‘Big Data for Health’ strategy at the Dalla Lana Faculty of Public Health

Prabhat Jha, on behalf of the big data for health working group of DLSPH
(Chairs David Henry and Prabhat Jha)

Prabhat.jha@utoronto.ca
Twitter: Countthedead

‘Wide’ and ‘Deep’ data

- Deep
  - Genome wide analysis
  - Proteomics, metabolomics, microbiomics
  - Functional MRI
  - Geospatial-linked exposures
- Wide
  - Population data, inc mortality
    - Linked at the level of the individual
    - Administrative data
    - Electronic health records
    - Registries
- COMMON to Both: *PLATFORMS* to enable *IMAGINATIVE* linkages by diverse brains

cghr

MILLION DEATH STUDY IN INDIA

1. Visit 1 M homes ("true snapshot" of India) with a recent death & ask standard questions and get a narrative
2. Use non-medical surveyors (electronic entry + GPS)
3. Web-based double coding by 500 doctors (guidelines + adjudication and other strict quality control)
4. Study all diseases, work with census dept, keep costs

MILLION DEATH STUDY: selected results

(M=Millions, K=thousands)

- **4-12M girls aborted before birth since 1980 (1/2 of these since 2000)**
- **1M smoking deaths (more than expected)** and 0.1M alcohol deaths
- **200K malaria deaths: WHO predicted only 15K**
- 100K HIV deaths: UNAIDS predicted 400K
- 60K pedestrian traffic deaths: Police estimate=9K
- 50K snakebite: WHO worldwide estimate=50K
- 33K cervical cancer: only 7K at Kashmir/Assam rate
- **Each common disease is rare somewhere in India, & hence is largely avoidable**

cghr

81 districts are home to 37% of the national deaths in children < 5 years

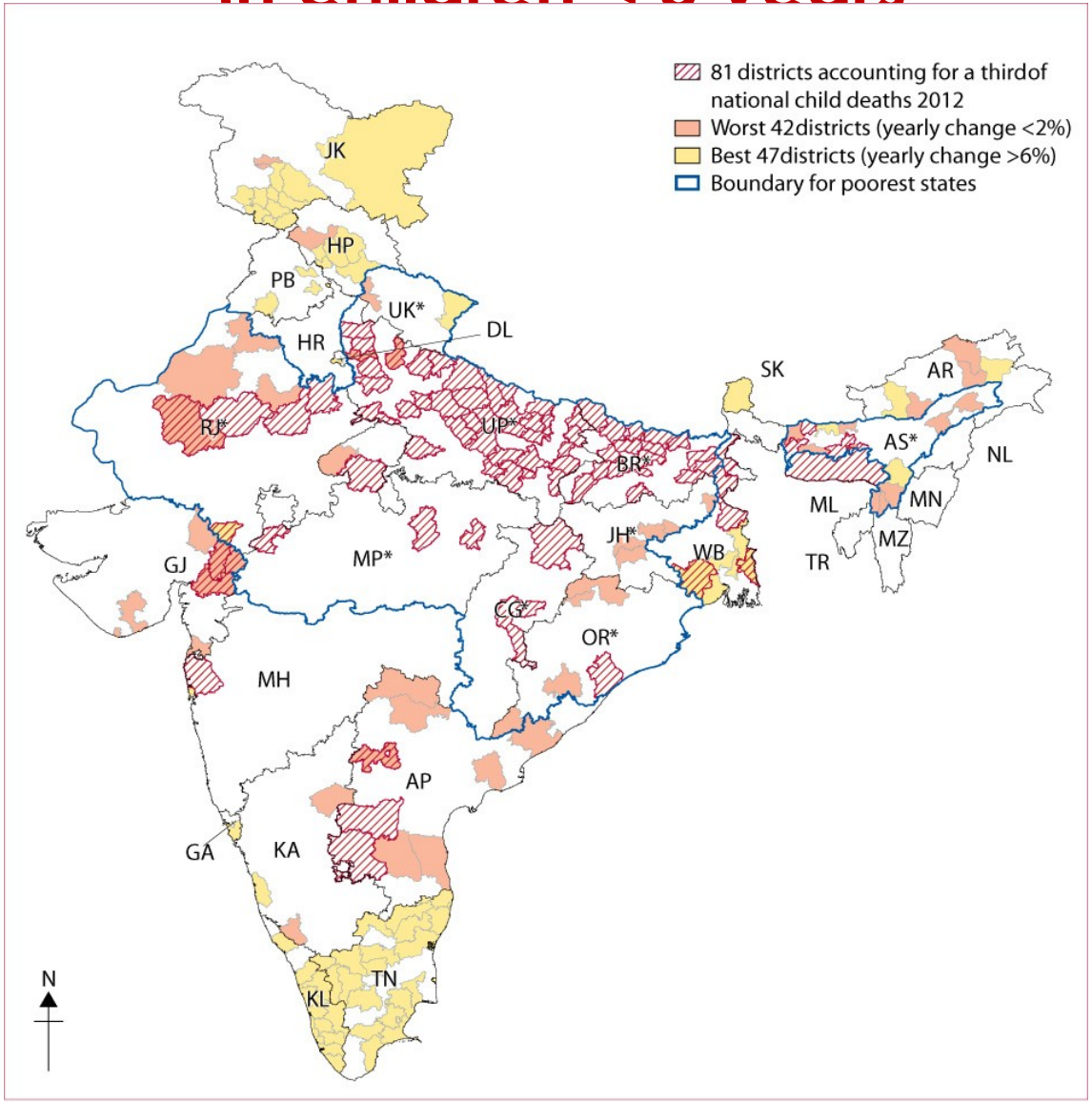
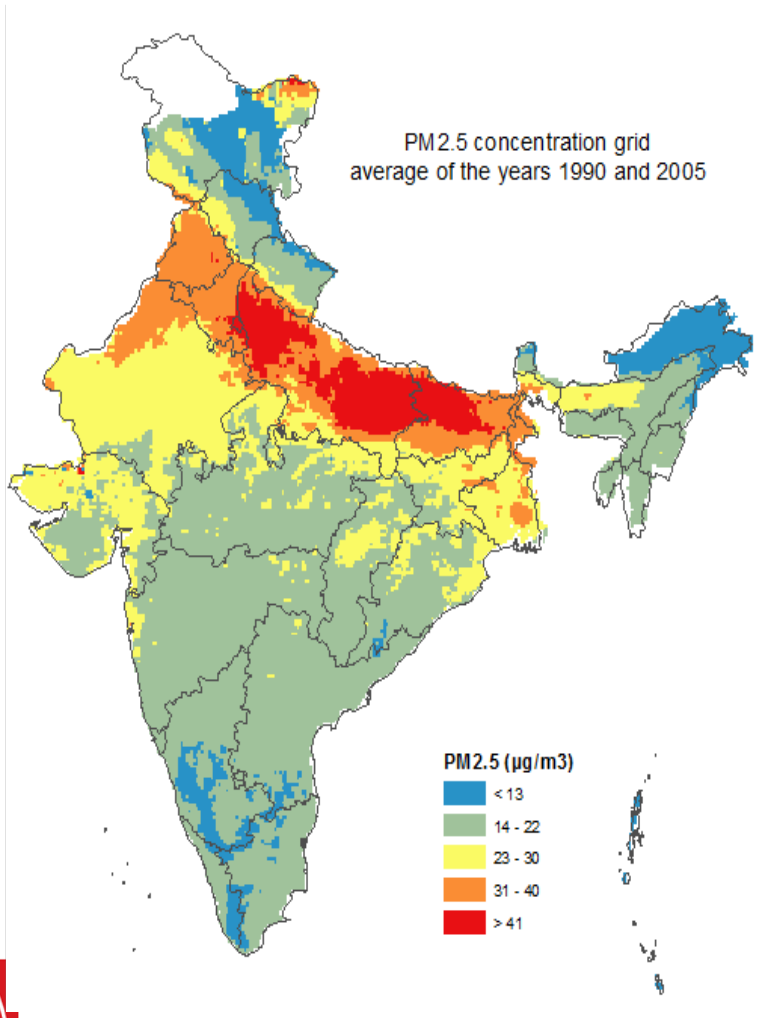


