

# **Obtaining phenotype and outcome data from health records: China Kadoorie Biobank experience**

**Zhengming CHEN**

CKB Principal Investigator  
Professor of Epidemiology  
Nuffield Dept. of Population Health  
University of Oxford, UK  
(*zhengming.chen@ndph.ox.ac.uk*)

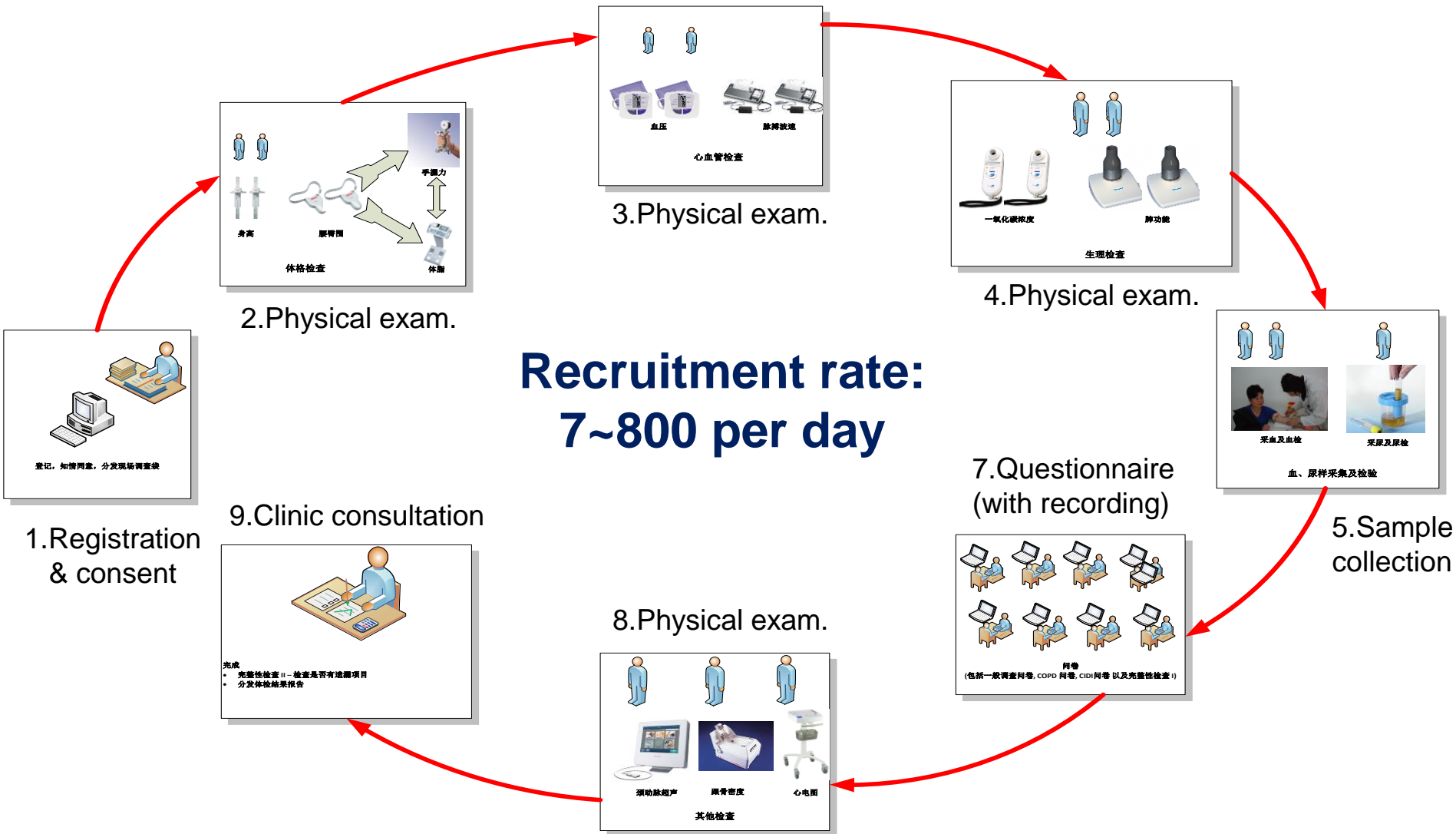
International Cohorts Summit, Duke University, USA,  
26-27 March 2018

