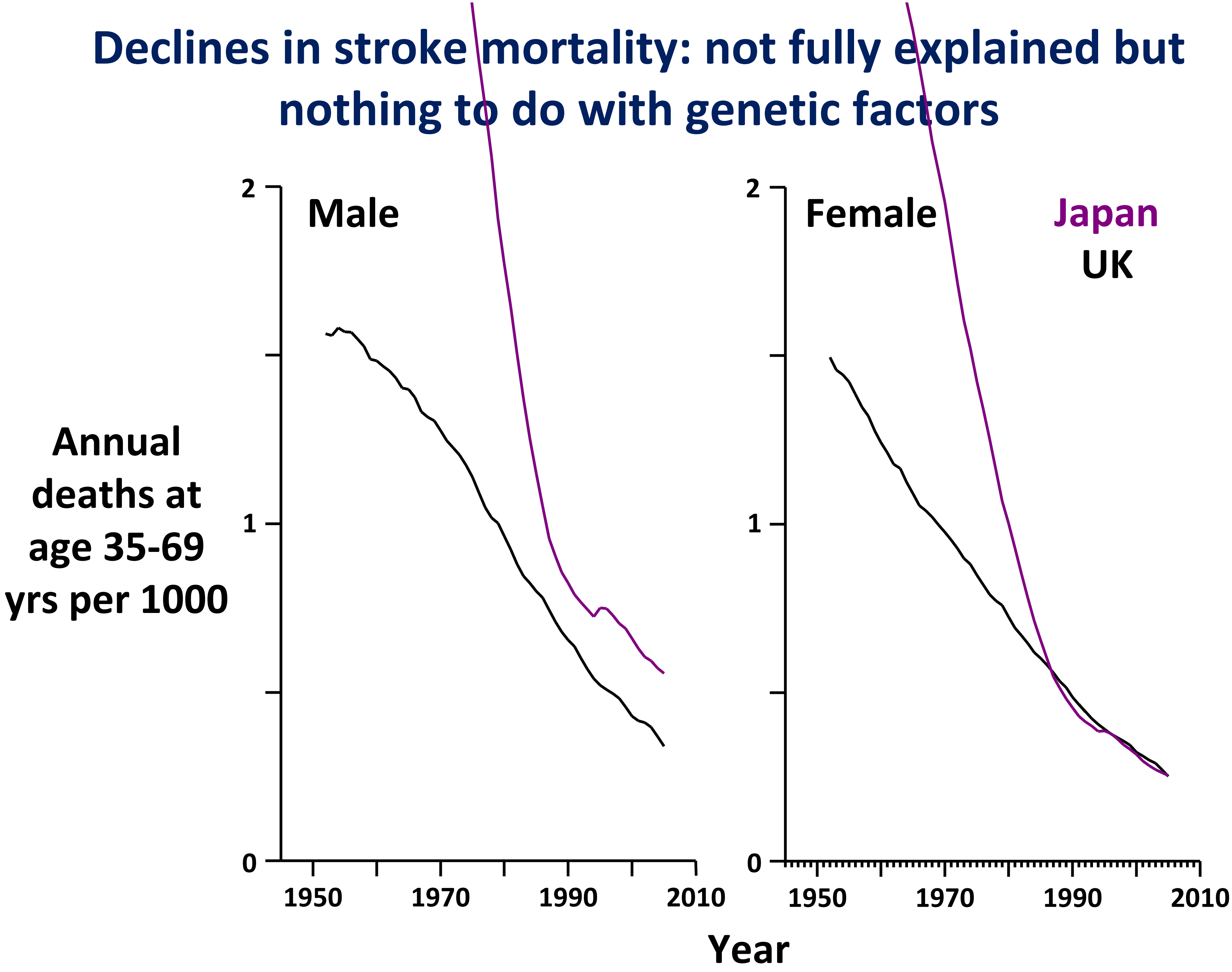
Big Data for Population Health and Personalised Medicine through EMR Linkages

Zheng-Ming CHEN
Professor of Epidemiology
Nuffield Dept. of Population Health,
University of Oxford

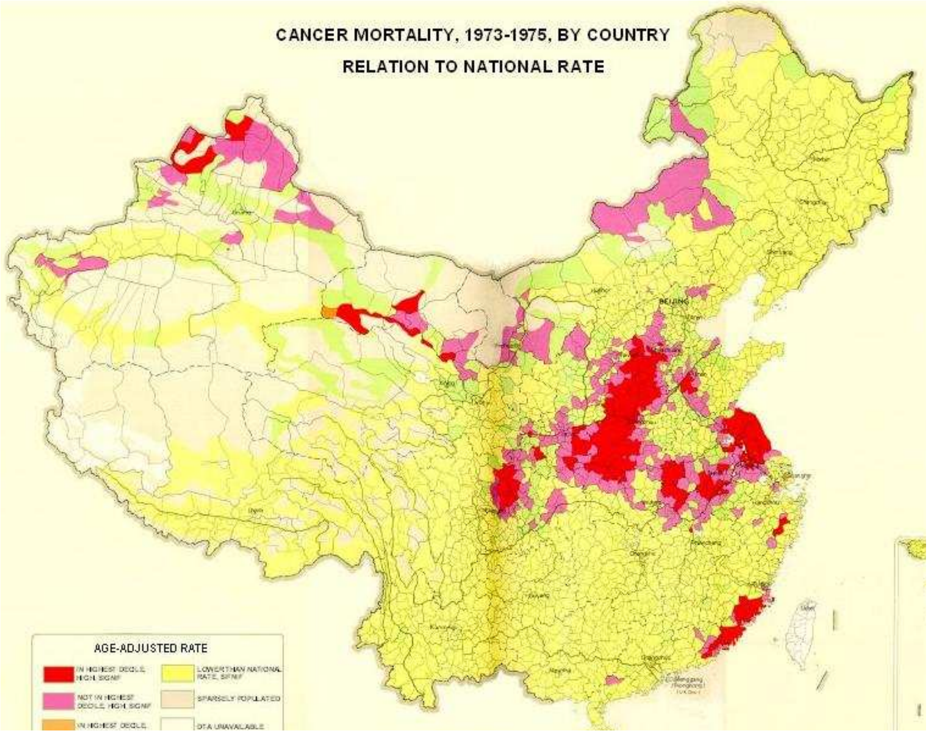
Big Data for Health Policy Workshop, Toronto, Canada
5 November 2014

Declines in stroke mortality: not fully explained but nothing to do with genetic factors

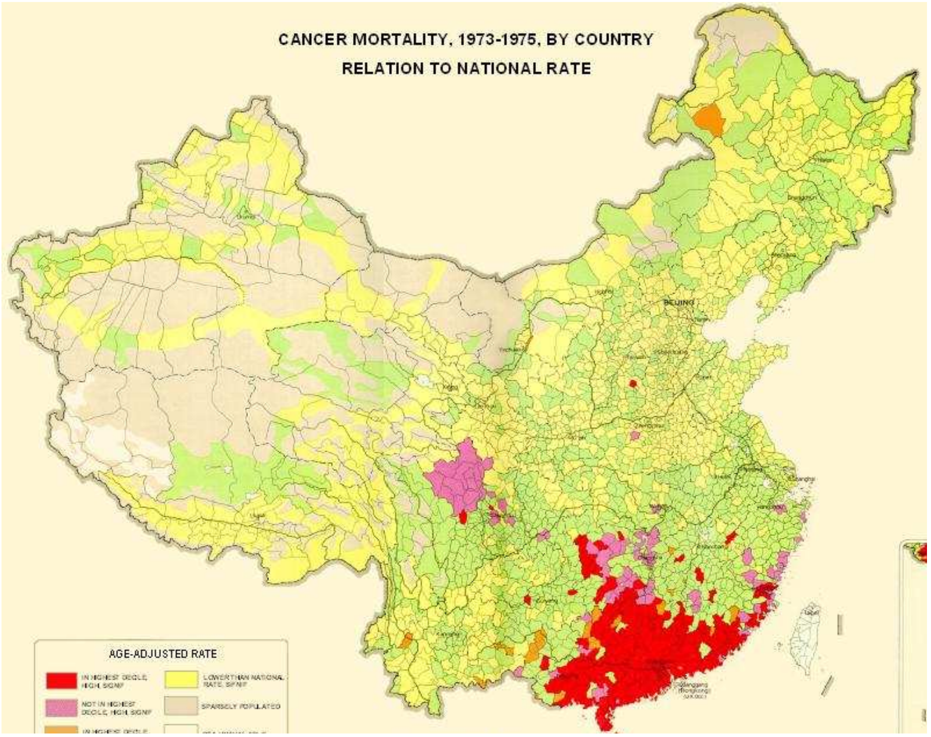


China: large, unexplained mortality variations

Oesophagus cancer

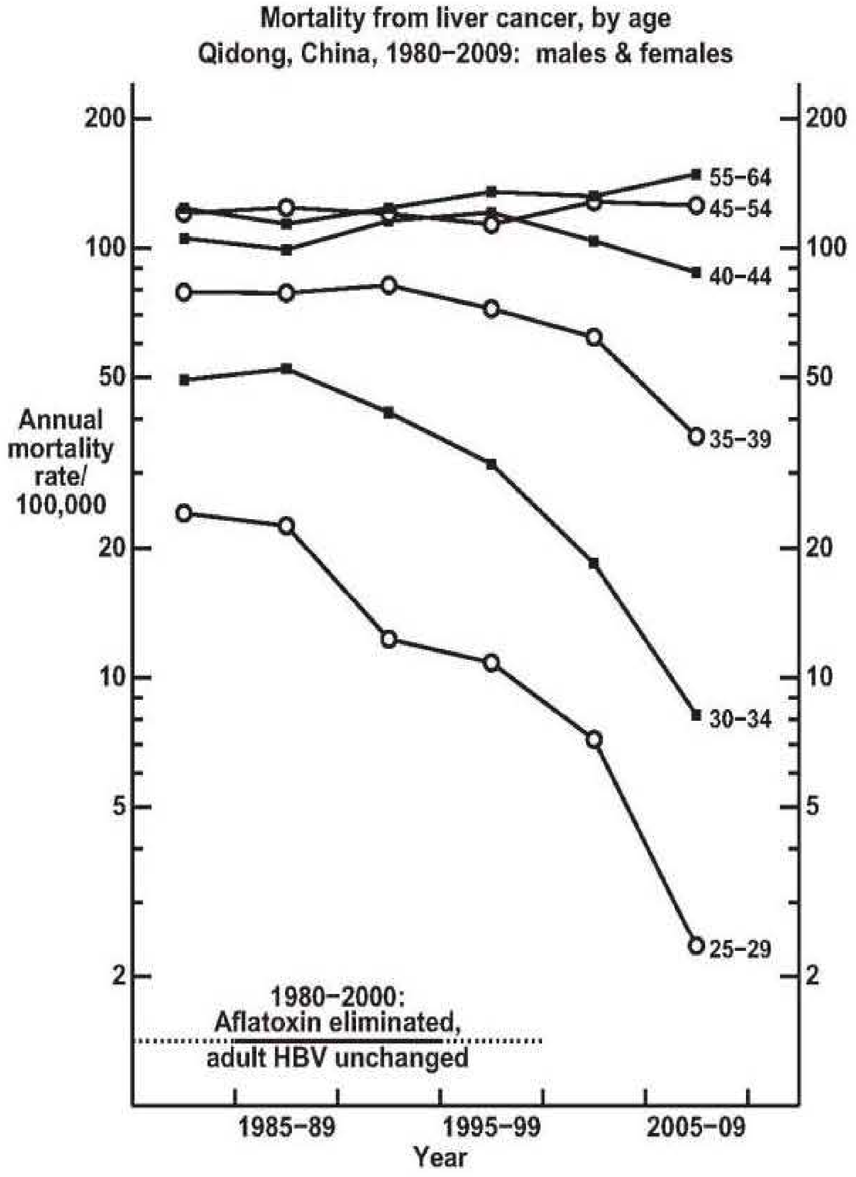

CANCER MORTALITY, 1973-1975, BY COUNTRY
RELATION TO NATIONAL RATE
BEIJING
AGE-ADJUSTED RATE
IN HIGHEST DECILE, HIGH SIGNIF
NOT IN HIGHEST DECILE, HIGH SIGNIF
IN HIGHEST DECILE
LOWER THAN NATIONAL RATE, SIGNIF
SPARSELY POPULATED
DTA UNAVAILABLE

