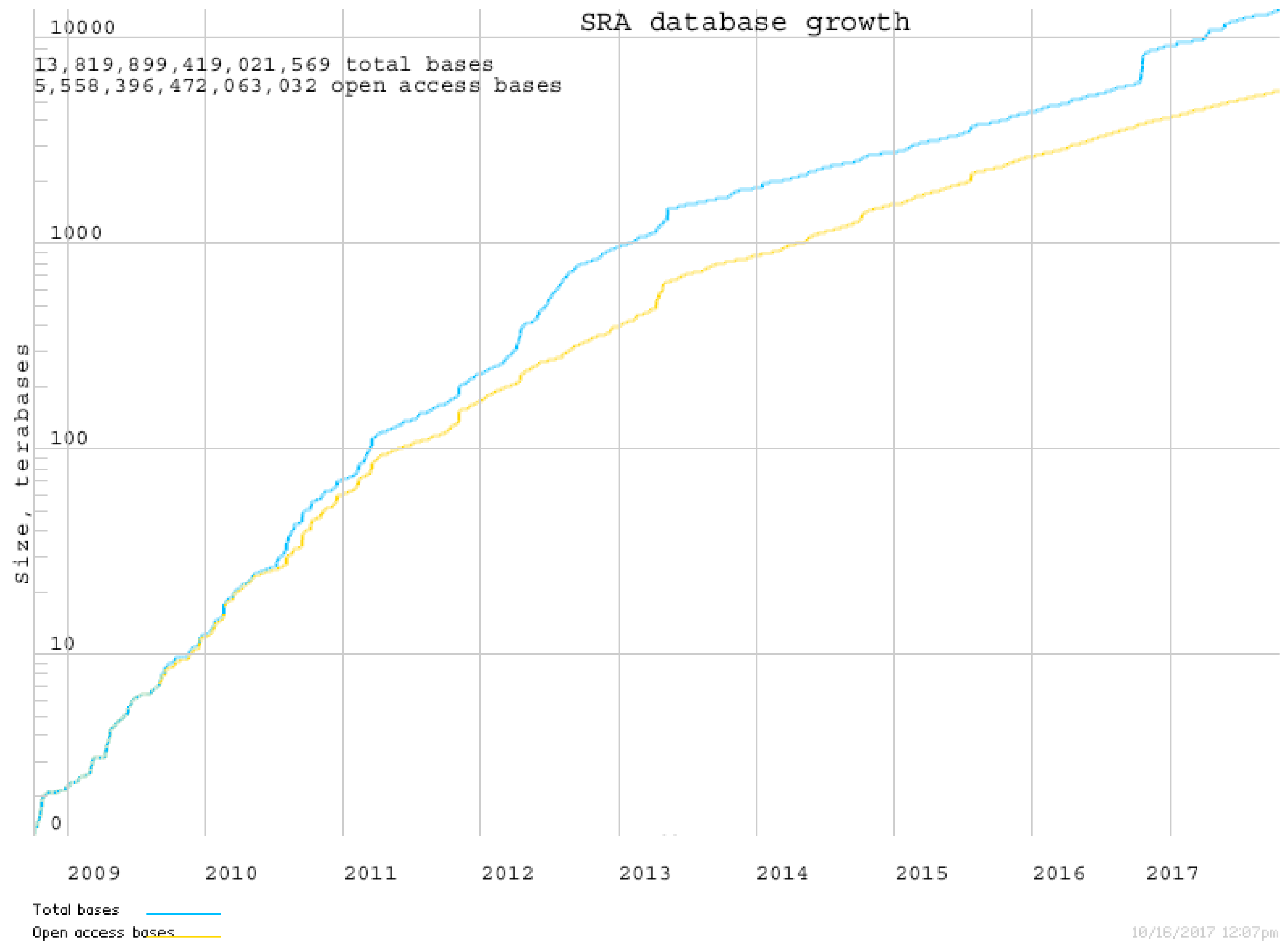
ASCO CANCERLINQ™
Learning Intelligence Network for Quality

International Coordination of Data Science Infrastructure: Some Insights from Biomedicine