# China Kadoorie Biobank (CKB)

- >512K recruited from 10 localities in 2004-08
- Participants interviewed, measured, and gave plasma and DNA (urine) for long-term storage
- All followed up indefinitely via electronic record linkage to deaths and ALL hospital episodes
- Periodic resurvey of 5% surviving participants (for enhancements and sources of variation)

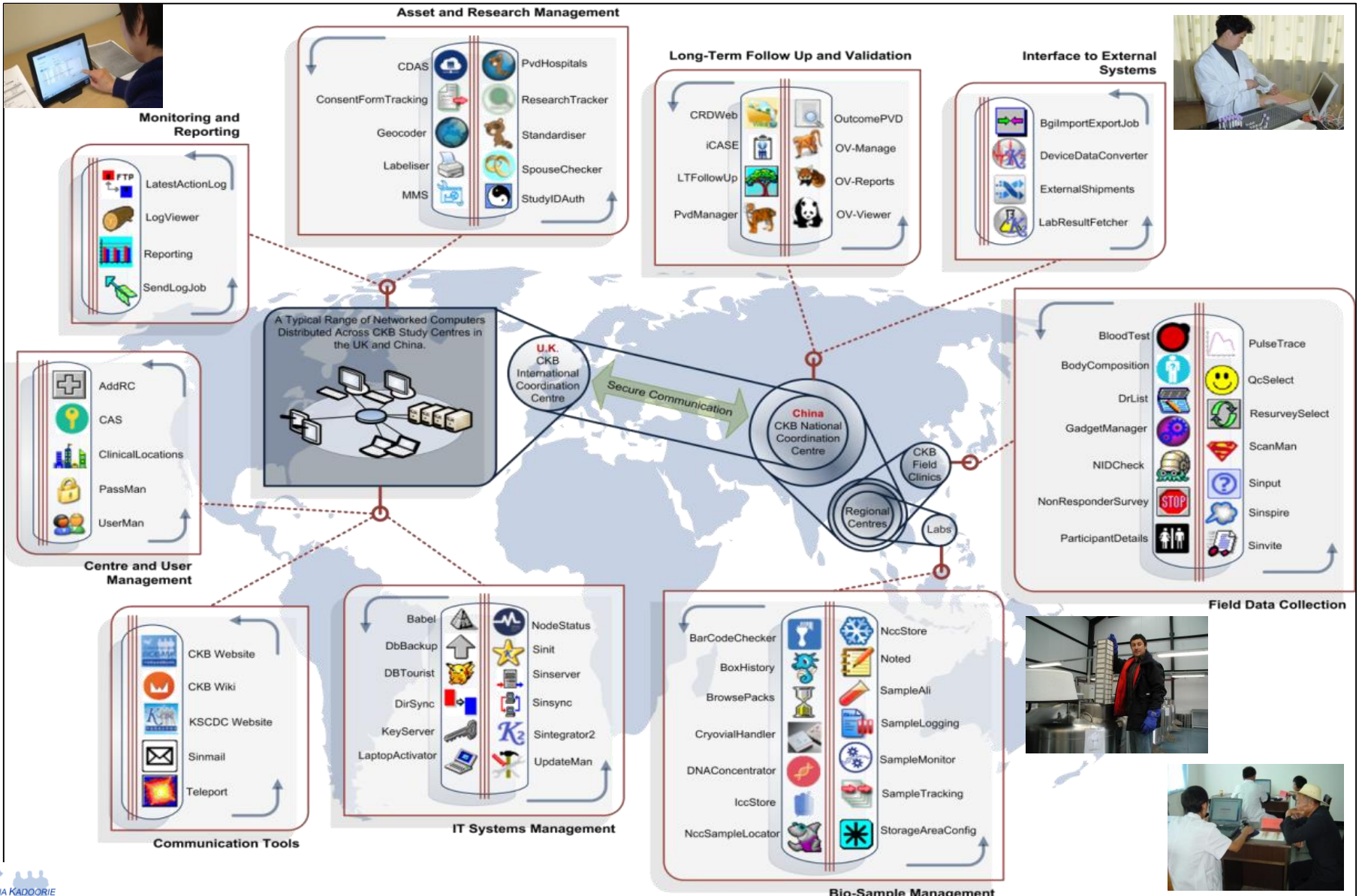
Informed consent for linkages to health records and unspecified research use of stored samples