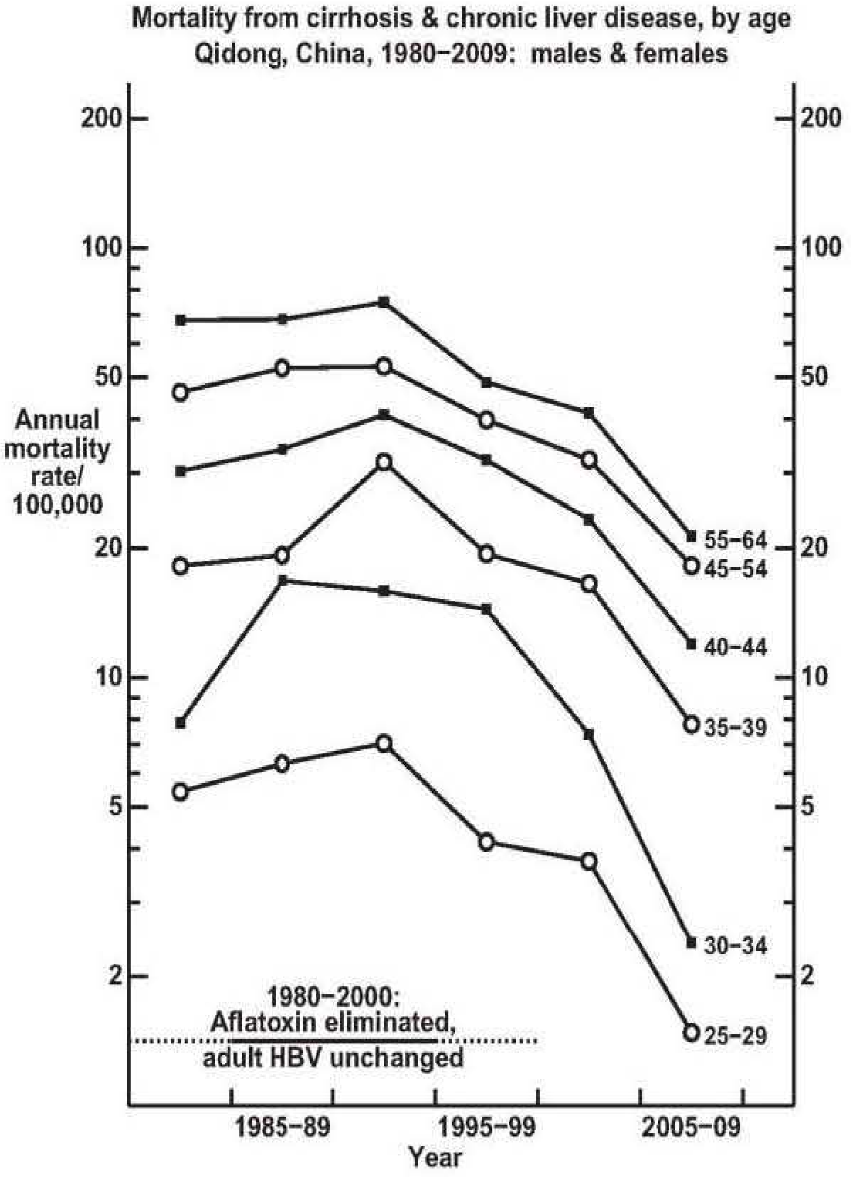
Nasopharynx cancer


CANCER MORTALITY, 1973-1975, BY COUNTRY
RELATION TO NATIONAL RATE
BEIJING
AGE-ADJUSTED RATE
IN HIGHEST DECILE, HIGH SIGNIF
NOT IN HIGHEST DECILE, HIGH SIGNIF
IN HIGHEST DECILE
LOWER THAN NATIONAL RATE, SIGNIF
SPARSELY POPULATED

Females only, hence little effect of tobacco or alcohol
(Red = high mortality is >10x yellow = low mortality)

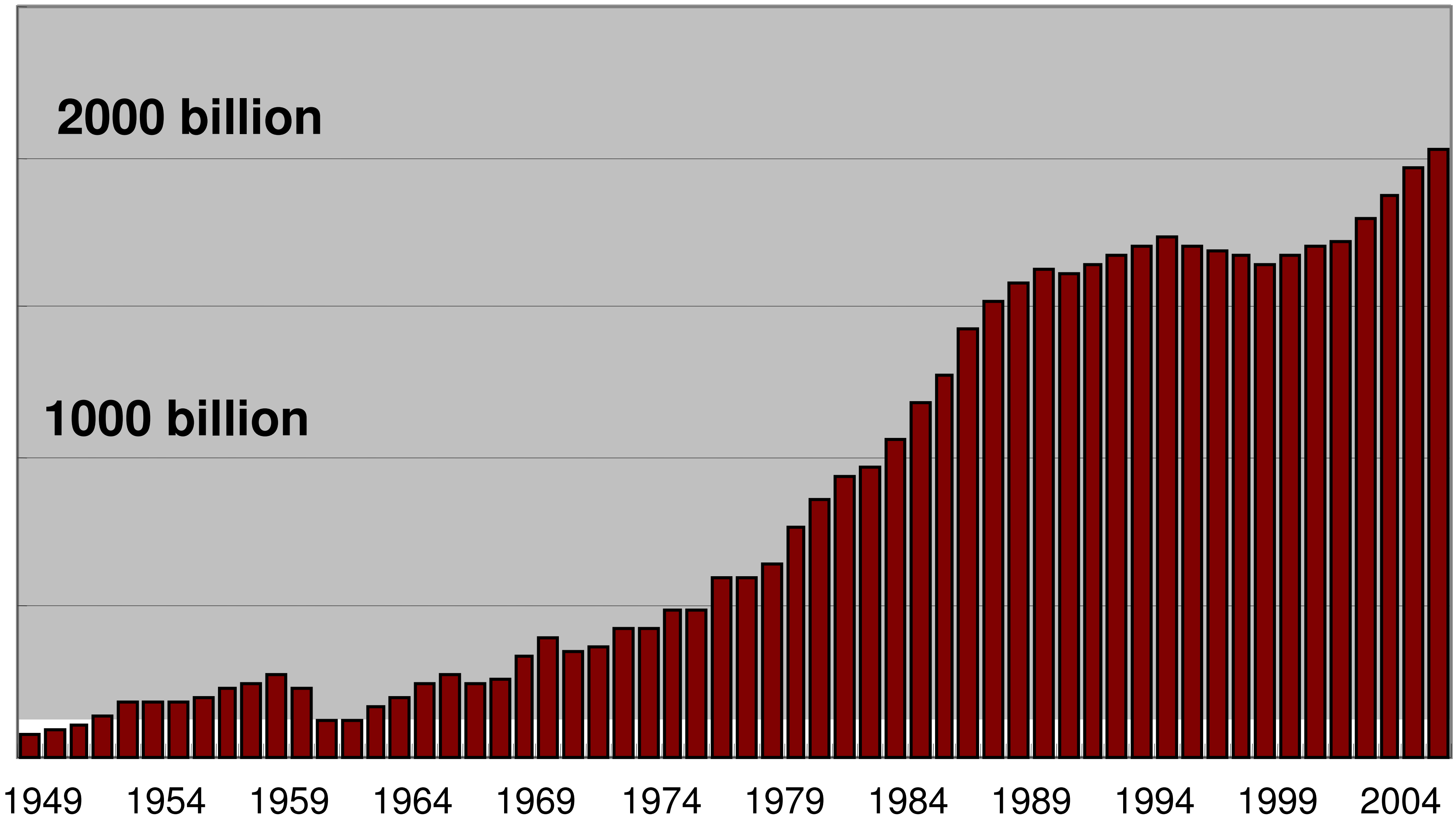
Age-specific trends in adult liver cancer and cirrhosis mortality in Qidong, China, 1980-2009