George Komatsoulis, Ph.D.
Chief, Bioinformatics
CancerLinQ

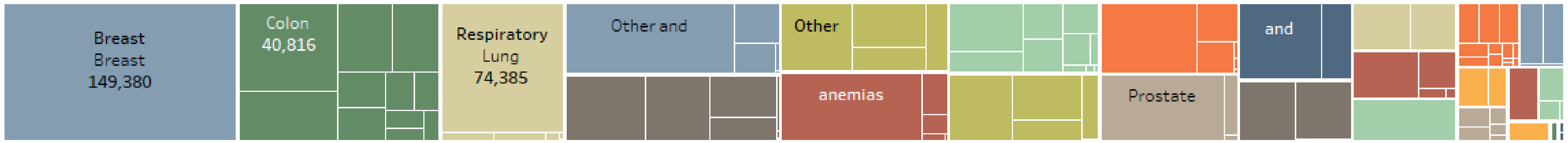

SRA database growth
13,819,899,419,021,569 total bases
5,558,396,472,063,032 open access bases
Size, terabases
10000
1000
100
10
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
Total bases
Open access bases
10/16/2017 12:07pm

CancerLinQ Patient Demographics

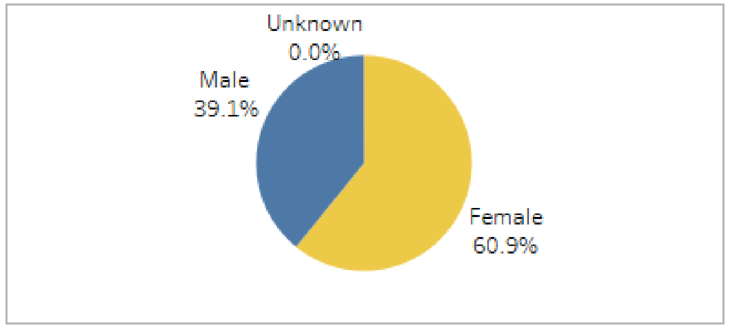
Patient Count: 710,754

Practice Count: 38

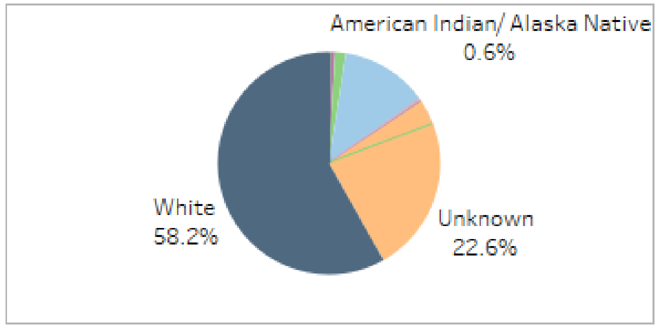
E&M Encounters: 4,900,216



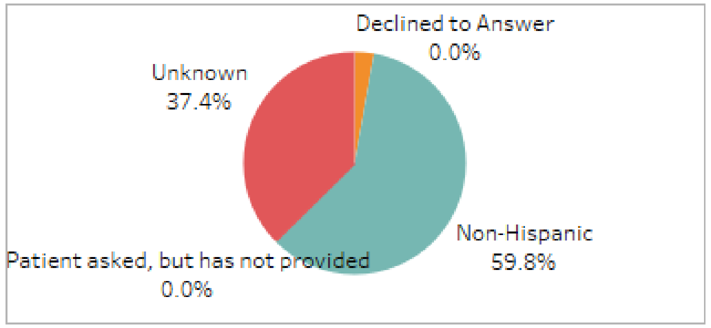
Gender



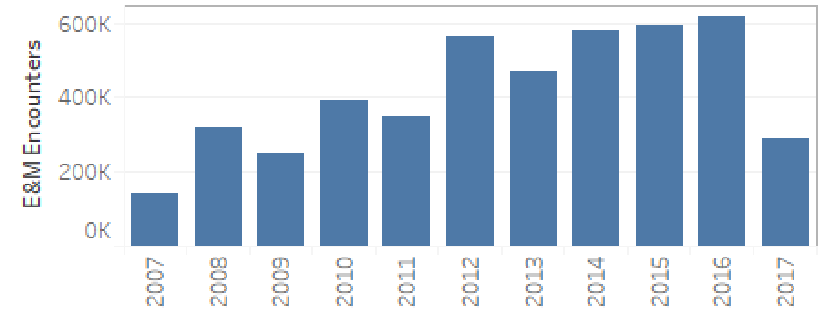
Race



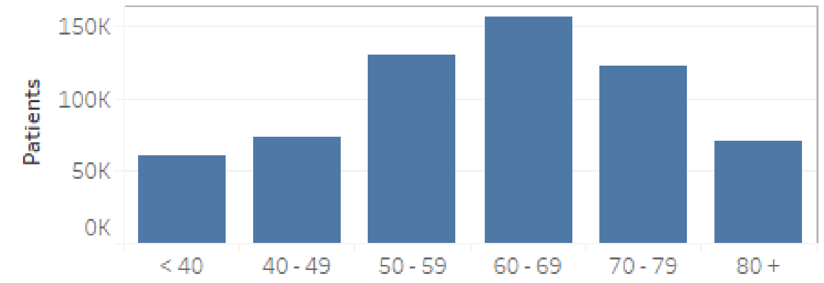
Ethnicity



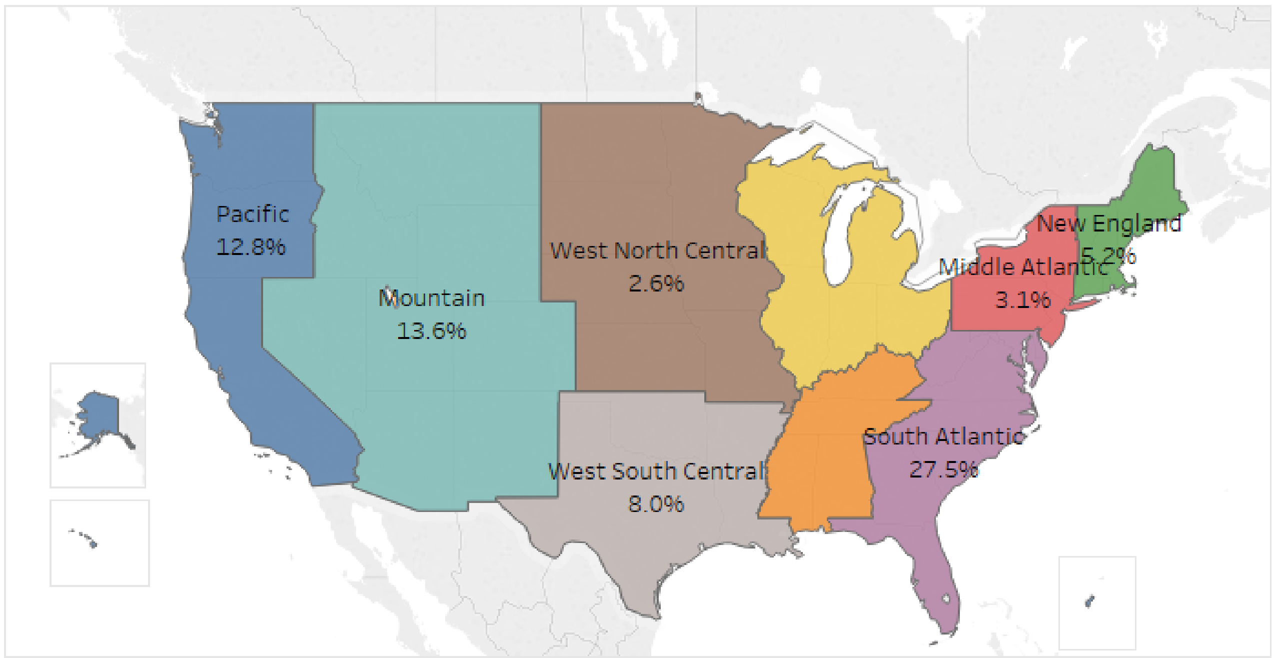
Year of First Encounter



Age at Diagnosis



Regional Distribution



DISEASES

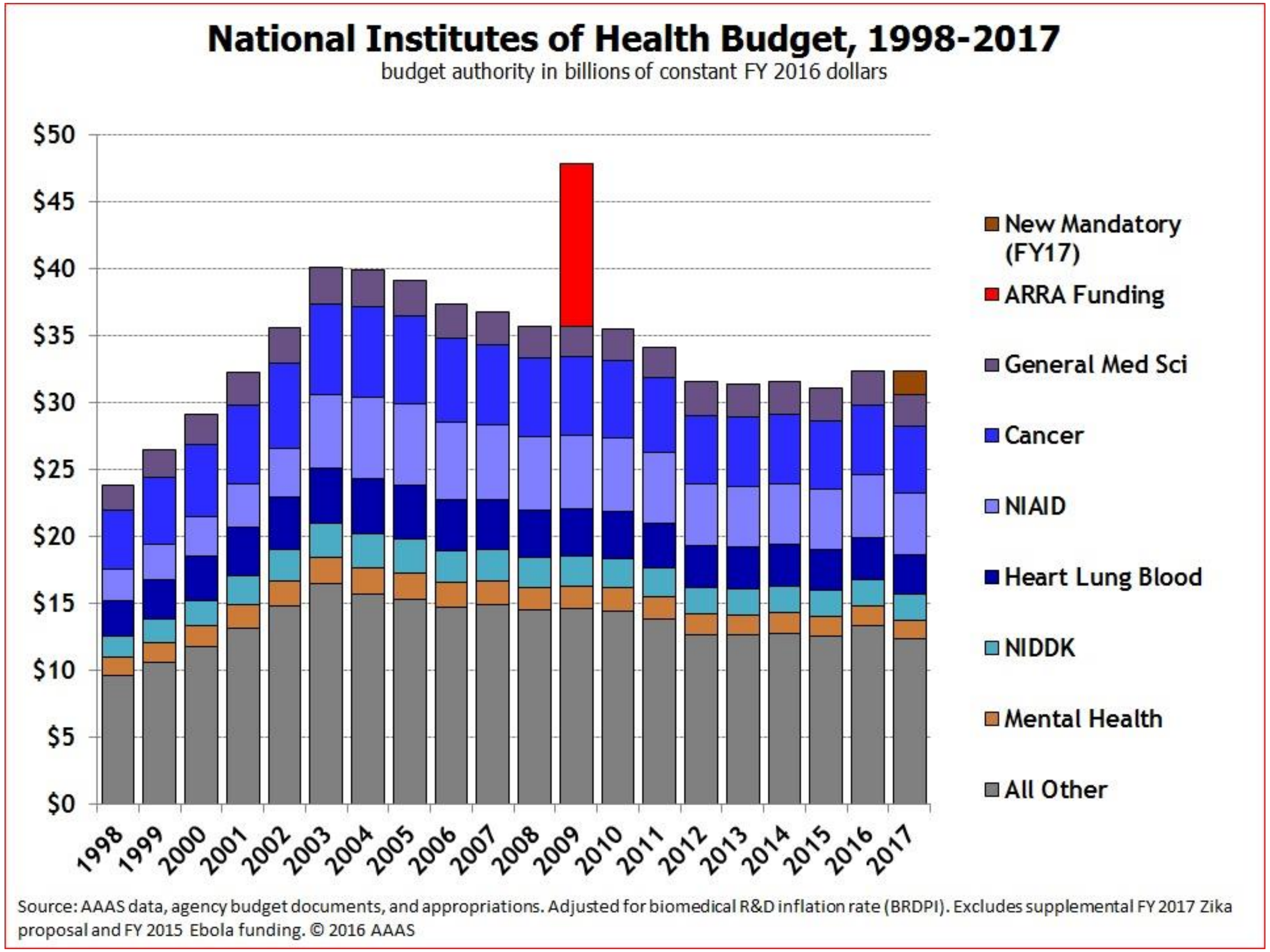