# CKB: Clinical stations at local assessment centre



**The clinic visit took 60-90 minutes, with daily statistical monitoring**

# CKB: Supported by >90 bespoke IT systems



# CKB: Fully established with 10-year follow-up

## Questionnaire

SES, smoking, alcohol, tea, diet, physical activity, indoor air pollution, sleep, reproductive patterns, medical history

## Measurements

Blood pressure, height, weight, lung function, heart rate, bone density, exhaled CO, ECG, cIMT, ambient temperature, ambient air pollution, blood lipids, metabolites, proteomics, infectious markers, genetics

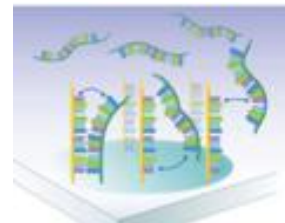
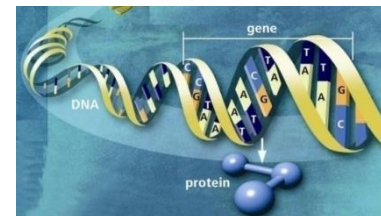
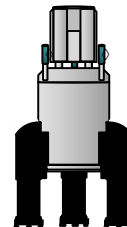
## Electronic health records

>1,300 different diseases, 43K deaths, <5K lost to follow-up, ~0.9 million hospitalizations, >100 million chargeable items

[www.ckbiobank.org](http://www.ckbiobank.org)



## Data is growing rapidly



# CKB: Follow-up through record linkages



Death registries

居民死亡医学证明书

南京市 (县) 区 (县) 街道 (乡) N: 0404003

姓名: 曹少华 身份证号: 407107195408180011

性别: 男 年龄: 56 文化: 小学 职业: 个体户

出生日期: 1954年8月18日 死亡日期: 2007年7月4日

家庭住址: 南京市江宁区... 联系电话: 88755870

致死的主要疾病诊断: 冠心病

直接导致死亡的其他重要情况: 心梗发作

引起 (a) 的疾病或情况: 冠心病

引起 (b) 的疾病或情况:

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 (省) (市) (县) (区) (市) 级 3 县 (区) 级 4 卫生 5 村 6 未报 7 其他及病最高诊断单位: 医院 医院 医院 医生 医生 不详

死者生前上述 (1) 尸检 (2) 临床 (3) 理化 (4) 临床 (5) 死因推断 (6) 不详

病最高诊断依据: 尸检 临床 理化 临床 死因推断 不详

住诊号: 33782 医师签名: 任国良 填报日期: 年 月 日

本死亡原因: ICD 编码: 统计分类号: 诊断专用章

外伤中毒的外部原因: Y 编码: 统计分类号: 诊断专用章

Active follow-up

Outcome Follow up in CKB

Disease registries

桐乡市城乡居民合作医疗保险管理系统

姓名	性别	人员类别	结算单号	结算次数	合作医疗卡号	就诊日期	结算日期	冲账日期	疾病名称
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-07-07	2007-07-14	2007-07-14	09:24:27 正常
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-07-05	2007-07-14	2007-07-14	09:24:27 正常
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-06-05	2007-07-05	2007-07-05	15:46:25 正常
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-07-15	2007-07-15	2007-07-15	11:28:06 正常
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-06-17	2007-07-13	2007-07-13	14:08:58 正常
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-07-01	2007-07-14	2007-07-14	10:01:29 正常
曹少华	男	普通	33040301000000000001	1	33040301000000000001	2007-06-03	2007-07-04	2007-07-04	07:38:18 正常

项目	项目单位	项目单价	项目数量	项目金额	自付比例	自付金额	统筹金额	ICD编码
19	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
20	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
21	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
22	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
23	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
24	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
25	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
26	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
27	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
28	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
29	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00
30	意外伤害赔偿	按有关规定	0.0000	0.00	0.0000	0.0000	0.0000	0.00