Mortality from liver cancer, by age
Qidong, China, 1980–2009: males & females

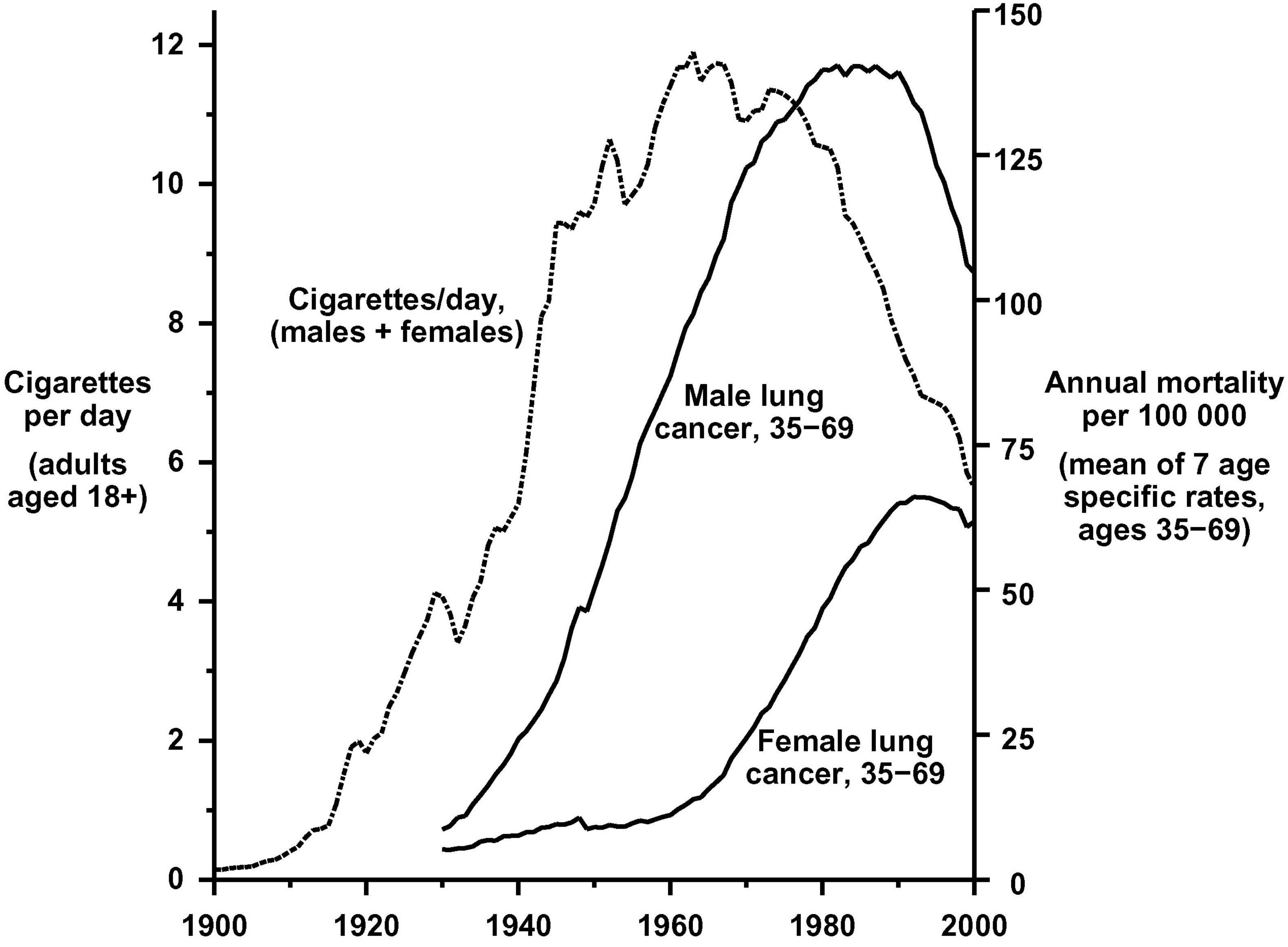

Mortality from cirrhosis & chronic liver disease, by age
Qidong, China, 1980–2009: males & females

Trend of annual cigarette production in China

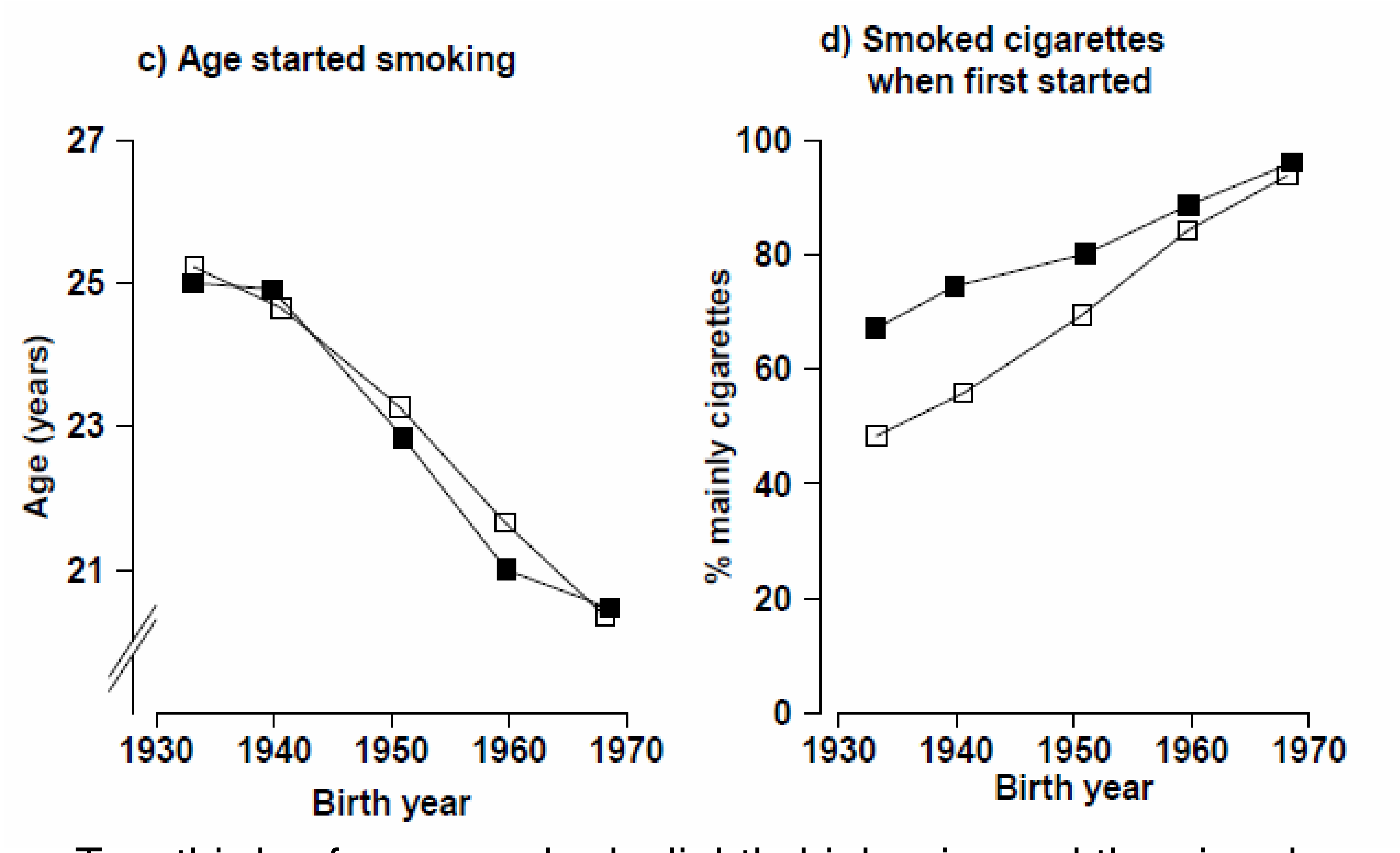
(5% annual increase since 1998)



Cigarette consumption & lung cancer in US



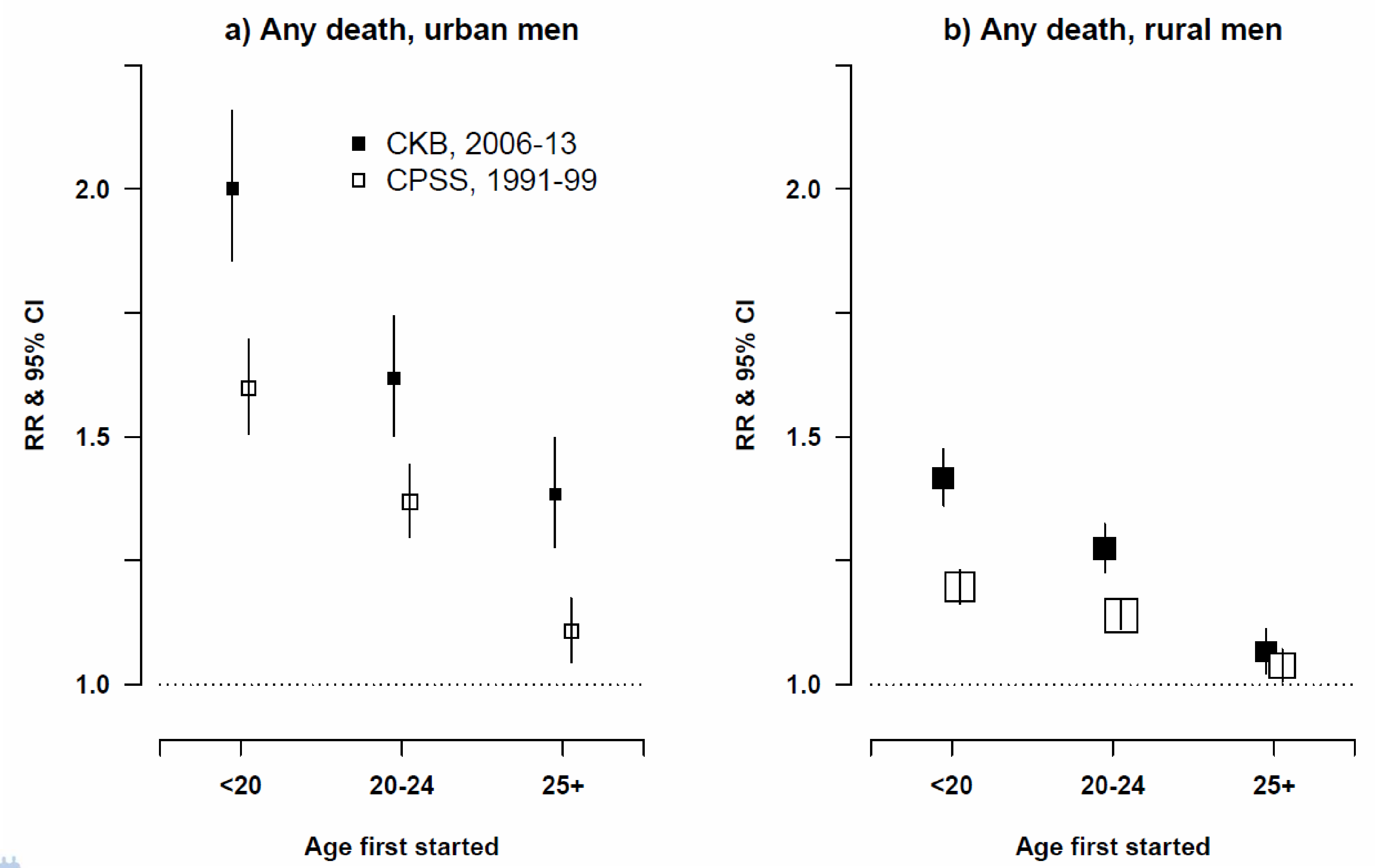
CKB: Smoking patterns by year of birth among men



Two thirds of men smoked, slightly higher in rural than in urban

CKB: Adjusted RR for total mortality by age started

(Tobacco-attributed death: 25% urban, 15% rural)



