-up, and phenotypic data from both baseline as well as from regular recontact through yearly follow-up q'aires.
- Decided to add to questionnaires, at minimal cost, wide range of self-reported outcomes: physician-diagnosed arthritis and other connective tissue disorders, diabetes, visual disorders (cataract, AMD), cognitive decline, venous thromboembolism, osteoporosis, neurologic conditions, migraine, etc.

Women's Health Study

- Also made sure had at baseline, extensive core group of demographic, lifestyle, and medical history variables relevant to a wide range of outcomes.
- Represent experiences as adults, as well as some full history variables (like reproductive, smoking, hormone use) , to allow assessment of prevalence at baseline and then change over follow-up.
- Also, obtained baseline plasma and buffy coat samples prerandomizaton from 28,345 participants: aliquoted, frozen and stored in nitrogen freezers.
- Then we conducted the trial, and meanwhile tried to leverage the resource….

Women’s Health Study

- Ancillary studies were funded for specific conditions (non-trial endpoints): to include deeper information on details of diagnoses (medical records, path reports, tissue blocks), send additional risk factor q’aires, and assay bloods for specific biomarkers.
- Ancillaries also funded to apply more updated assessment methods, such as accelerometers for physical activity.
- No problem with burden of recontact – losses to cohort in WHS through death, not LTFU.

Women's Health Study

- Received nonfederal foundation support to maintain and expand biorepository:
  - All plasma samples assayed for expanded array of biochemical markers (lipids, inflammatory, hematologic).
  - Extracted DNA on all samples and evaluated individual hypotheses.
  - Received additional funding from NHLBI to add genetic components to risk development score model, that allowed us to conduct GWAS on all blood samples (completed in 2008).

Women's Health Study

- Didn't know what the questions of importance were going to be. Picked GWAS as believed best approach we could do at the time for the money. Illumina chip, more than 360,000 SNPs including additional panel of CVD-relevant markers.
- Did on entire sample. Could have kept samples in freezer, conducted nested case-control studies as needed. But by adding genetic component then, and on everyone, we could be ready to go as needed. Also allowed us to do a case-cohort analysis, which has more power for multiple outcomes, risk prediction.

Women’s Health Study

- Was embedding this genetic component into mature cohort too late?
- Assessing risk and identifying interactions requires sufficient f-up for adequate number of new cases to accrue. GWAS completed about 4 years ago; stage of cohort where numbers of events increase exponentially because population aging.
- Total cancers: 3735 confirmed events between 1992 and 2008 (16 years), now 5063 (1328 new cases in 4 years).
- Important vascular events: 1285 (16 years) to 1663 (378 new confirmed cases in 4 years).

Women’s Health Study

- How fruitful were the data collected? Look at participation in consortium activities as an indicator - and for what range of outcomes.
- For some, contributed fully and had sufficient levels of data for all risk factors needed (eg. not only for cancer and CVD, but consortia such as BMI/mortality, physical activity, and hypertension).
- For others, not all variables but more than enough for basic set: reproductive (menopause, menarche, parity), migraine, glaucoma, neurologic outcomes, glioma, liver cancer, pancreatic cancer.
- For others, invited but could not provide data without supplemental information requiring recontact (fertility, hearing loss).
- Last 2 full years: 2011 – 20% of our publications were consortial; 2010 – 36%

Women's Health Study

- Here example of one study being broad or deep for particular phenotypes.
- Exogenous hormones and breast cancer: We had full history at baseline for all participants. On each yearly follow-up questionnaire, we assessed change in menopausal status and change in hormone use over time; did an hormone biomarker panel in nested case-control design for cases of breast cancer; got pathology report plus pathology slides and tissue block. Obesity is a confounder: we had weight and height at baseline, repeated questions every 2 years, and sent tape measure to a subgroup to get waist:hip ratio. Depending on the question and analysis, could utilize these variables as broadly or as deeply as needed.

Harmonization

- Not having standardized measures between studies of phenotypic and environmental information results in significant challenges in merging data in a valid way. Need to limit analysis to "lowest common denominator" of available variables.
- But important to remember that this set of variables is usually sufficient - and variables can be externally harmonized.
- Harmonization is not mandated questions or wording. Studies provide core information to central group, used for similar but not identical definitions of exposure and disease measures to permit pooling of data. Eg. if cohorts all have height and weight, harmonize actual data to define BMI with common cutpoints. Same for smoking, drug use, etc.
- Need to understand the question being evaluated, and assess not what would be perfect, but what will be adequate (ie. smoking, history for lung cancer vs current for CVD) .

Women’s Health Study

- We had to do harmonization for our data for cancer consortium . Data for million people so far in cancer consortium have been harmonized , and it didn’t take long nor was costly. It is do-able, both scientifically and logistically.
- Remember: the time and money invested in a cohort study is in the setting up the population and collecting the data. After that, using the data to contribute to various activities takes reasonable amounts of time and money .

Lessons to be Considered from WHS

1. Many individual cohorts have ability to contribute to analyses of multiple outcomes: added value situation.

2. Genetic component such as sequencing can be added to existing studies, as long as have adequate follow-up to accrue sufficient numbers of outcomes. Maybe can’t do on entire cohort, but powerful even in just a subset.

3. Best guide for adequate levels of phenotypes (how broad vs. deep) comes if the outcomes and hypotheses to be evaluated are known.

Lessons to be Considered from WHS

- But even if hypotheses to be evaluated in future can’t yet be specified, can take a middle road.
- We have experience with set of core basic variables we will need for exposures – anthropometric, smoking, ETOH, medical history, reproductive history, family history, physical activity. Build these into every study.
- In general, if raising hypotheses, broader is more appropriate; testing hypotheses, deeper.

Lessons to be Considered from WHS

4. There are scientific and logistic disadvantages of existing cohort studies. Sometimes can be addressed by obtaining additional data and specimens. eg. recontacting the participants, or, as our WHS population all reach age 65, can begin to use Medicare/Medicaid tapes to assess hospitalizations and diagnoses.

- There are advantages to initiating new cohorts. But first need to carefully identify and address the gaps to be filled, and how the new cohorts will be used to do so. Have to consider adequate sample size and nature of the population in light of questions to be addressed. Eg. adolescents – represented? or in adequate numbers to compare them vs children vs adults?

Lessons to be Considered from WHS

- To establish research portfolio to answer many potential questions in future, I believe we should leverage existing cohorts while developing any new ones. Knowledge will come from accumulated data, not one study.
- We don't need or want to wait – we don't have everything on all cohorts, but we do have enough on most.