✓ Benign Hematology
✓ Cancer

SYSTEMS

✓ Aplastic anemias
✓ Bone and Connective ..
✓ Brain/CNS
✓ Breast
✓ Coagulation Defects
✓ Digestive
✓ Female Genital
✓ Head and Neck
✓ Hemolytic anemias
✓ Illdefined
✓ Leukemia
✓ Lymphoma
✓ Male Genital
✓ Melanoma/Skin
✓ Mesothelial/Soft Tiss..
✓ Multiple Myeloma
✓ Neuroendocrine
✓ Nutritional anemias
✓ Occular
✓ Other disorders of bl..
✓ Other Specified Type
✓ Respiratory
✓ Thyroid/Endocrine
✓ Unspecified

ANATOMIC SITE

✓ Acquired hemolytic a..
✓ Acquired pure red cel..
✓ Acute posthemorrha..
✓ Anemia in chronic dis..
✓ Appendix
✓ Bladder
✓ Bone
✓ Brain/CNS
✓ Breast
✓ Carcinoid Tumor
✓ Cervix


National Institutes of Health Budget, 1998-2017
budget authority in billions of constant FY 2016 dollars
$50
$45
$40
$35
$30
$25
$20
$15
$10
$5
$0
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
New Mandatory (FY17)
ARRA Funding
General Med Sci
Cancer
NIAID
Heart Lung Blood
NIDDK
Mental Health
All Other
Source: AAAS data, agency budget documents, and appropriations. Adjusted for biomedical R&D inflation rate (BRDPI). Excludes supplemental FY 2017 Zika proposal and FY 2015 Ebola funding. © 2016 AAAS

Increase ROI: What we *can* do

- Increase the efficiency with which resources are used to archive, store, manage and compute on biomedical data
  - Reduce unnecessary redundancy
  - Embrace technology that increases efficiency
- Extract more knowledge from each research effort
  - Implies that the data does not become meaningless electrons
- Recognize that there exists a data lifecycle and move various resources to different parts of the lifecycle based on scientific priorities and resource requirements
- Embrace FAIR principles