To be submitted to the Lancet, 2014

China Kadoorie Biobank: design

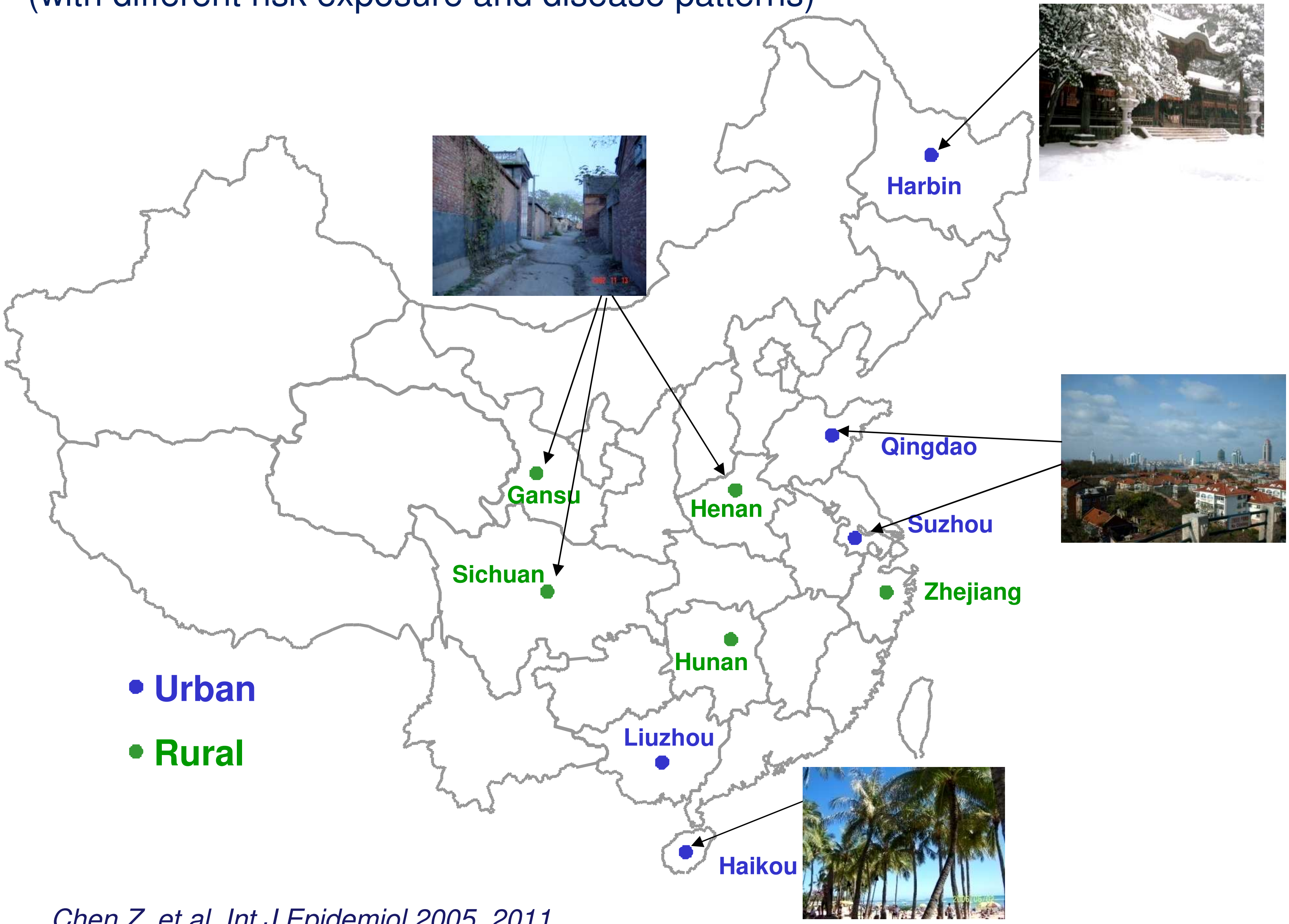
(genetic & other causes of common disease)

- 500K recruited from 10 localities in 2004-08
- Participants interviewed, measured, and gave 10 mL blood for long-term storage
- Periodic resurvey of 5% (for regression dilution)
- All followed up indefinitely via electronic record linkage to deaths and ALL hospital episodes

General consent for access to health record for unspecified medical research

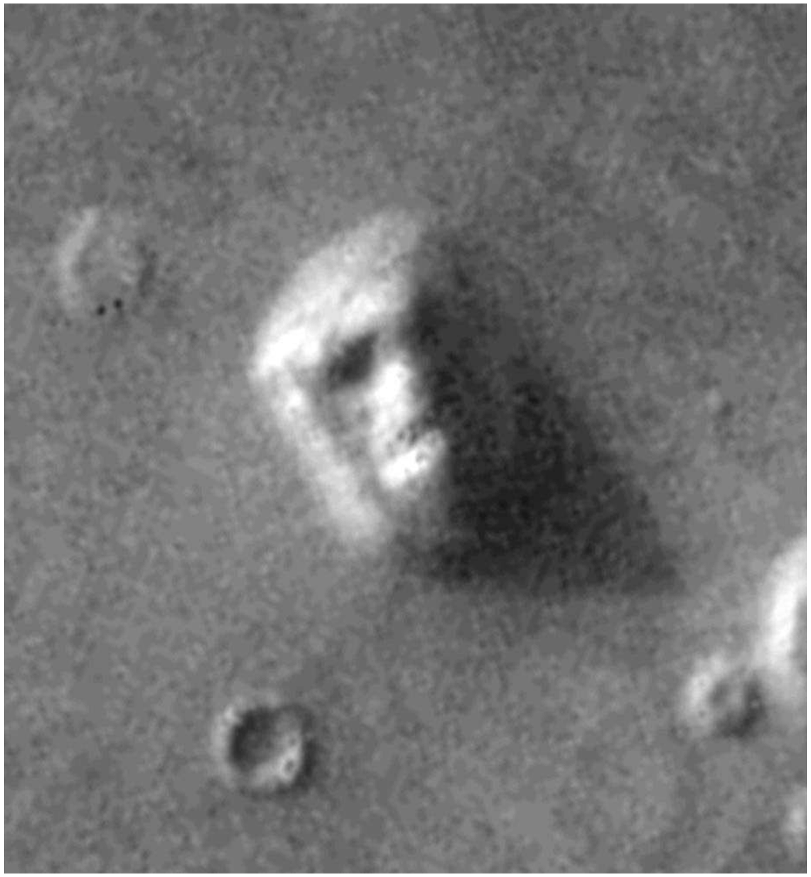
CKB: Location of the 10 survey sites in China

(with different risk exposure and disease patterns)

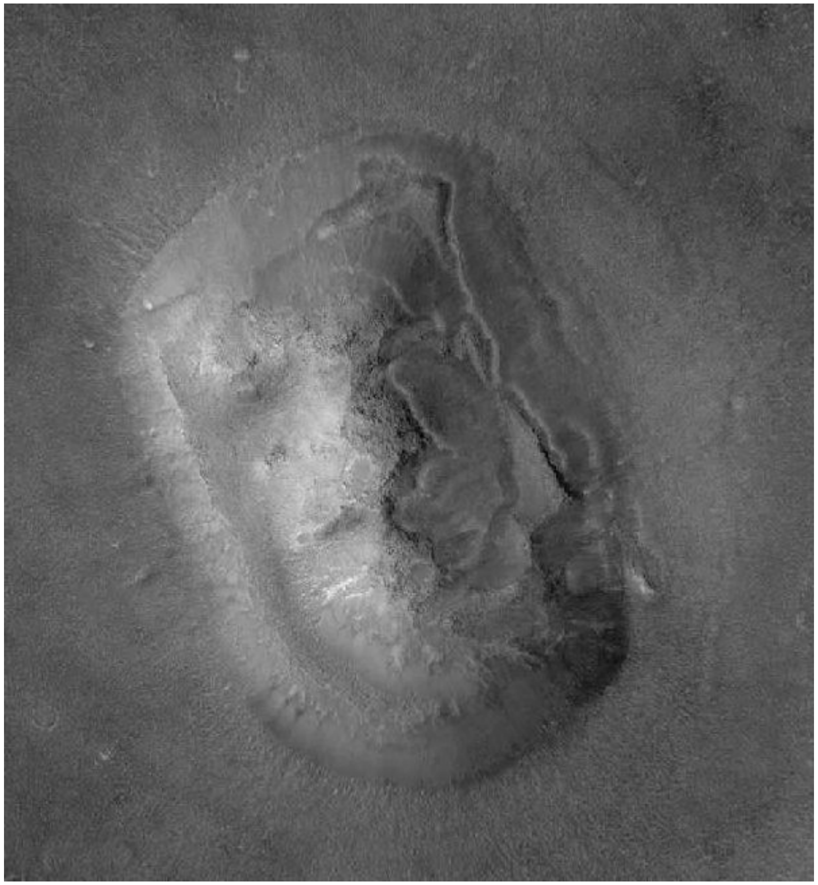


Chen Z, et al. Int J Epidemiol 2005, 2011

A human face on Mars?



1976: Viking Orbiter

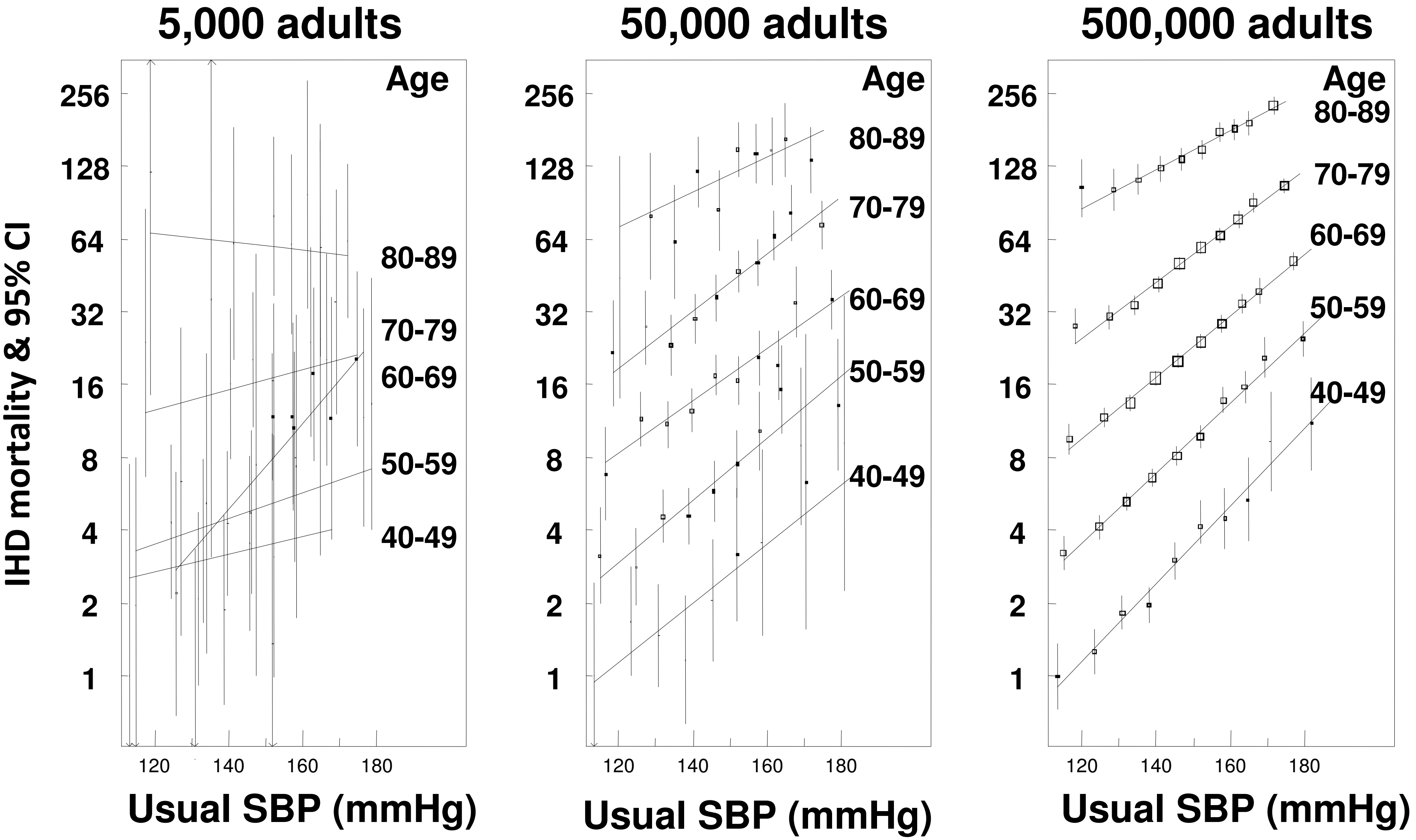


2001: Mars Orbiter

More observations allow a clearer, more precise, and more detailed picture of reality – also makes it less likely that we see patterns when none exist