National health insurance

哈尔滨市慢性非传染疾病报告卡

门诊号: 01020 住院号: 01020 病种: 死亡 更正 卡片编号: 894

患者: 曹少华 性别: 男 年龄: 56 民族: 汉族

职业: 个体户 文化程度: 小学

工作单位: 个体户 电话: 52628377

患者: 曹少华 性别: 男 年龄: 56 民族: 汉族

职业: 个体户 文化程度: 小学

工作单位: 个体户 电话: 52628377

疾病日期: 04年8月10日

诊断日期: 04年8月10日

死亡日期: 04年8月10日

诊断依据: 临床, 病理, 心电图, X光, 超声, CT, 内镜

报告单位: 曹少华 报告人: 曹少华 (病种: 死亡)

报告日期: 04年8月10日

修正病名: 冠心病

# National health insurance system in China

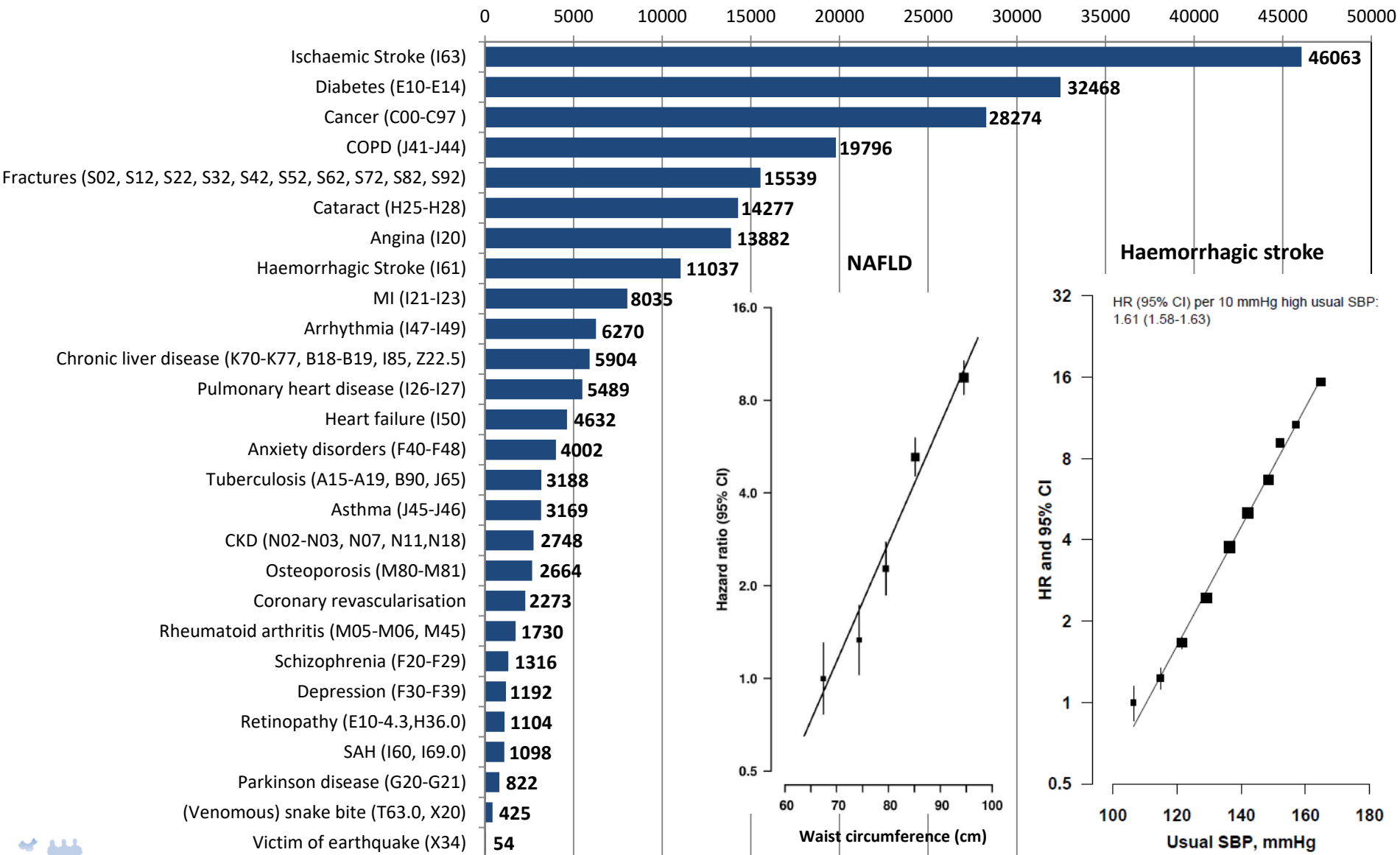
(supplementing death and cancer registries)