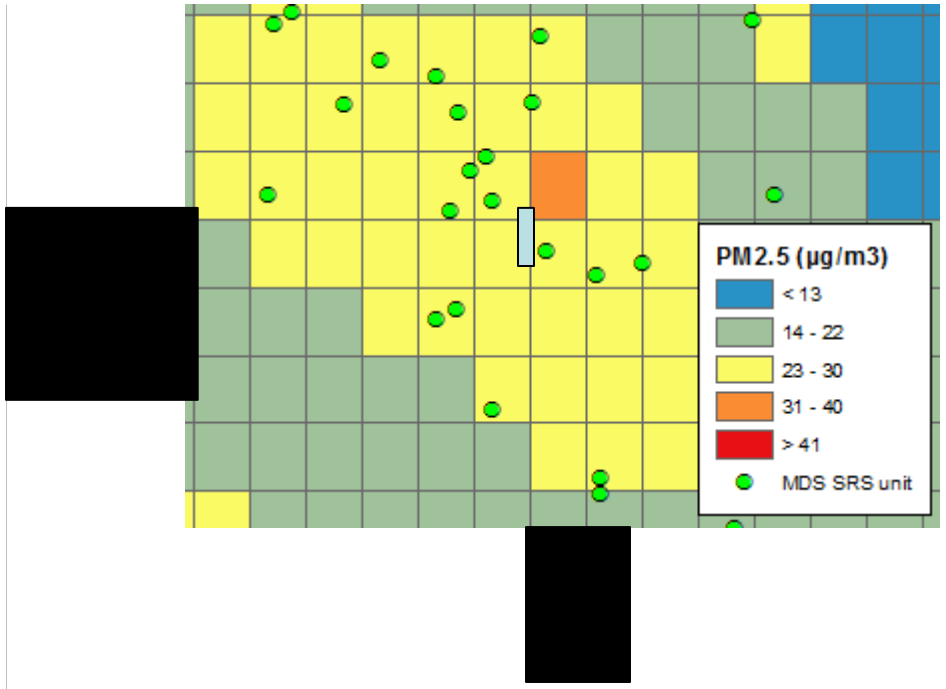
Figure 2: Best and worst Indian districts according to change in under-5 mortality (2001–12), and districts with the highest third of under-5 mortality in 2012

68 of these 81 districts are in poorer states

cghr

Geo-code this!
Air pollution and mortality





Nationally representative estimates of ambient and indoor air pollution mortality (Half about WHO estimates!)