SIZE matters: SBP vs IHD mortality, by age

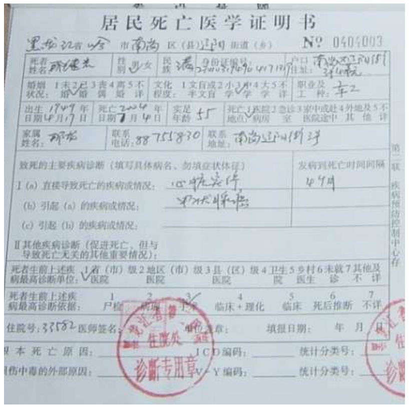
5K, 50K & 500K randomly chosen from PSC*



** Prospective Studies Collaboration, Lancet 2002*

CKB: Main data sources for linkage

Death registries



居民死亡医学证明书

黑龙江省 哈 市 南岗 区（县）辽河 街道（乡）　№ 0404003

死者姓名 [illegible]	性别 男／女	民族 汉	身份证编号 [illegible]	户口住址 南岗辽河街[illegible]
婚姻状况：1未婚 [✓ 2已婚] 3丧偶 4离婚 5不详	文化程度：1文盲或半文盲 2小学 [✓ 3中学] 4大学 5不详	职业及工种：[illegible]	
出生日期 1949 年 4 月 17 日	死亡日期 2004 年 7 月 4 日	实足年龄 55	死亡地点：[✓ 1医院病房] 2急诊室 3家中或赴医院途中 4外地及其他 5不详
家属姓名 邢[illegible]	联系电话 88751830	联系地址 南岗辽河街[illegible]	

致死的主要疾病诊断（填写具体病名、勿填症状体征）　发病到死亡时间间隔

I (a) 直接导致死亡的疾病或情况：[illegible]　4年

(b) 引起 (a) 的疾病或情况：[illegible]

(c) 引起 (b) 的疾病或情况：

II 其他疾病诊断（促进死亡，但与导致死亡无关的其他重要情况）：

死者生前上述疾病最高诊断单位：[✓ 1省（市）级医院] 2地区（市）级医院 3县（区）级医院 4卫生院 5乡村医生 6未就诊 7其他及不详

死者生前上述疾病最高诊断依据：1 尸检 2 病理 [✓ 3 手术] 4 临床＋理化 5 临床 6 死后推断 7 不详

住院号：33582 医师签名：　单位盖章：　填报日期：　年　月　日

根本死亡原因：＿＿＿ ICD编码：＿＿＿ 统计分类号：＿＿＿

损伤中毒的外部原因：＿＿＿ V－Y编码：＿＿＿ 统计分类号：＿＿＿

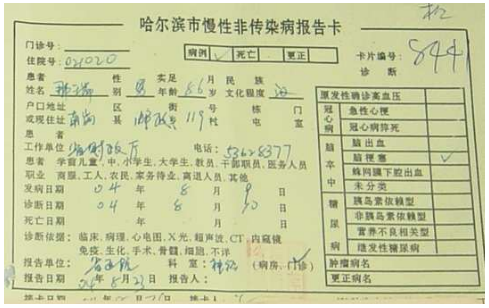
第二联 疾病预防控制中心存

黑龙江省[illegible] 住院处 诊断专用章

Outcome Follow up in CKB

Active follow-up

Disease registries



哈尔滨市慢性非传染病报告卡

门诊号：

住院号：021020

病例 ✓ 死亡 　更正 　卡片编号：844

患者姓名 [illegible] 性别 男 实足年龄 86 岁 民族 　文化程度 [illegible]

户口地址或现住址 南岗 区县 [illegible] 街村 119 号 栋屯 门室

患者工作单位 省财政厅 电话：53628377

患者职业 学前儿童、中、小学生、大学生、教员、干部职员、医务人员、商服、工人、农民、家务待业、离退人员、其他

发病日期 04 年 8 月 9 日

诊断日期 04 年 8 月 10 日

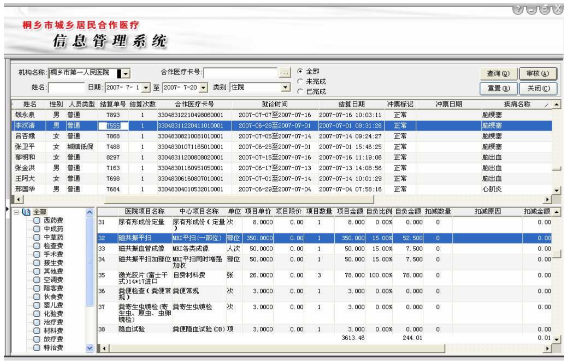
死亡日期 　年 　月 　日

诊断依据：临床、病理、心电图、X光、超声波、CT、内窥镜、免疫、生化、手术、骨髓、细胞、不详

报告单位：省医院 科室：神经（病房、[✓ 门诊]）

报告日期 04 年 8 月 23 日 报告人：

诊断		
原发性确诊高血压		
冠心病	急性心梗	
	冠心病猝死	
脑卒中	脑出血	
	脑梗塞	✓
	蛛网膜下腔出血	
	未分类	
糖尿病	胰岛素依赖型	
	非胰岛素依赖型	
	营养不良相关型	
	继发性糖尿病	
肿瘤病名		
更正病名		



桐乡市城乡居民合作医疗

信息管理系统

机构名称：桐乡市第一人民医院　合作医疗卡号：　● 全部 ○ 未完成 ○ 已完成

姓名：　日期：2007- 7- 1 至 2007- 7-20　类别：住院

姓名	性别	人员类型	结算单号	结算次数	合作医疗卡号	就诊时间	结算日期	冲票标记	冲票日期	疾病名称
钱永泉	男	普通	7893	1	330483122104980600001	2007-07-07至2007-07-16	2007-07-16 10:03:11	正常		脑梗塞
李汉清	男	普通	7555	1	330483112204110100001	2007-06-28至2007-07-01	2007-07-01 09:31:26	正常		脑梗塞
吕杏娥	女	普通	7868	1	330483082100810100001	2007-07-05至2007-07-14	2007-07-14 09:24:27	正常		脑梗塞
张卫平	女	城镇低保	7488	1	330483010711650100001	2007-06-25至2007-07-01	2007-07-01 15:46:25	正常		脑梗塞
郁明和	女	普通	8297	1	330483112008080200001	2007-07-15至2007-07-16	2007-07-16 11:19:06	正常		脑出血
张金洪	男	普通	7163	1	330483011609510500001	2007-06-17至2007-07-13	2007-07-13 14:08:56	正常		脑出血
王阿大	女	普通	7698	1	330483061608070100001	2007-07-01至2007-07-14	2007-07-14 10:01:29	正常		脑出血
邢国华	男	普通	7684	1	330483040105320100001	2007-06-29至2007-07-04	2007-07-04 07:58:16	正常		心肌炎

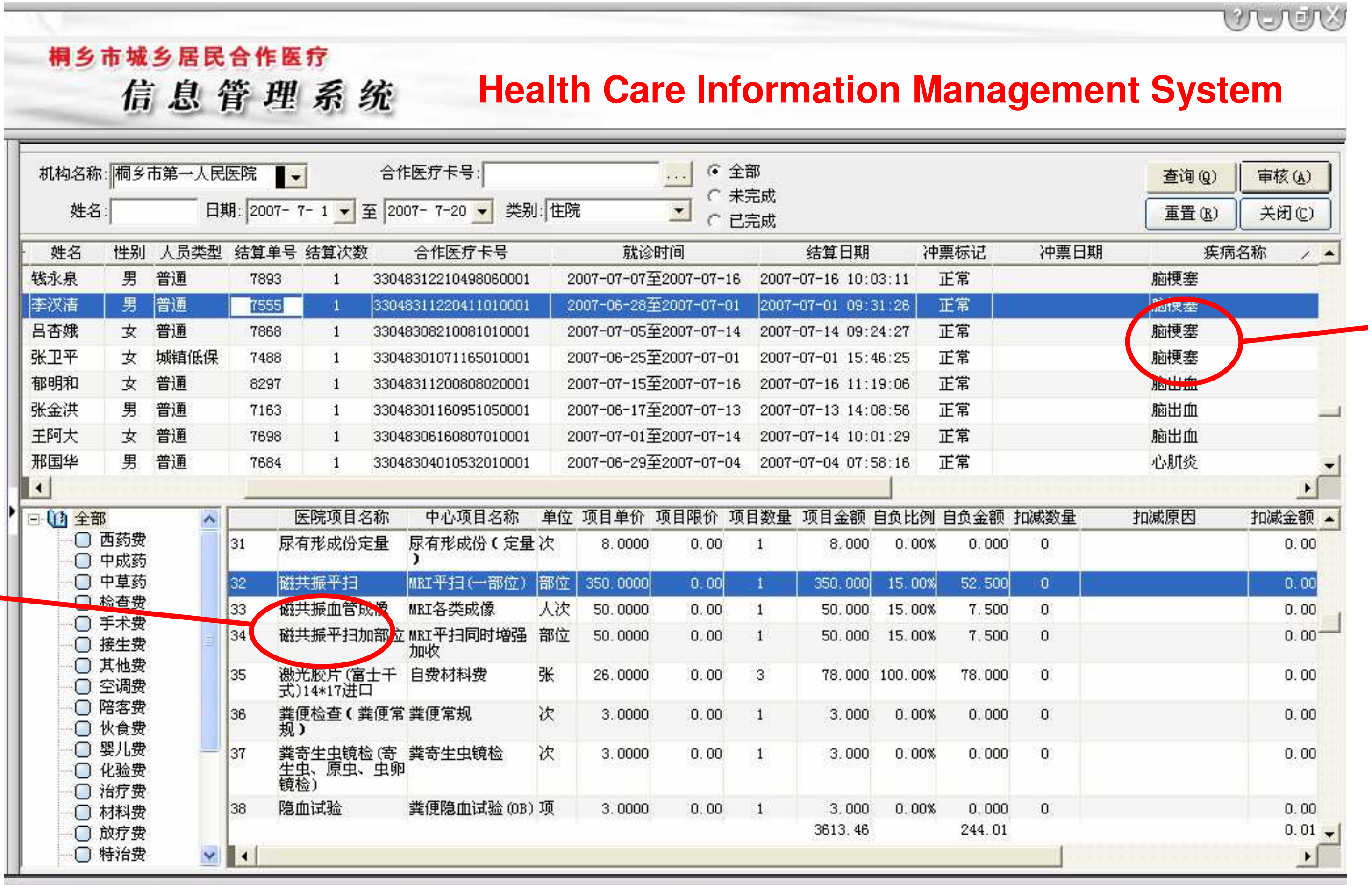
	医院项目名称	中心项目名称	单位	项目单价	项目限价	项目数量	项目金额	自负比例	自负金额	扣减数量	扣减原因	扣减金额
31	尿有形成份定量	尿有形成份（定量）	次	8.0000	0.00	1	8.000	0.00%	0.000	0		0.00
32	磁共振平扫	MRI平扫（一部位）	部位	350.0000	0.00	1	350.000	15.00%	52.500	0		0.00
33	磁共振血管成像	MRI各类成像	人次	50.0000	0.00	1	50.000	15.00%	7.500	0		0.00
34	磁共振平扫加部位	MRI平扫同时增强加收	部位	50.0000	0.00	1	50.000	15.00%	7.500	0		0.00
35	激光胶片（富士干式）14*17进口	自费材料费	张	26.0000	0.00	3	78.000	100.00%	78.000	0		0.00
36	粪便检查（粪便常规）	粪便常规	次	3.0000	0.00	1	3.000	0.00%	0.000	0		0.00
37	粪寄生虫镜检（寄生虫、原虫、虫卵镜检）	粪寄生虫镜检	次	3.0000	0.00	1	3.000	0.00%	0.000	0		0.00
38	隐血试验	粪便隐血试验（OB）	项	3.0000	0.00	1	3.000	0.00%	0.000	0		0.00
							3613.46		244.01			0.01