- Introduced during 2004-6, with almost universal coverage in CKB areas by 2010
- Multiple disease diagnoses, with ICD-10 codes plus disease descriptions and >2,500 procedure codes
- Managed electronically at city level, with detailed chargeable items for reimbursement purposes
- Lacks certain details (e.g. cancer pathology) required for disease sub-phenotyping

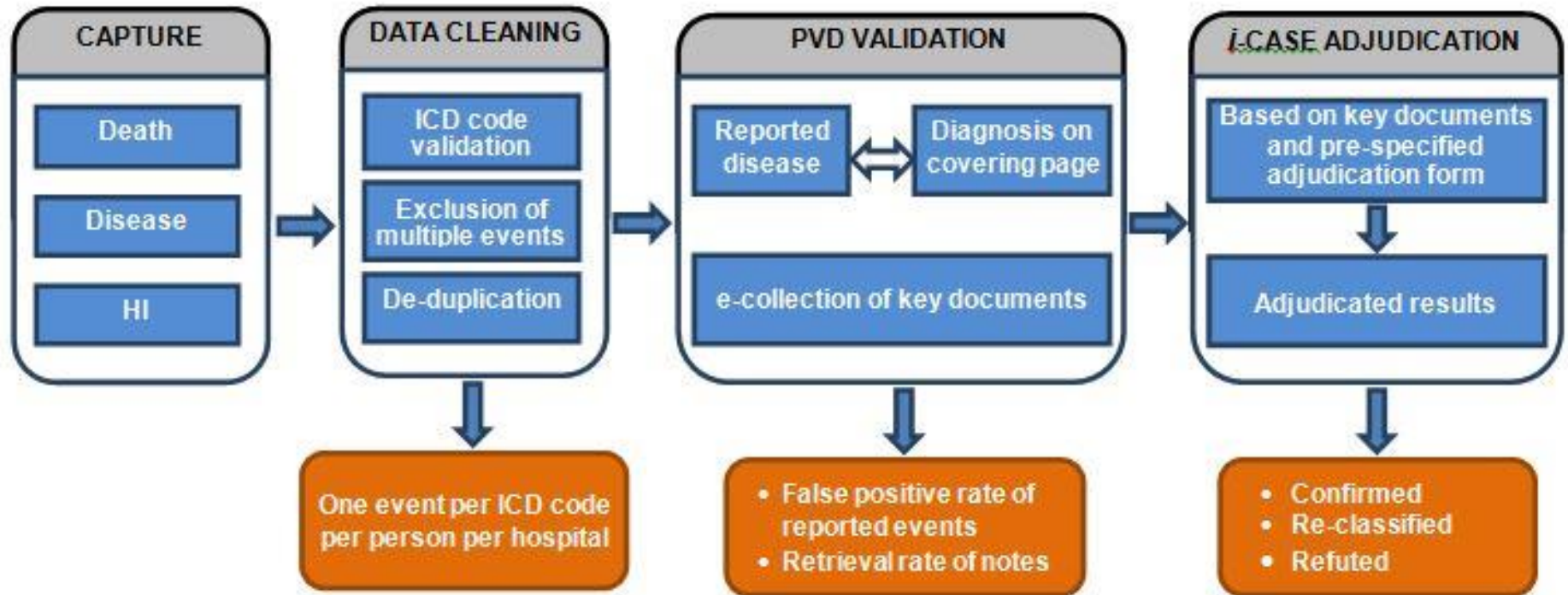
Nearly all CKB participants now linked to the health insurance databases via unique national ID number

# CKB: Participants with selected diseases in 10 years

(43K deaths, 0.9M hospital admissions; 2017 HI data are being processed)



# CKB: Procedures for improving disease phenotyping



Pilot study of ~1000 cases for specific disease before deciding whether to undertake systematic adjudication