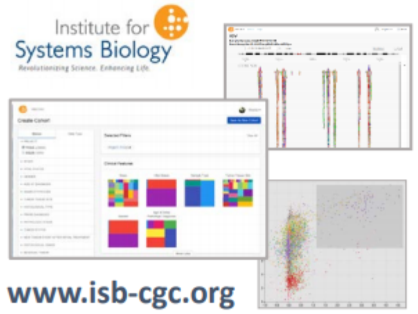
The *Commons* (Phil Bourne - 2014)

- Is *scalable* and exploits new computing models
- Is more *cost effective* given digital growth
- *Simplifies* sharing digital research objects such as data, software, metadata and workflows
- Makes digital research objects more *FAIR*: Findable, Accessible, Interoperable and Reusable

- **Uses Cloud Computing to Provide Scalable and Cost Effective Infrastructure**

ASCO CANCERLINQ Learning Intelligence Network for Quality

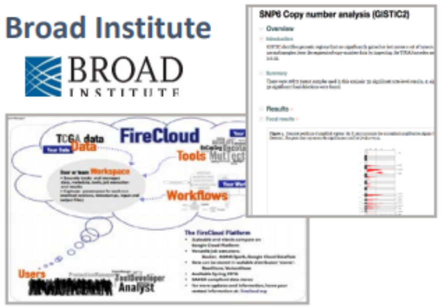
Institute for Systems Biology



www.isb-cgc.org

The Institute for Systems Biology (ISB) Cloud provides interactive and programmatic access to data, leveraging many aspects of the Google Cloud Platform. The **interactive ISB-CGC web-app** allows scientists to interactively define and compare cohorts, examine underlying molecular data for specific genes or pathways of interest, and share insights with collaborators. For computational users, programmatic interfaces and GCP tools such as **BigQuery, Genomics, and Compute Engine** allow users to perform complex queries from R or Python scripts, or run Dockerized workflows on sequence data available in cloud storage.

Broad Institute



www.firecloud.org

Broad Institute FireCloud is modeled after their **Firehose analysis infrastructure** and facilitates collaboration and provides a robust, scalable platform accessible to the community at-large. Using the elastic compute capacity of Google Cloud, FireCloud empowers analysts, tool developers, and production managers to perform large-scale analysis, engage in data curation, and store or publish results. Users can upload their own analysis methods and data to workspaces or run the **Broad's best practice tools and pipelines** on pre-loaded data.