Health insurance (national)

UNIVERSITY OF OXFORD

CHINA KADOORIE BIOBANK 中国慢性病前瞻性研究

National Health Insurance system in China



桐乡市城乡居民合作医疗
信息管理系统

Health Care Information Management System

机构名称: 桐乡市第一人民医院
合作医疗卡号:
● 全部
○ 未完成
○ 已完成
姓名:
日期: 2007- 7- 1 至 2007- 7-20
类别: 住院

姓名	性别	人员类型	结算单号	结算次数	合作医疗卡号	就诊时间	结算日期	冲票标记	冲票日期	疾病名称
钱永泉	男	普通	7893	1	33048312210498060001	2007-07-07至2007-07-16	2007-07-16 10:03:11	正常		脑梗塞
李汉清	男	普通	7555	1	33048311220411010001	2007-06-28至2007-07-01	2007-07-01 09:31:26	正常		脑梗塞
吕杏娥	女	普通	7868	1	33048308210081010001	2007-07-05至2007-07-14	2007-07-14 09:24:27	正常		脑梗塞
张卫平	女	城镇低保	7488	1	33048301071165010001	2007-06-25至2007-07-01	2007-07-01 15:46:25	正常		脑梗塞
郁明和	女	普通	8297	1	33048311200808020001	2007-07-15至2007-07-16	2007-07-16 11:19:06	正常		脑出血
张金洪	男	普通	7163	1	33048301160951050001	2007-06-17至2007-07-13	2007-07-13 14:08:56	正常		脑出血
王阿大	女	普通	7698	1	33048306160807010001	2007-07-01至2007-07-14	2007-07-14 10:01:29	正常		脑出血
邢国华	男	普通	7684	1	33048304010532010001	2007-06-29至2007-07-04	2007-07-04 07:58:16	正常		心肌炎

全部
- 西药费
- 中成药
- 中草药
- 检查费
- 手术费
- 接生费
- 其他费
- 空调费
- 陪客费
- 伙食费
- 婴儿费
- 化验费
- 治疗费
- 材料费
- 放疗费
- 特治费

	医院项目名称	中心项目名称	单位	项目单价	项目限价	项目数量	项目金额	自负比例	自负金额	扣减数量	扣减原因	扣减金额
31	尿有形成份定量	尿有形成份（定量）	次	8.0000	0.00	1	8.000	0.00%	0.000	0		0.00
32	磁共振平扫	MRI平扫(一部位)	部位	350.0000	0.00	1	350.000	15.00%	52.500	0		0.00
33	磁共振血管成像	MRI各类成像	人次	50.0000	0.00	1	50.000	15.00%	7.500	0		0.00
34	磁共振平扫加部位	MRI平扫同时增强加收	部位	50.0000	0.00	1	50.000	15.00%	7.500	0		0.00
35	激光胶片(富士干式)14*17进口	自费材料费	张	26.0000	0.00	3	78.000	100.00%	78.000	0		0.00
36	粪便检查（粪便常规）	粪便常规	次	3.0000	0.00	1	3.000	0.00%	0.000	0		0.00
37	粪寄生虫镜检(寄生虫、原虫、虫卵镜检)	粪寄生虫镜检	次	3.0000	0.00	1	3.000	0.00%	0.000	0		0.00
38	隐血试验	粪便隐血试验(OB)	项	3.0000	0.00	1	3.000	0.00%	0.000	0		0.00
							3613.46		244.01			0.01

Ischaemic stroke

Confirmed by MRI

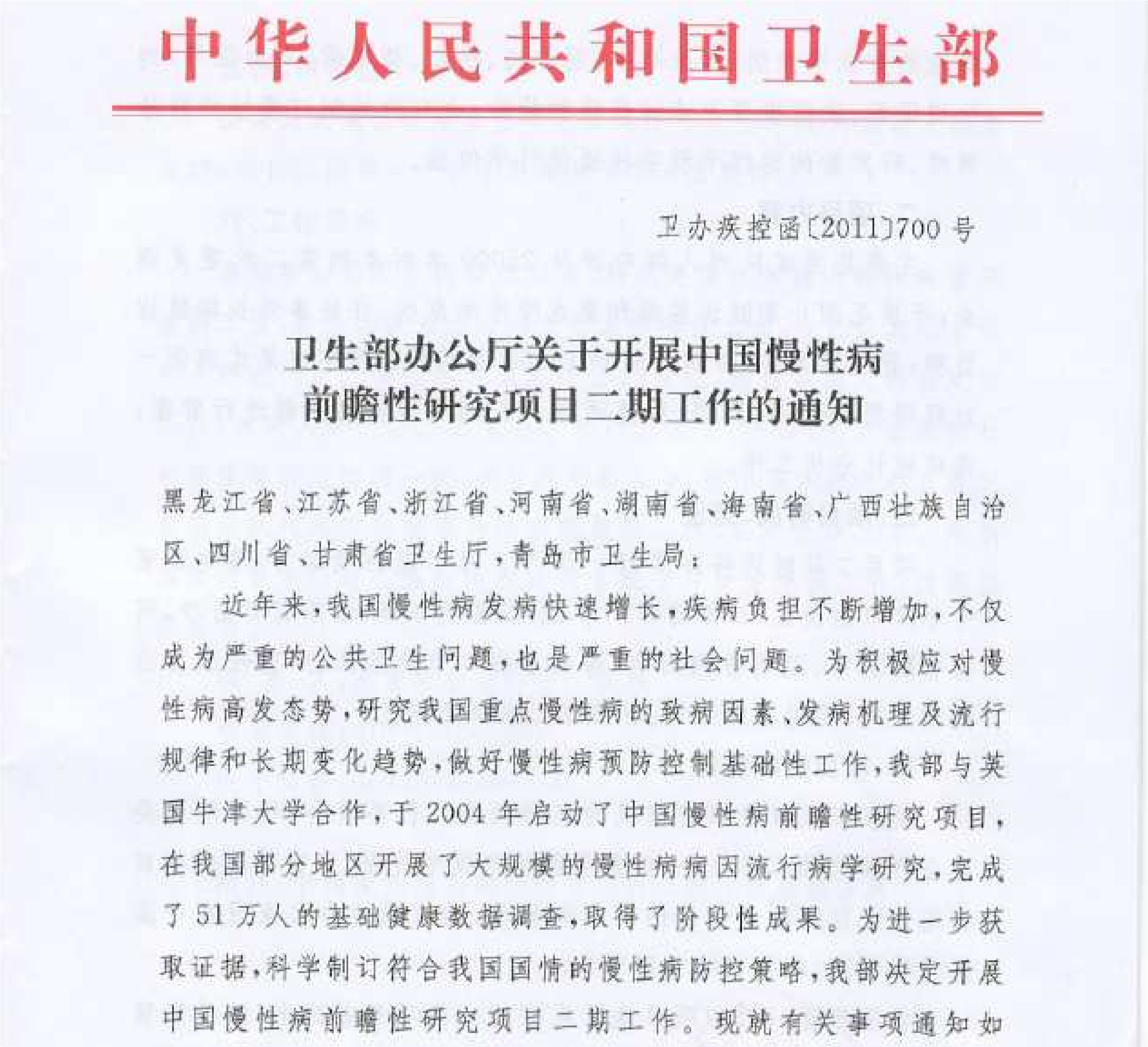
By 1.1.2014, >98% of participants had been linked to the HI databases through unique national ID number

National health insurance system in China

- Introduced during 2004-6 with almost universal coverage by 2010
- Diagnosis ICD-10 coded, plus disease descriptions and >2,000 procedure codes
- Managed electronically at city or county levels, mainly for financial purposes (& itemised cost)

In CKB ~1.6M episodes, ~20M procedures/tests, ~3500 diseases had been recorded during 2006-14

Strong political support within China



中华人民共和国卫生部

卫办疾控函〔2011〕700号

卫生部办公厅关于开展中国慢性病前瞻性研究项目二期工作的通知

黑龙江省、江苏省、浙江省、河南省、湖南省、海南省、广西壮族自治区、四川省、甘肃省卫生厅，青岛市卫生局：

近年来，我国慢性病发病快速增长，疾病负担不断增加，不仅成为严重的公共卫生问题，也是严重的社会问题。为积极应对慢性病高发态势，研究我国重点慢性病的致病因素、发病机理及流行规律和长期变化趋势，做好慢性病预防控制基础性工作，我部与英国牛津大学合作，于2004年启动了中国慢性病前瞻性研究项目，在我国部分地区开展了大规模的慢性病病因流行病学研究，完成了51万人的基础健康数据调查，取得了阶段性成果。为进一步获取证据，科学制订符合我国国情的慢性病防控策略，我部决定开展中国慢性病前瞻性研究项目二期工作。现就有关事项通知如