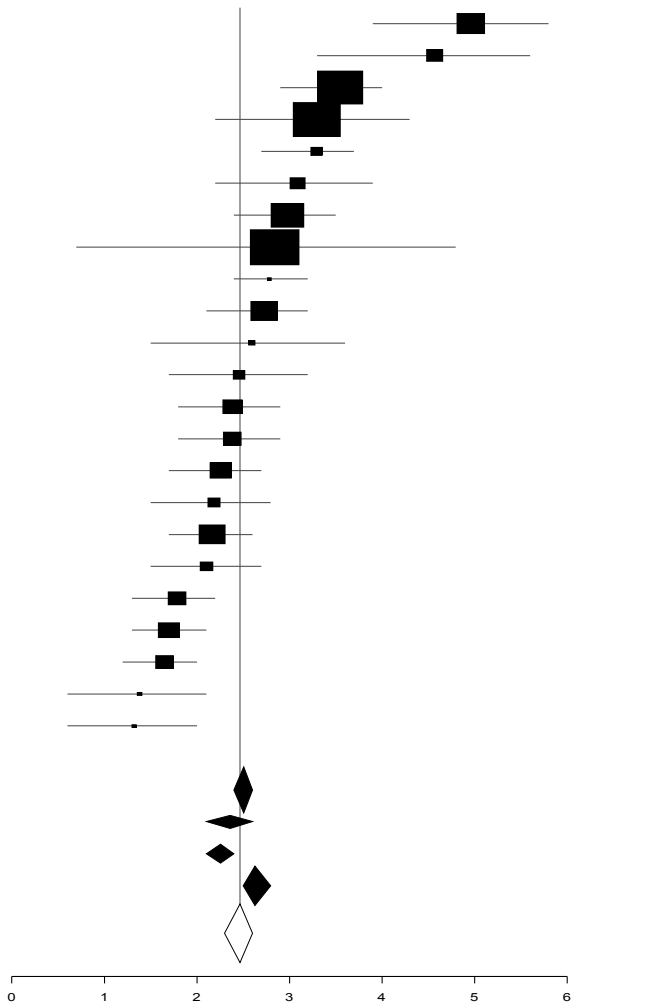
Source: Yurie Maher, in press

cghr

Cancer (non tobacco/non infection):