# CKB: Disease standardisation and coding tool

Standardization Tool

File Automatic Manual Verify Options

Split Code Translate Delete Copy Paste Undo Redo Find... Filter... Sort...

	description	primary_code	secondary_code	translation	review_notes
4591	心肌病(急性)				
4592	心肌病(急性)				
4593	心肌病,肥厚性				
4594	心肌病、冠心病				
4595	心肌病上消化道出血心力衰竭 NOS慢性阻塞性肺病伴有急性加重 NOS高尿酸血症				
4596	心肌病	I42.900		Cardiomyopath	
4597	上消化道出血	K92.208		Upper gastroin	
4598	心力衰竭	I50.900		Heart Failure	
4599	慢性阻塞性肺病	J44.900		Chronic obstruc	
4600	伴有急性加重			With acute exa	
4601	高尿酸血症	E79.000		Hyperuricemia	
4602	心肌病心律失常				
4603	心肌病心律失常房颤房早室早				
4604	心肌病心功能II级肺部感染				
4605	心肌病心功能IV级肝功能异常肺部感染低钾血症高尿酸血症心功能III级				
4606	心肌病急性支气管炎				

Rows: 41570 Ready

# CKB: Verifying reported diagnosis



## Overview of Case Validation

Personal Information		
Name	smith	
Sex	Male	
Date of birth	1952-04-03	
Hospitalisation number	12	

Hospital Admissions		
Admission date	Discharge date	Is this an electronic note?
2008-05-05	2008-05-15	Yes
2010-01-04	2010-01-07	No

Vital Status	
What is the patient's vital status at discharge from the last note based on the discharge date in this hospital?	Alive

Main Diagnosis	
Please choose the disease sub-type from the	Ischemic stroke

The following section shows the updated information for this patient. Please check it carefully to confirm that the details are all correct.

Is this information correct?

Yes  No

Please enter your user name and password and select 'Finish' to sign-off the validation. After this you will not be able to make any changes.

User name

Password

Finish

Previous

# CKB: Adjudicating & phenotyping major diseases

(>70K adjudicated: 30K stroke, 25K IHD, 15K cancer, >3K CKD)

**CKB Event Adjudication Form**

Medical Notes CT Page 1

Page 1

**彭州市中医医院**

[REDACTED]

姓名	胡庆云	性别	男	年龄	64岁	科室	
扫描技师		扫描方式		检查部位	头部		
门诊号		住院号		床号		仪器	Emotion 16 (2007)

**检查所见及其解释**

双侧基底节-放射冠区见多发小片状低密度影，边界欠清楚；双侧脑室周围脑白质区多发小片状、斑片状稍低密度灶，左右大致对称分布。  
脑室系统稍扩大，大部分脑沟、裂增宽。  
中线结构居中。  
颅骨未见明确骨质异常。

**诊断意见**

- 1、双侧基底节区多发腔隙性脑梗塞。
- 2、双侧脑白质变性。
- 3、老年性脑改变。

建议复查。

Study ID520366057

### 8. Clinical examinations

CT  Yes  No/Unknown

Test done before  Yes  No  
(If you are not clear whether the test is done before or after please select After admission.)

How long ago? <1 months

Diagnosis of prior test Lacunar infarct (LACI)

Test done in this admission  Yes  No

How many tests? 1

Date of the test 2010 - Jan - 22  
(If you are not clear about the date please enter the admission date.)

Test report Formal test report

Main findings Soften lesion (old lesion)

#### Location of brain lesion

Frontal lobe  Yes  No/Unknown

Side Left

Were any abnormalities consistent in age and site to the clinical presentation?  Yes  No

Temporal lobe  Yes  No/Unknown

Parietal lobe  Yes  No/Unknown

Occipital lobe  Yes  No/Unknown

Deep regions (Any of internal Capsule, Thalamus, Claustrum, Basal ganglia, Corona radiata, Corpus callosum)  Yes  No/Unknown