CKB: **examples of new research using EMR**

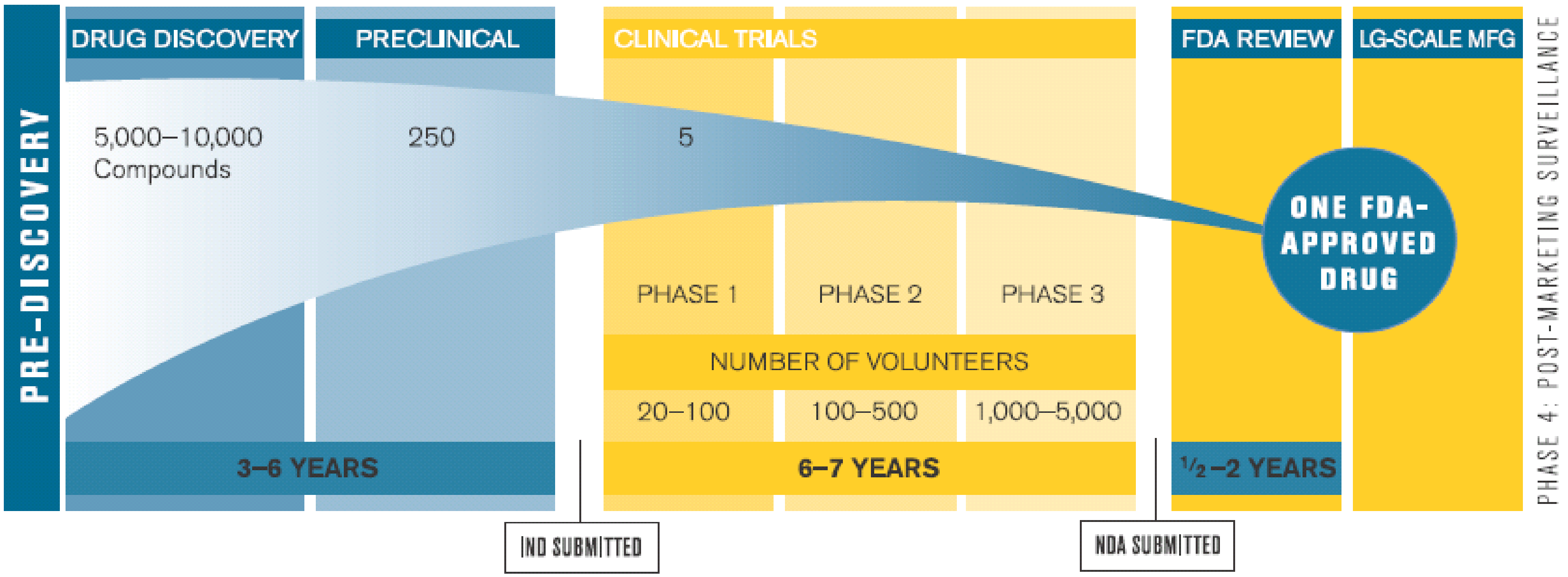
- Infective causes of cancer (*WHO IARC, France)*
- Genetics to aid drug development (*GSK, Merck)*
- Multi-omics biomarker discovery *(Oulu,SomaLogic)*
- Effects of air pollution (*Fudan University, China)*
- Healthcare delivery in China *(Oxford & Fudan)*

Plus conventional epidemiological research

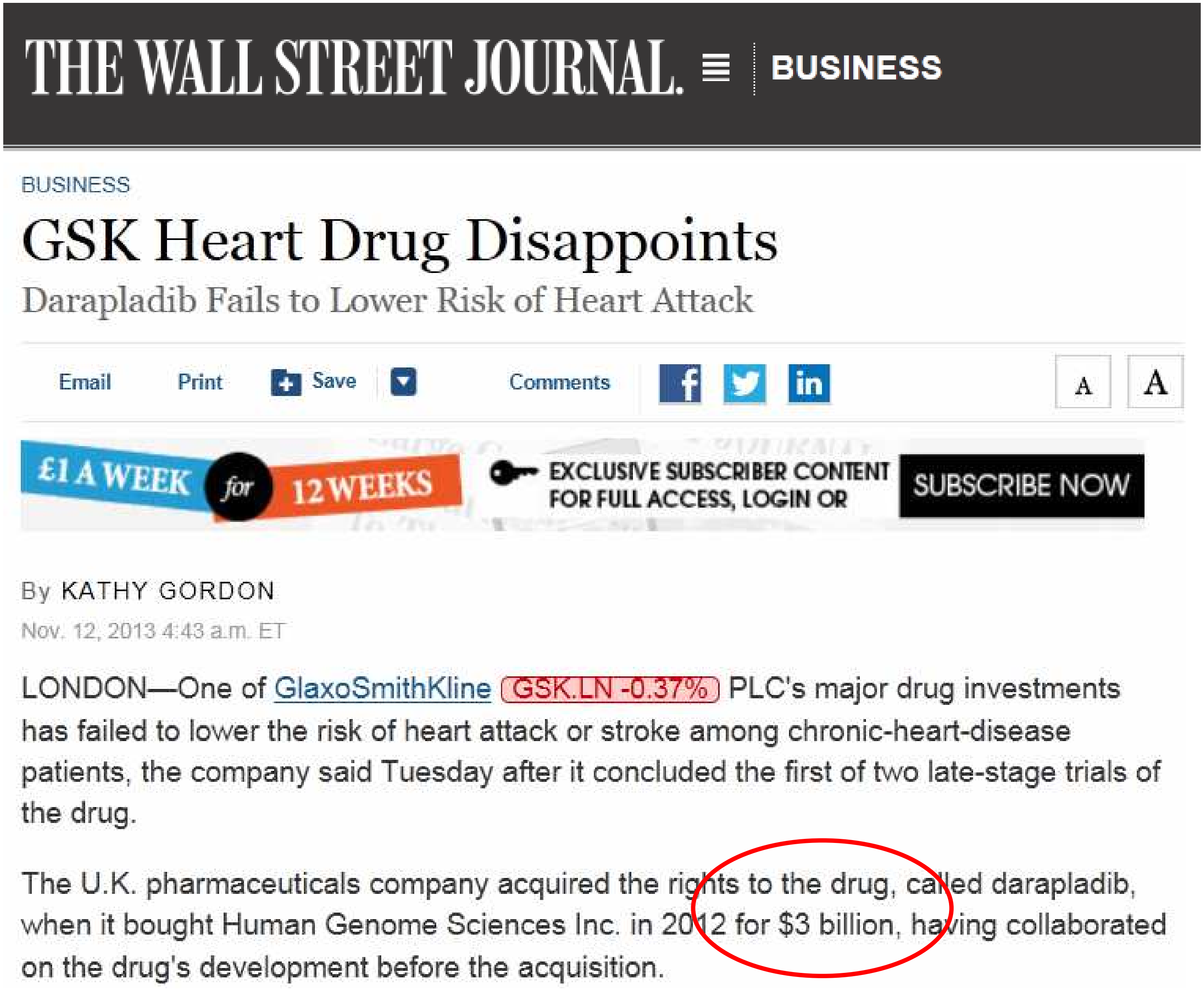
CHINA KADOORIE
BIOBANK
中国慢性病前瞻性研究

Drug Development Across the Industry: From Discovery to Approval

FIGURE 11: **The R&D Process: Long, Complex, and Costly**



- **For 5-10,000 compounds discovered, only 1 becomes a FDA-approved drug**
- **It takes 10-15 years to develop a new drug, costing ~US$1.3 billion**
- **Despite soaring cost, the annual No. of approved drugs halved since 1996**



THE WALL STREET JOURNAL. ≡ | BUSINESS

BUSINESS

GSK Heart Drug Disappoints

Darapladib Fails to Lower Risk of Heart Attack

By KATHY GORDON

Nov. 12, 2013 4:43 a.m. ET

LONDON—One of GlaxoSmithKline (GSK.LN -0.37%) PLC's major drug investments has failed to lower the risk of heart attack or stroke among chronic-heart-disease patients, the company said Tuesday after it concluded the first of two late-stage trials of the drug.

The U.K. pharmaceuticals company acquired the rights to the drug, called darapladib, when it bought Human Genome Sciences Inc. in 2012 for $3 billion, having collaborated on the drug's development before the acquisition.

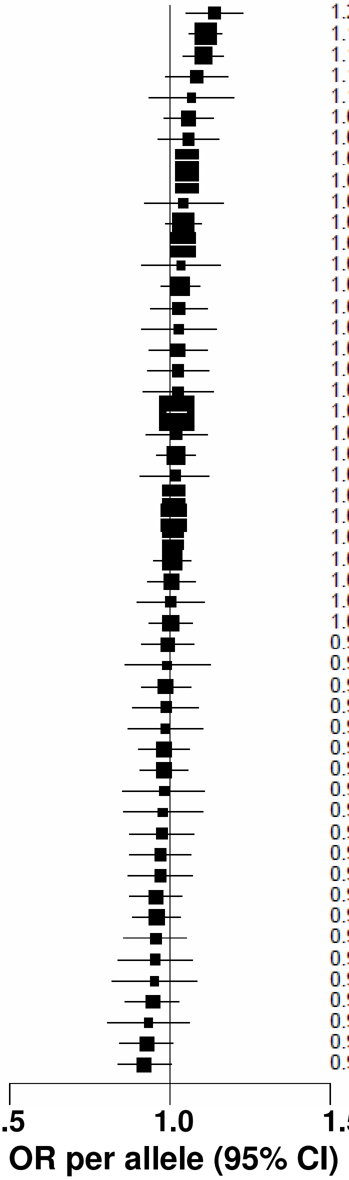
Lp-PLA_2

- A phospholipase enzyme carried on LDL and macrophages in atherosclerotic plaques
- Elevated activity predicts CVD risk, but causal effect uncertain
- Null variants in PLA2G7F (found only in East Asians), gene encoding Lp-PLA_2, reduces enzyme activity
- In animal models inhibitors of Lp-PLA_2 (darapladib) reduced coronary atherosclerosis
- Two trials assessed the effects of darapladib in 30,000 patients

CKB: using PheWAS approach to assess the efficacies and safeties of the inhibition of Lp-PLA_2 in 100K participants