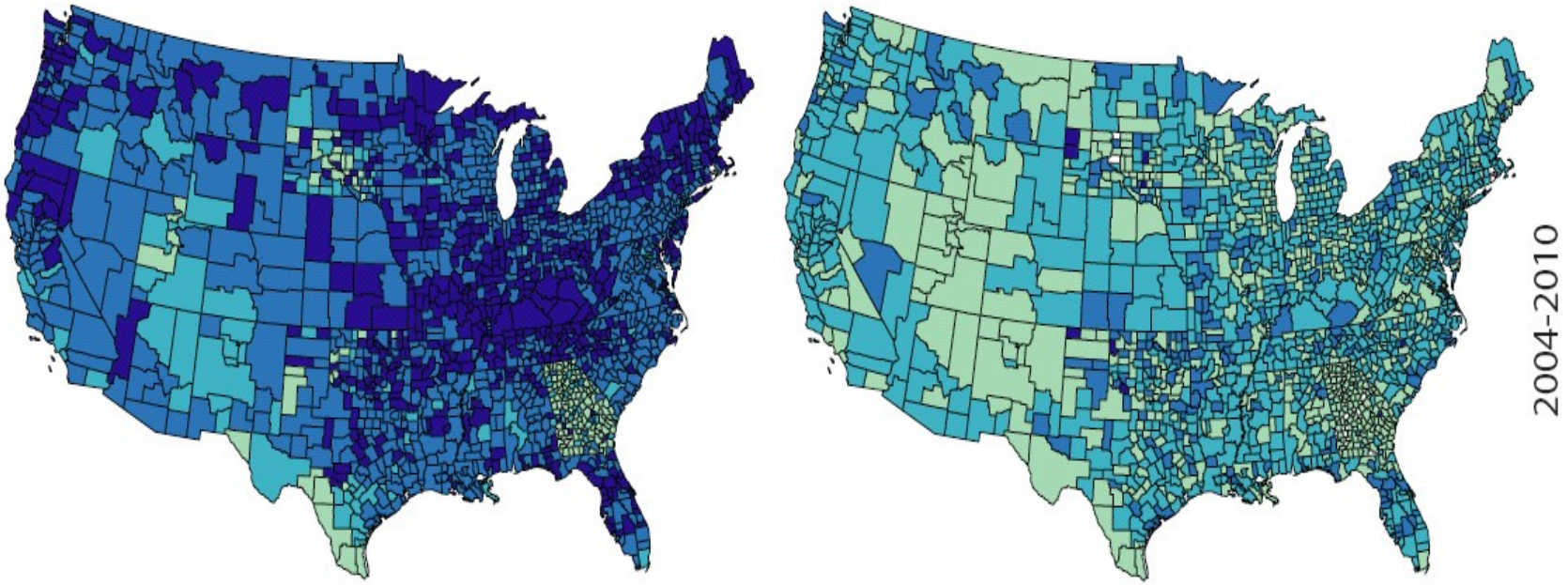
BOTH GENDERS aged 30-69 years

State or Region	Study deaths of non-tobacco related cancers common to both genders: Total / Male / Female	Estimated cancer deaths (000s) in2010	Age standardised cancer mortality rate per 100 000 (99% CI)	Cumulative risk (99%CI)
Northeast states	165 / 89 / 76	4·6	103·1	5·0 (3·9 – 5·8)
Jammu & Kashmir	96 / 50 / 46	3·3	89·3	4·6 (3·3 – 5·6)
Kerala	273 / 183 / 90	10·9	71·2	3·5 (2·9 – 4·0)
West Bengal	280 / 152 / 128	23·1	69·2	3·3 (2·2 – 4·3)
Himachal Pradesh	65 / 29 / 36	1·7	65·8	3·3 (2·7 – 3·7)
Assam	91 / 50 / 41	6·2	63·6	3·1 (2·2 – 3·9)
Karnataka	195 / 111 / 84	14·3	62·9	3·0 (2·4 – 3·5)
Uttar Pradesh	291 / 126 / 165	35·2	60·1	2·8 (0·7 – 4·8)
Uttarakhand	12 / 5 / 7	1·7	53·5	2·8 (2·4 – 3·2)
Andhra Pradesh	159 / 82 / 77	19·3	58·4	2·7 (2·1 – 3·2)
Delhi	37 / 16 / 21	3·2	53·8	2·6 (1·5 – 3·6)
Punjab	69 / 29 / 40	4·8	49·0	2·5 (1·7 – 3·2)
Gujarat	114 / 63 / 51	10·7	50·4	2·4 (1·8 – 2·9)
Madhya Pradesh	107 / 56 / 51	10·5	48·2	2·4 (1·8 – 2·9)
Rajasthan	125 / 64 / 61	9·6	47·4	2·3 (1·7 – 2·7)
Haryana	72 / 42 / 30	3·9	47·0	2·2 (1·5 – 2·8)
Tamil Nadu	156 / 86 / 70	14·1	46·9	2·2 (1·7 – 2·6)
Other states	75 / 50 / 25	1·0	45·3	2·1 (1·5 – 2·7)
Maharashtra	106 / 56 / 50	15·9	38·2	1·8 (1·3 – 2·2)
Bihar	125 / 59 / 66	11·5	37·3	1·7 (1·3 – 2·1)
Odissa	106 / 56 / 50	5·8	37·0	1·7 (1·2 – 2·0)
Chhattisgarh	21 / 9 / 12	2·0	27·9	1·4 (0·6 – 2·1)
Jharkhand	23 / 14 / 9	3·3	30·5	1·3 (0·6 – 2·0)
Rural	2157 / 1136 / 1021	155·4	53·2 (50·0 - 56·3)	2·5 (2·4 – 2·6)
Urban	606 / 341 / 265	61·0	50·0 (45·3 - 54·7)	2·4 (2·1 – 2·6)
Poor	901 / 439 / 462	85·7	48·2 (44·4 - 52·1)	2·3 (2·1 – 2·4)
Rich	1862 / 1038 / 824	130·7	56·0 (52·4 - 59·6)	2·6 (2·5 – 2·8)
Total	**2763 / 1477 / 1286**	**216·4**	**52·7 (50·0**	**2·5 (2·3 – 2·6)**



Source: Dikshit et al, Lancet 2012

Tobacco deaths (% of total), WOMEN, US, 2004-2010 by county (0.4M deaths/year



Evaluate: smoking, obesity, health insurance (Obamacare) on mortality changes
Extend to 1.7 M deaths in Ontario and 4 M deaths in Mexico