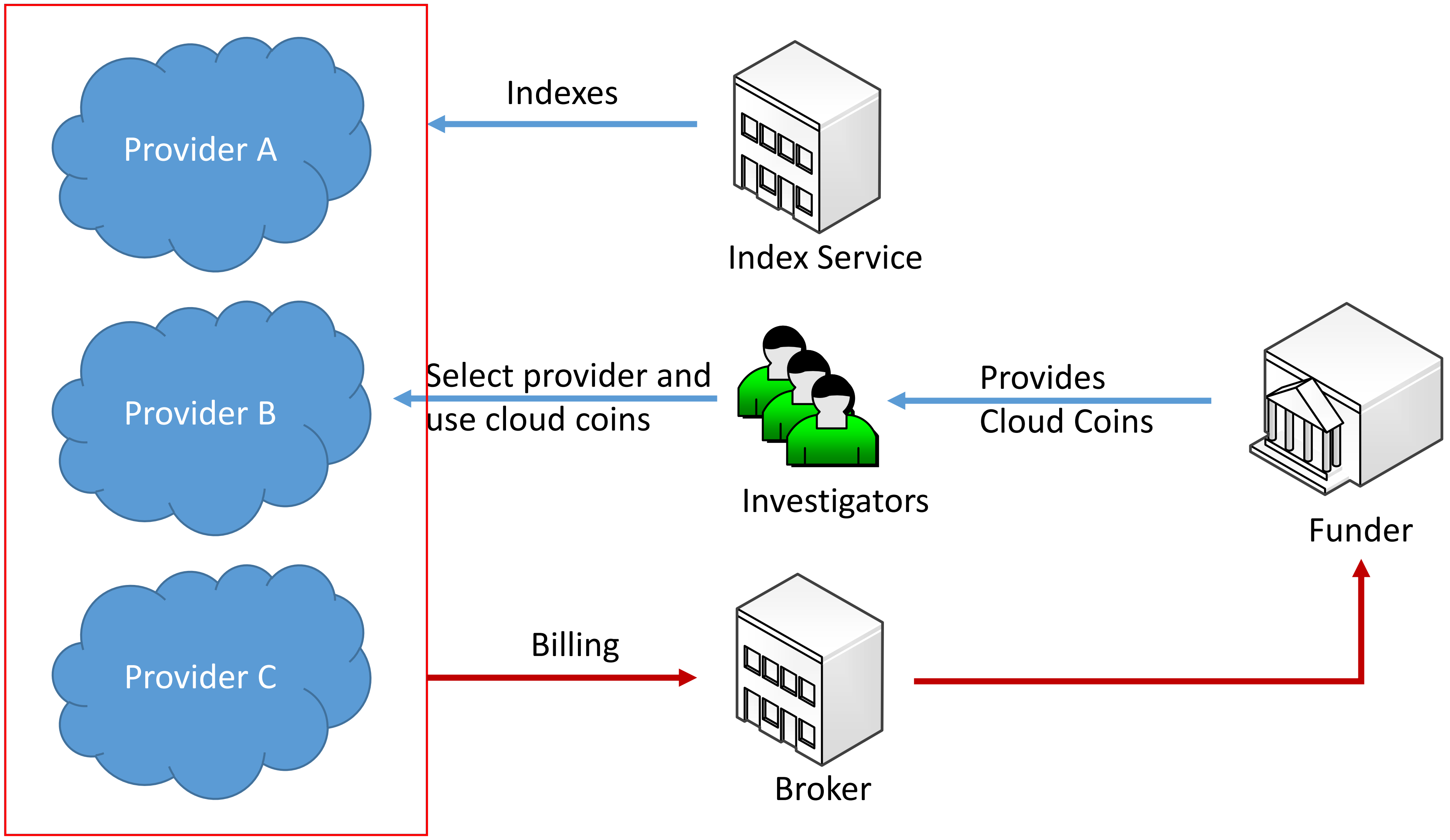
Seven Bridges

www.cancergenomicscloud.org

Seven Bridges Cancer Genomics Cloud enables researchers to collaborate on the analysis of large cancer genomics datasets in a secure, reproducible, and scalable manner. A **rich query system** allows researchers to find the exact data of interest and combine it with their own private data. Native implementation of the **Common Workflow Language specification** makes it easy for developers, analysts, and bench biologists to deploy, customize and run reproducible analysis methods to learn from genomics data faster.

NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS

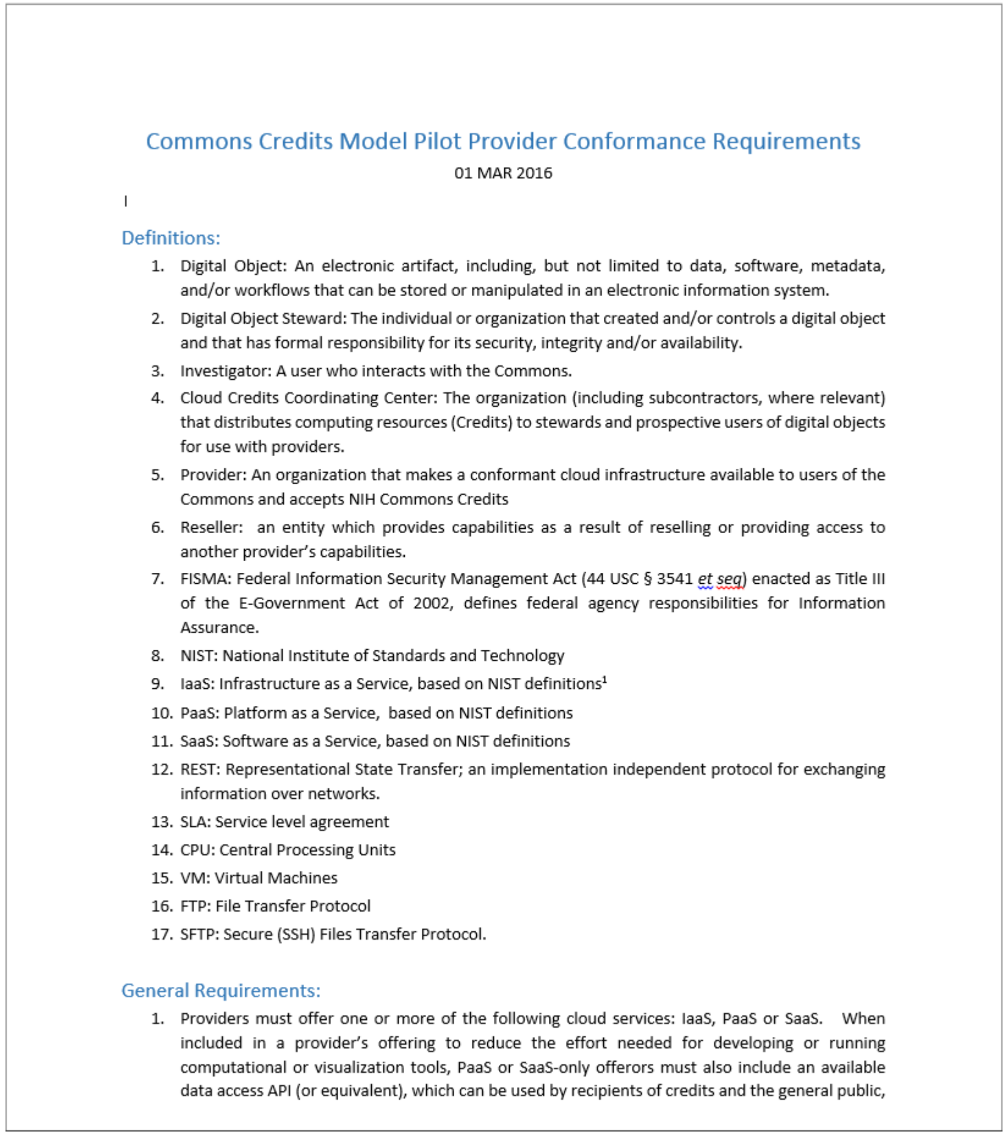
#NCIGDC


Provider A
Provider B
Provider C
Indexes
Index Service
Select provider and use cloud coins
Investigators
Provides Cloud Coins
Funder
Billing
Broker

How do cloud coins work from the point of view of an investigator?

- Investigators receive cloud coins worth a certain amount (in dollars, euros, etc.) that **can be used at the conformant provider(s) of their choice**
- Cloud coins are pre-purchased and applied to the account of the investigator with the relevant provider(s)
- As the investigator uses services with a conformant provider, the provider debits the value of the investigators usage against the pre-loaded credit amount
- **INVESTIGATORS ARE NOT BILLED BY PROVIDERS** AS LONG AS THEY DO NOT EXCEED THEIR CLOUD COIN ALLOCATION.

Who can provide cloud resources?



Commons Credits Model Pilot Provider Conformance Requirements

01 MAR 2016

I

Definitions:

1. Digital Object: An electronic artifact, including, but not limited to data, software, metadata, and/or workflows that can be stored or manipulated in an electronic information system.
2. Digital Object Steward: The individual or organization that created and/or controls a digital object and that has formal responsibility for its security, integrity and/or availability.
3. Investigator: A user who interacts with the Commons.
4. Cloud Credits Coordinating Center: The organization (including subcontractors, where relevant) that distributes computing resources (Credits) to stewards and prospective users of digital objects for use with providers.
5. Provider: An organization that makes a conformant cloud infrastructure available to users of the Commons and accepts NIH Commons Credits
6. Reseller: an entity which provides capabilities as a result of reselling or providing access to another provider's capabilities.
7. FISMA: Federal Information Security Management Act (44 USC § 3541 *et seq*) enacted as Title III of the E-Government Act of 2002, defines federal agency responsibilities for Information Assurance.
8. NIST: National Institute of Standards and Technology
9. IaaS: Infrastructure as a Service, based on NIST definitions[1]
10. PaaS: Platform as a Service, based on NIST definitions
11. SaaS: Software as a Service, based on NIST definitions
12. REST: Representational State Transfer; an implementation independent protocol for exchanging information over networks.
13. SLA: Service level agreement
14. CPU: Central Processing Units
15. VM: Virtual Machines
16. FTP: File Transfer Protocol
17. SFTP: Secure (SSH) Files Transfer Protocol.

General Requirements:

1. Providers must offer one or more of the following cloud services: IaaS, PaaS or SaaS. When included in a provider's offering to reduce the effort needed for developing or running computational or visualization tools, PaaS or SaaS-only offerors must also include an available data access API (or equivalent), which can be used by recipients of credits and the general public,

- "Conformant Providers"
- Can be IaaS, PaaS, SaaS
- Meet standards promulgated by the cloud coin authority for:
  - Capacity
  - Accessibility
  - Interfaces
  - Identifiers and Metadata
  - Networking
  - Authentication/Authorization
  - Information Assurance

Cloud Coin Pilot (NIH)

- 2 year (originally planned to be 3 year) pilot to test this business model to facilitate researcher use of cloud resources (enhance data sharing and potentially reduce costs).
- Contract with the CMS Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Center (FFRDC) managed by the MITRE corporation
  - FFRDCs are special purpose, government-owned but contractor-managed entities that meet R&D needs that can't be well managed by traditional grants and contracts
  - Examples: National Labs and organizations like RAND
- Pilot **will not directly interact with the existing grant system**.
  - Instead is modeled on the mechanisms being used to gain access to NSF and DOE national resources (HPC, light sources, etc.)
- The only required qualification for applying for cloud coins was that the **investigator must have an existing NIH grant**