CKB: Examples of PheWAS of genetic variant or GRS

Endpoint	No. of cases	No. of controls	Odds Ratio (95% CI)	P for trend
Non−insulin−dependent diabetes (E11.−)	1061	90262	1.21 (1.07, 1.37)	0.00253
Diabetes mellitus (E10−E14)	3290	88033	1.17 (1.09, 1.26)	3.3e−05
Unspecified diabetes mellitus (E14.−)	2185	89138	1.16 (1.06, 1.26)	0.0013
Dizizness and giddiness (R42)	1091	90232	1.12 (0.98, 1.28)	0.318
Other intervertebral disc disorders (M51.−)	498	90825	1.10 (0.91, 1.32)	0.328
Dyspepsia (K30)	1638	89685	1.09 (0.98, 1.21)	0.599
Unspecified chronic bronchitis (J42)	1008	90315	1.08 (0.95, 1.24)	0.235
Chronic ischaemic heart disease (I25.−)	4128	87195	1.08 (1.01, 1.16)	0.0267
Ischaemic stroke (I63)	3932	87391	1.08 (1.01, 1.16)	0.0367
Other disorders of bone (M89.−)	601	90722	1.06 (0.90, 1.26)	0.491
Cerebral infarction (I63.−)	3308	88015	1.06 (0.98, 1.14)	0.13
Any stroke (I60−I61,I63−I64)	4897	86426	1.06 (0.99, 1.13)	0.073
Acute laryngitis and tracheitis (J04.−)	639	90684	1.05 (0.88, 1.25)	0.963
Essential hypertension (I10)	3046	88277	1.05 (0.96, 1.14)	0.275
Other soft tissue disorders [not specified](M79.−)	1232	90091	1.04 (0.92, 1.18)	0.529
Spondylosis (M47.−)	750	90573	1.04 (0.89, 1.22)	0.642
Transient cerebal ischaemic attacks (G45.−)	1168	90155	1.04 (0.92, 1.18)	0.568
Other respiratory disorders (J98.−)	1058	90265	1.04 (0.91, 1.18)	0.584
Other joint disorders (M25.−)	841	90482	1.04 (0.89, 1.20)	0.64
Cardiovascular (I00−I09,I16−I25,I27−I88,I95−I99)	12070	79253	1.03 (0.99, 1.08)	0.189
Other special examinations and investigations (Z01.−)	1199	90124	1.03 (0.90, 1.17)	0.887
Acute upper respiratory infections (J06.−)	4408	86915	1.03 (0.95, 1.12)	0.525
Other inflammation of vagina and vulva (N76.−)	1050	90273	1.02 (0.88, 1.18)	0.427
Gastritis and duedenitis (K29.−)	5383	85940	1.02 (0.95, 1.08)	0.599
Cerebrovascular disease (I60−I69)	6350	84973	1.02 (0.96, 1.08)	0.569
Dorsalgia (M54.−)	4327	86996	1.02 (0.95, 1.09)	0.673
Malignant neoplasms (C00−C97)	2762	88561	1.01 (0.93, 1.10)	0.784
COPD (J41−J44)	1755	89568	1.01 (0.91, 1.12)	0.848
Acute pharyngitis (J02.−)	943	90380	1.00 (0.87, 1.16)	0.963
Bronchitis[not specified as acute or chronic](J40)	2233	89090	1.00 (0.91, 1.10)	0.964
Injury of unspecified body region (T14.−)	1516	89807	0.99 (0.88, 1.11)	0.887
Cancer of bronchus and lung (C34.−)	546	90777	0.99 (0.82, 1.19)	0.914
Acute nasopharyngitis (J00)	2180	89143	0.98 (0.89, 1.09)	0.0594
Other dermatitis (L30.−)	980	90343	0.98 (0.85, 1.13)	0.808
Urethritis and urethral syndrome (N34.−)	792	90531	0.98 (0.84, 1.15)	0.427
Cholecystitis (K81.−)	1649	89674	0.98 (0.88, 1.09)	0.648
Other non−infective gastroenteritis and colitis (K52.−)	1858	89465	0.97 (0.88, 1.08)	0.614
Cholelithiasis (K80.−)	607	90716	0.97 (0.82, 1.16)	0.766
Other gastroenteris (A09.−)	641	90682	0.97 (0.82, 1.15)	0.745
Other COPD (J44.−)	946	90377	0.97 (0.84, 1.11)	0.626
Gingivitis and peridontal disease (K05.)	1184	90139	0.96 (0.84, 1.10)	0.561
Haemorraghic stroke (I61)	1014	90309	0.96 (0.84, 1.10)	0.561
Abdominal and pelvic pain (R10.−)	1577	89746	0.94 (0.84, 1.06)	0.318
Other arthritis (M13.−)	1962	89361	0.94 (0.85, 1.04)	0.256
Pneumonia organism unspecified (J18.−)	1036	90287	0.94 (0.82, 1.08)	0.38
Calculus of kidney and ureter (N20.−)	745	90578	0.94 (0.80, 1.10)	0.427
Pain in throat and chest (R07.−)	576	90747	0.94 (0.78, 1.12)	0.427
Malaise and fatigue (R53)	1794	89529	0.93 (0.83, 1.04)	0.0851
Acute bronchitis (J20.−)	613	90710	0.91 (0.77, 1.09)	0.301
Headache (R51)	1537	89786	0.90 (0.81, 1.02)	0.0851
Other cerebrovascular disease (I67.−)	1565	89758	0.90 (0.80, 1.01)	0.0594



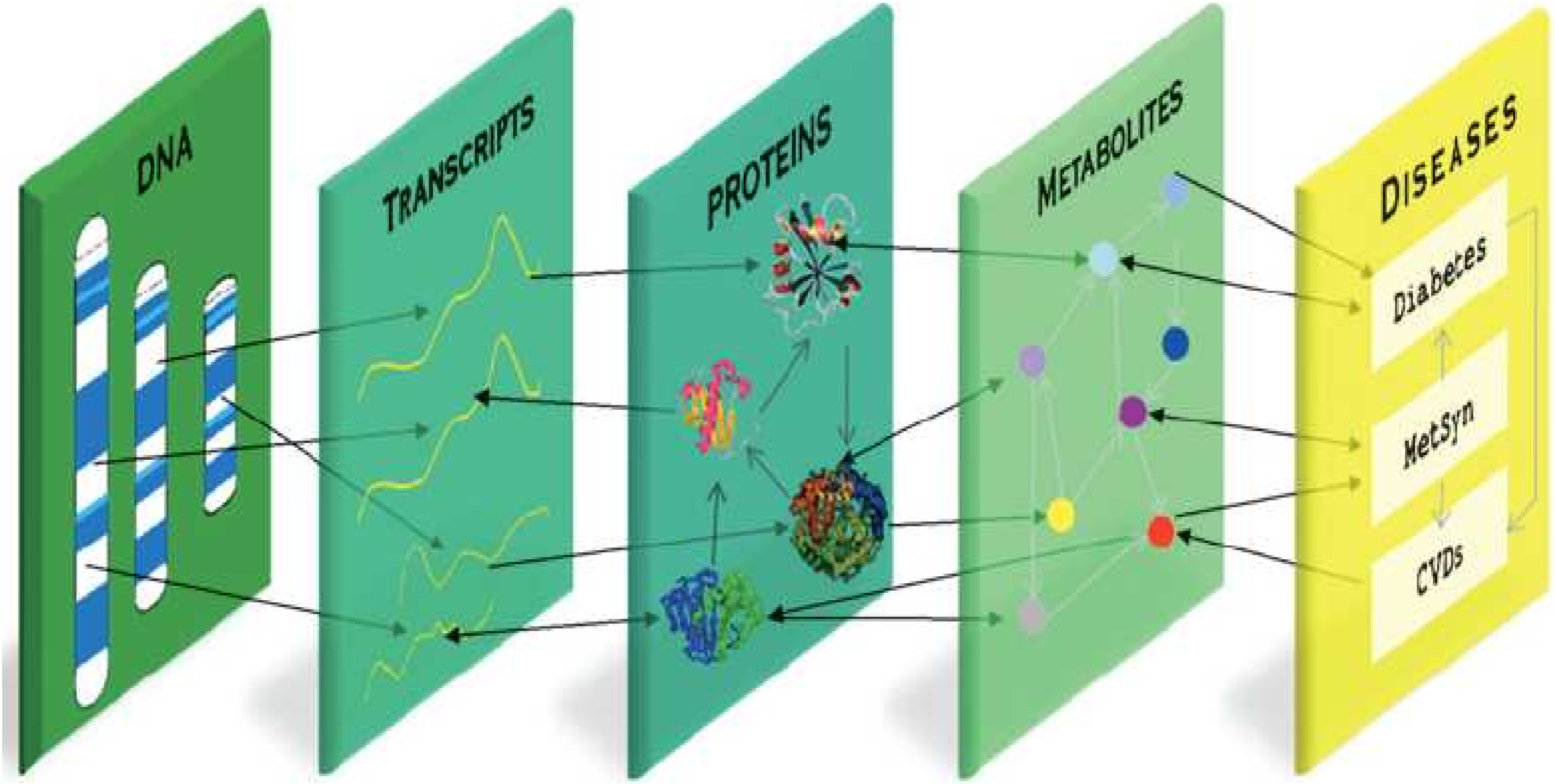
To compare disease risk between extreme thirds of a gene score based on all SNPs

China Kadoorie BIOBANK

Unpublished results

CKB: opportunities for multi-*omics* research

Genome | Transcriptome | Proteome | Metabolome | Phenome



~40,000,000 | ~20,000 | ~100,000 | ~5,000 | ~5,000

We aim to genotype 510,000 samples using customised array

CHINA KADOORIE BIOBANK

Genomics in medicine

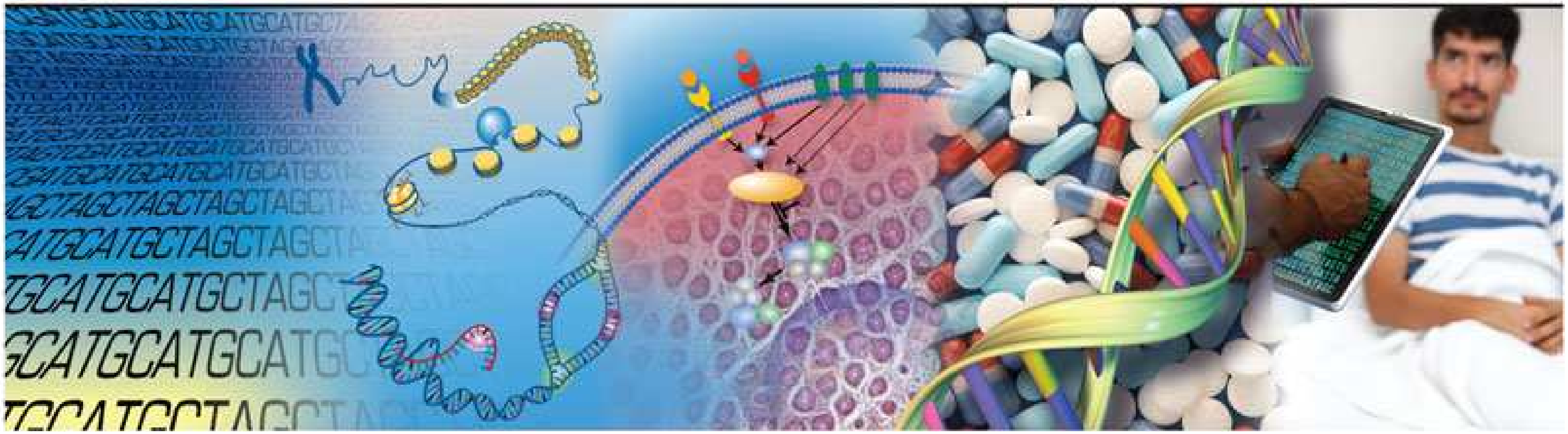
Understanding the structure of genomes

Understanding the biology of genomes

Understanding the biology of diseases

Advancing the science of medicine

Improving the effectiveness of healthcare



DNA sequence
Genes variation

Gene regulation
Gene function

Pathways
Mechanisms

Diagnosis
Treatment
Prevention

Risk prediction
Targeted therapy

ED Green *et al. Nature* 2011; **470**:204-213

nature

CKB: Opportunities for BIG DATA using EMR and multi-*omics* information

- Great increase in the range of diseases that can be studied
- Improved power, disease classification & patient stratifications
- Better understanding of genetic factors on multiple diseases with shared pathways/mechanisms
- Further exploration of causative genes at loci discovered previously from trans-ethnic studies
- Identification of novel biomarkers as therapeutic targets
- Better predication of drug response and prognosis

Need novel tools for data handling, analyses and interpretation

Oxford Big Data Institute