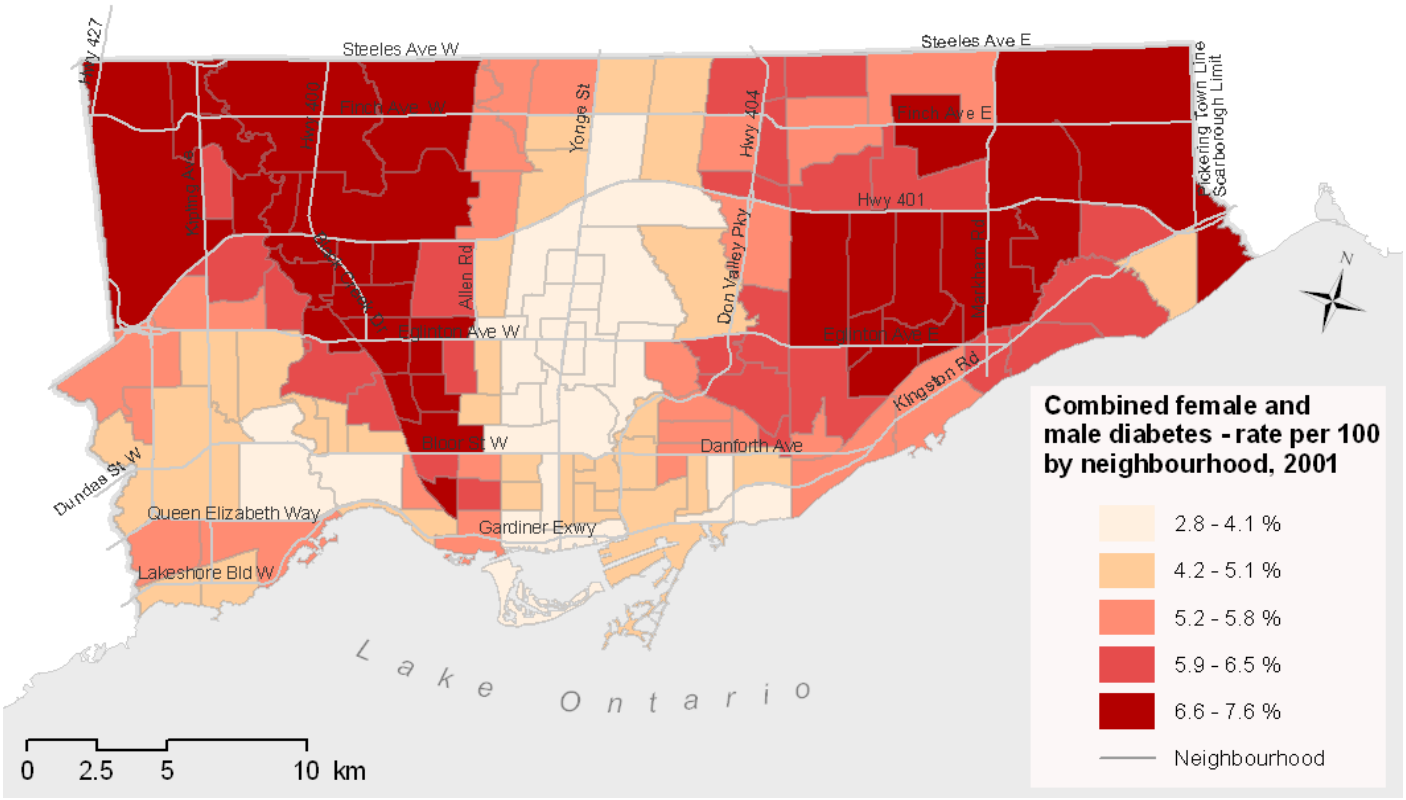
Source: Jha et al, CMAJ submitted

Tobacco deaths (% of total), MEN, Canada, 1986-2006, by income

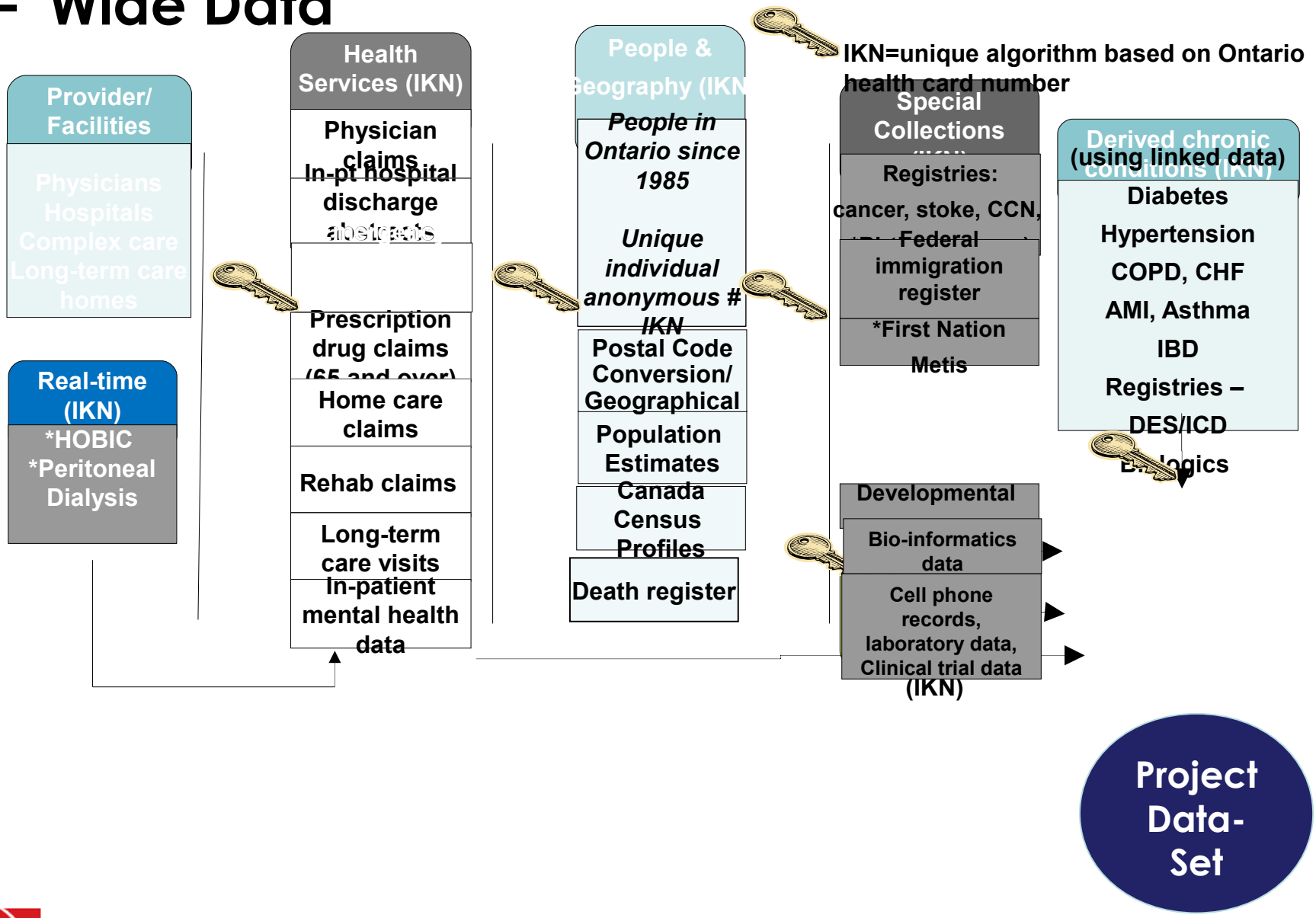
Source: Jha et al, CMAJ submitted

cghr

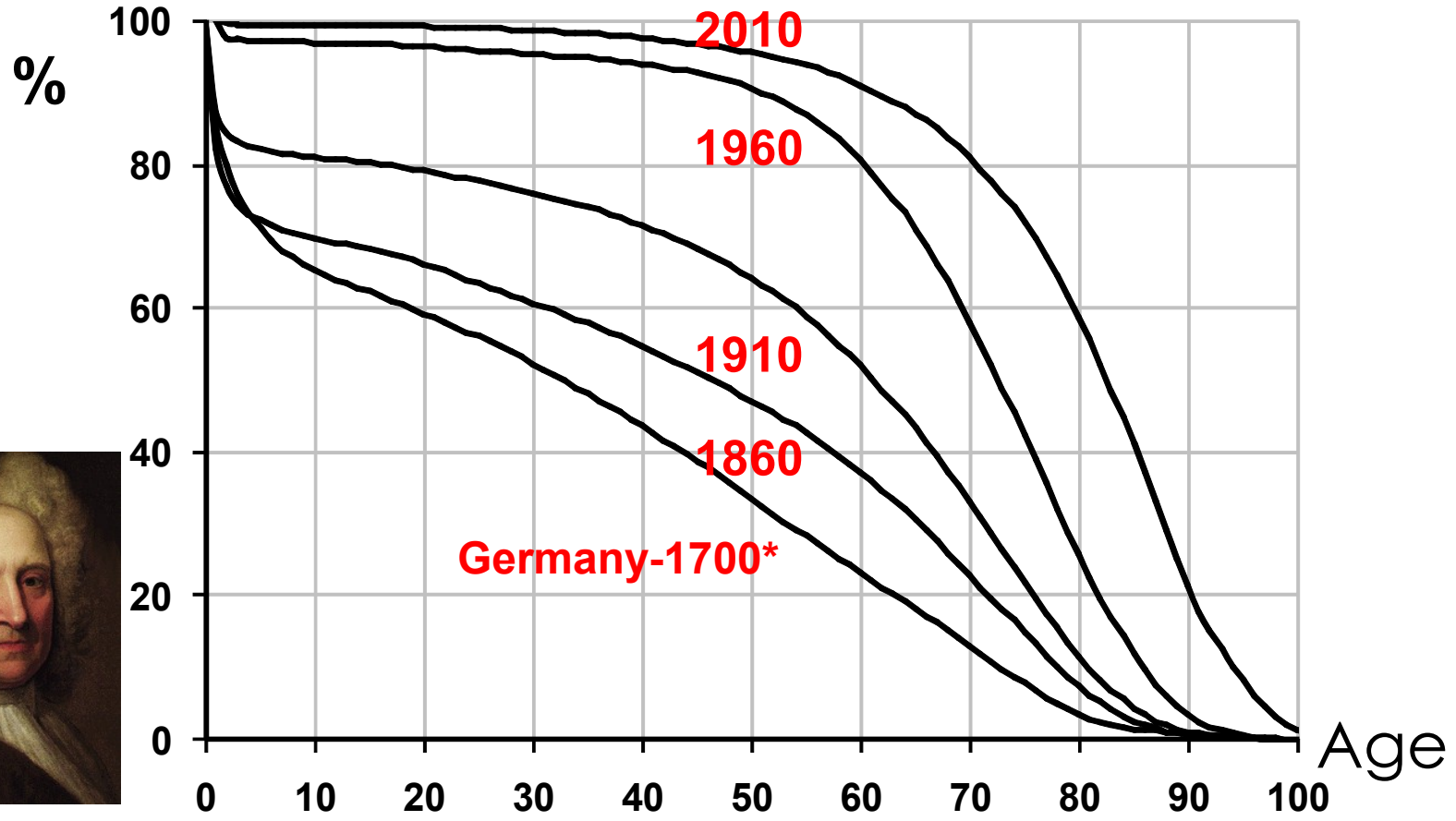
etes is more common in
ds with poor walkability



– 'Wide Data'


Provider/ Facilities
Physicians
Hospitals
Complex care
Long-term care homes
Real-time (IKN)
*HOBIC
*Peritoneal Dialysis
Health Services (IKN)
Physician claims
In-pt hospital discharge abstracts
Prescription drug claims (65 and over)
Home care claims
Rehab claims
Long-term care visits
In-patient mental health data
People & Geography (IKN
People in Ontario since 1985
Unique individual anonymous # IKN
Postal Code Conversion/ Geographical
Population Estimates
Canada Census Profiles
Death register
IKN=unique algorithm based on Ontario health card number
Special Collections
Registries:
cancer, stoke, CCN,
Federal immigration register
*First Nation
Metis
Derived chronic conditions (IKN)
(using linked data)
Diabetes
Hypertension
COPD, CHF
AMI, Asthma
IBD
Registries –
DES/ICD
Developmental
Bio-informatics data
Cell phone records, laboratory data, Clinical trial data
(IKN)
Project Data-Set