Cerebellum  Yes  No/Unknown

Brain stem (midbrain, Pons, medulla)  Yes  No/Unknown

# CKB: “traffic” light approach for outcome data

Clipboard		Font		Alignment		Number		Styles		Cells	
B15      fx      Unknown and unspecified causes of morbidity (R69)											
	A	B	C	D							
1	icd10	desc	event_count	participant_count	pvd						
3	I25	Chronic ischaemic heart disease (I25)	141900	46187							
4	I63	Cerebral infarction (I63)	164372	46014							
5	K52	Other noninfective gastroenteritis and colitis (K52)	105305	40866							
6			51989	36064							
7	K29	Gastritis and duodenitis (K29)	62053	32990							
8	I67	Other cerebrovascular diseases (I67)	41861	22621							
9	E14	Unspecified diabetes mellitus (E14)	91025	20778							
10	G45	Transient cerebral ischaemic attacks and related syndromes (G45)	27600	18799							
11	E11	Non-insulin-dependent diabetes mellitus (E11)	34733	18773							
12	J98	Other respiratory disorders (J98)	27109	16926							
13	J00	Acute nasopharyngitis [common cold] (J00)	33758	16629							
14	M54	Dorsalgia (M54)	27798	16589							
15	R69	Unknown and unspecified causes of morbidity (R69)	32567	16436							
16	M51	Other intervertebral disc disorders (M51)	26480	15944							
17	J18	Pneumonia, organism unspecified (J18)	24013	15592							
18	J40	Bronchitis, not specified as acute or chronic (J40)	32628	14232							
19	J44	Other chronic obstructive pulmonary disease (J44)	45249	14018							
20	M47	Spondylosis (M47)	23142	13968							
21	I20	Angina pectoris (I20)	27723	13755							
22	R42	Dizziness and giddiness (R42)	23106	12023							
23	M13	Other arthritis (M13)	20172	11631							
24	T14	Injury of unspecified body region (T14)	17998	11595							
25	J11	Influenza, virus not identified (J11)	21770	11454							
26	R53	Malaise and fatigue (R53)	22566	11300							
27	I61	Intracerebral haemorrhage (I61)	42288	10887							
28	N20	Calculus of kidney and ureter (N20)	18386	10777							
29	H26	Other cataract (H26)	14797	9927							
30	K81	Cholecystitis (K81)	14869	9769							
31	K80	Cholelithiasis (K80)	15580	9685							
32	J42	Unspecified chronic bronchitis (J42)	17705	9536							
33	J20	Acute bronchitis (J20)	14208	9339							
34	R10	Abdominal and pelvic pain (R10)	13600	9285							
35	K30	Dyspepsia (K30)	11704	8804							
36	I21	Acute myocardial infarction (I21)	25137	7865							
37	J02	Acute pharyngitis (J02)	14508	7824							
38	K05	Gingivitis and periodontal diseases (K05)	12304	7265							
39	I84	Haemorrhoids (I84)	9495	7193							
40	N76	Other inflammation of vagina and vulva (N76)	17261	7132							
41	I69	Sequelae of cerebrovascular disease (I69)	19670	6762							

# Future work for disease phenotyping

- Standardising and ICD-10 coding new events collected
- Processing and incorporating >100M chargeable items data to enhance disease phenotyping
- Extending outcome adjudication to several other diseases (e.g. heart failure, chronic liver disease)
- Developing automated algorithm to sub-phenotype stroke and other diseases according to clinical criteria
- Piloting collection of discharge summary pages and tumour tissue samples

# CKB: Open data access platform

(www.ckbiobank.org)



## Data Access

- [Data Overview](#) ▶
- [Data Access Policy](#) ▶
- [Access Procedures](#) ▶
- [FAQs](#) ▶
- [Related Documents](#) ▶
- [Login - Register](#) ▶

## Data Access Overview

The China Kadoorie Biobank (CKB) is a global resource for the investigation of lifestyle, environmental, blood biochemical and genetic factors as determinants of common diseases. The CKB study group is committed to making the cohort data available to the scientific community in China, the UK and worldwide to advance knowledge about the causes, prevention and treatment of disease. Detailed information on the CKB is given in the [About the Study](#), [Study Resources](#) and [Research](#) pages of this website.

This Data Access section provides:

- A [Data Overview](#) describing available datasets
- Details of current [Data Access Procedures](#).
- Information on [Data Access Policy and Principles](#)
- A page of [Frequently Asked Questions](#) and responses
- A link to [Register / Login](#) to the CKB Data Access System for researchers to apply for data