Males, England & Wales, % survival at period rates



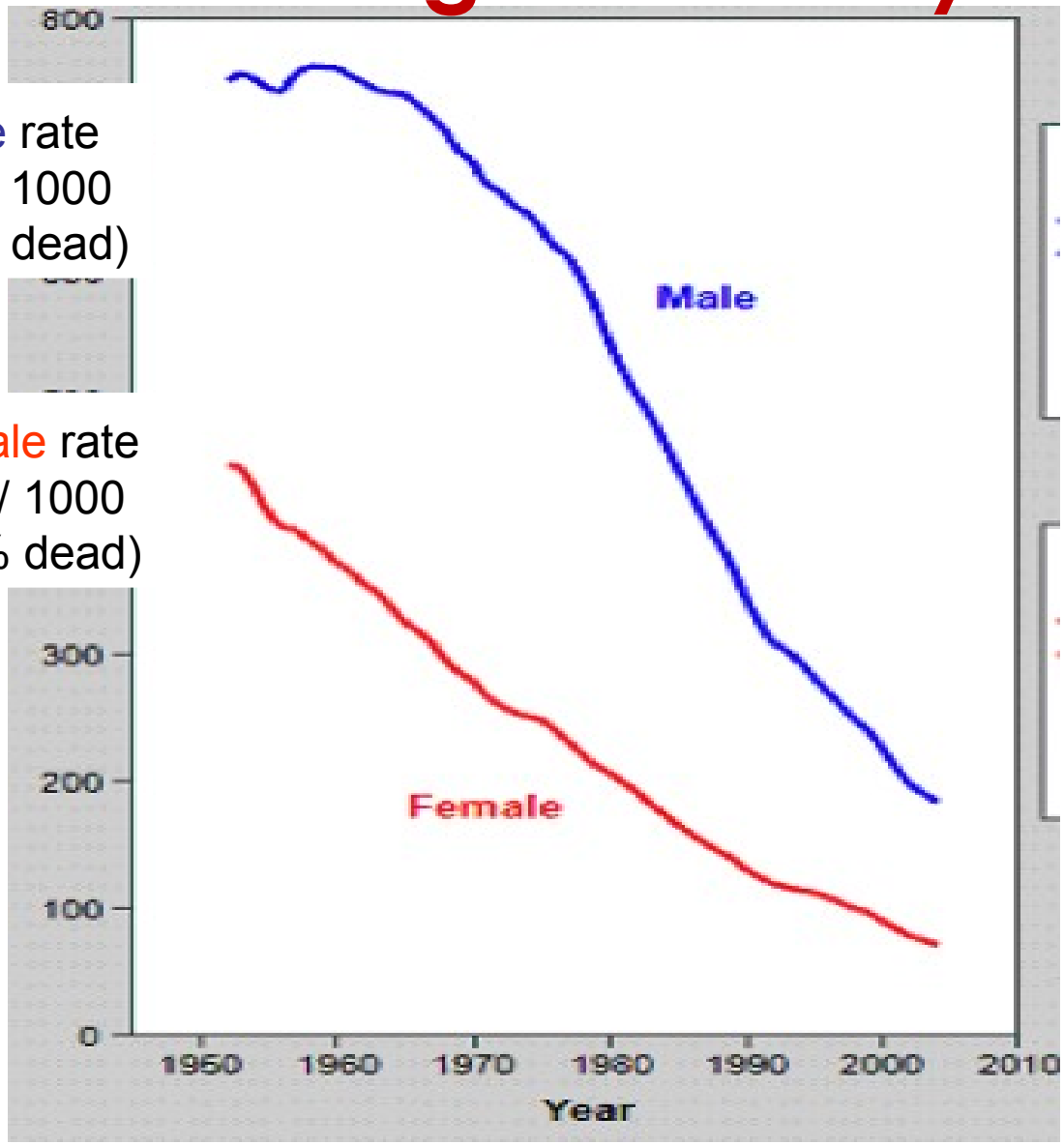
Source: Gary Whitlock, CTSU from Registrar-General reports and Human Mortality Database

* Males and females combined- from Edmond Halley, 1693 for Breslaw, Germany

Canada: vascular death rates in middle age over 50 years

Male rate
7.5 / 1000
(25% dead)

Female rate
4.5 / 1000
(15% dead)



Vascular death at ages 35-69, Canada 2005:
7% Male,
3% Female

cghr

Marginal costs for maximal child survival are falling

GDP per capita ($2005, PPP, per day)

1.25
1
0.75
0.5
0.25
0

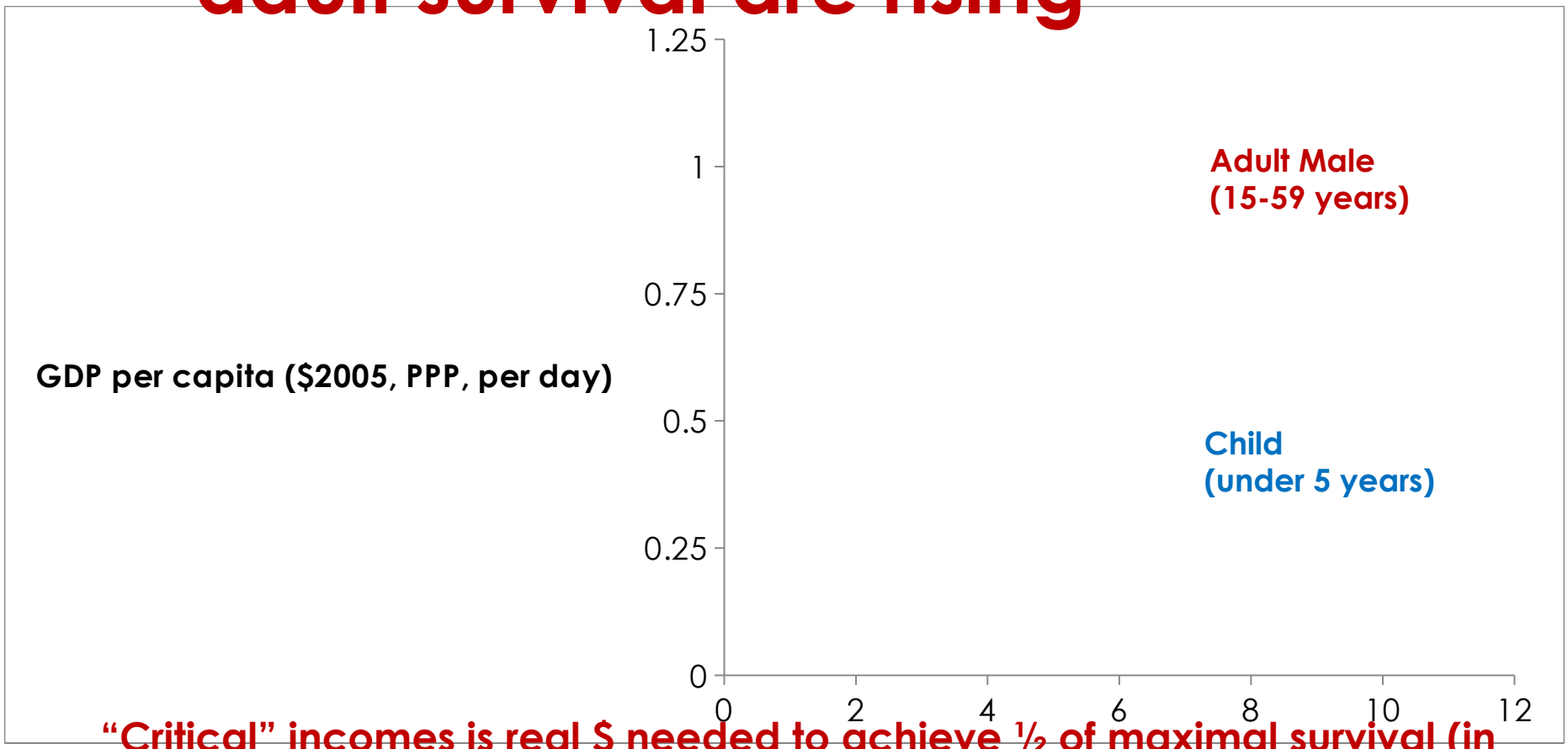
0 2 4 6 8 10 12

Child
(under 5 years)

**"Critical" incomes is real $ needed to achieve ½ of maximal survival (in that year)
from 1970 to 2007**

Source: Hum et al, eLife 2012

Marginal costs for maximal adult survival are rising



"Critical" incomes is real $ needed to achieve ½ of maximal survival (in that year)
from 1970 to 2007; note higher adult costs due in part to HIV and tobacco

Source: Hum et al, eLife 2012

cghr

Linkage of Wide and Deep data

- With large cohorts (100,000s) the administrative data are an efficient way of determining outcomes
- Identifying phenotypes in the administrative data to enable true population-based approaches
- Linking EMR/Admin data to bio-banked samples has great potential
- Overlaying environmental exposures or nutritional maps on geocoded disease distributions
- Great opportunity to study the 'Exposome'

Analytical approaches

- Complex algorithms to analyse genomic/ proteomic/ metabolomic data
- Data Mining: extract information from a data set and transform it into an understandable structure for further use
- Complex algorithms and models: image analysis, facial recognition, weather prediction
- Machine learning
- Measuring associations between exposures and outcomes
  - Propensity scores
  - Inverse probability of treatment weighting
  - Instrumental variable analysis

cghr

Discussions at the U of Toronto

- Big Data Committees in Faculty of Medicine and in the DLSPH
  - Medicine – concentration on deeper analytical capability: genomics proteomics….
  - DLSPH – discussions around linkage of broad data-sets: surveys (OHS, CCHS etc), public health data, social services data, education data etc
  - DLSPH – Focus on exposomics
  - Also the potential for establishment of a Health Observatory in Ontario using linked electronic health and administrative data
  - Support for the data platform from CIHR/SPOR

A few big themes of this meeting

- Substantial capacity in genomics/biological correlates
- Unique representative population (Ontario, n=13M)
- Converging vision- ICES to population, genomics to large datasets
- Need for computing/analytic other PLATFORMS (think LHC plus computing grid)
- GLOBAL relevance- esp with chronic disease/mental health
- Need to train young scientists across disciplines

cghr