Approved Vendors from NIH Pilot

Company	Direct/reseller	Type of Service
IBM	Selling own infrastructure, “SoftLayer”	IaaS
DLT	Reseller of Amazon Web Services infrastructure	IaaS
Onix	Reseller of Google infrastructure, and pathway to Broad Institute and Institute for Systems Biology service offerings	IaaS, PaaS, SaaS
Seven Bridges Genomics	Software as a Service provider operating on Amazon Web Services infrastructure	SaaS
MolBioCloud	Software as a Service provider operating on Amazon Web Services infrastructure, and Amazon Web Services infrastructure reseller	IaaS, SaaS
REAN Cloud	Reseller of Amazon Web Services infrastructure and Platform as a Service provider	IaaS, PaaS
Omnibond	“CloudyCluster” Platform as a Service provider and reseller of Amazon Web Services infrastructure	IaaS, PaaS
CDW-G	Reseller of Microsoft Azure infrastructure	IaaS

Issues During Initial Cloud Coin Distribution

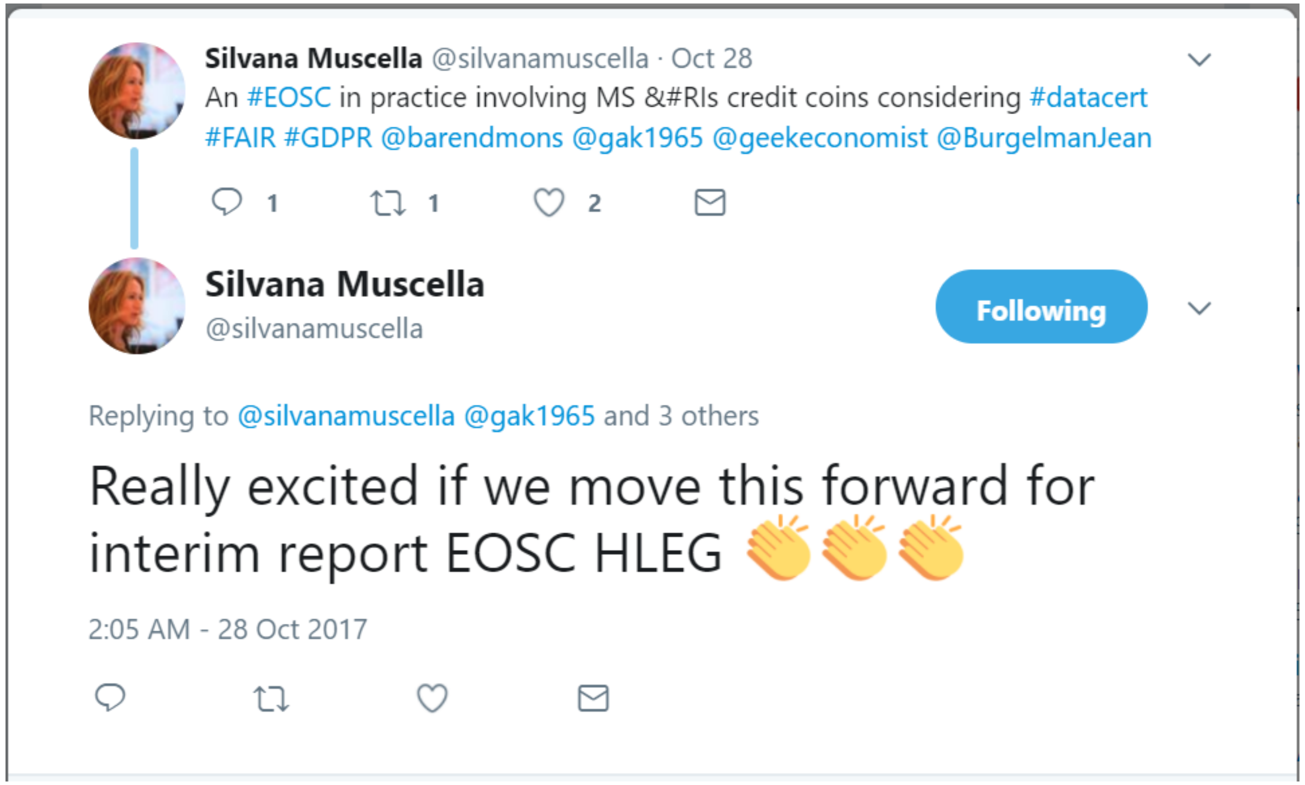
- Onboarding
  - Delays caused by university process to obtain accounts with providers
  - Secondary surety issues – providers generally preferred credit cards, universities prefer PO's
  - Resolution: Guidance to new applicants to start account provisioning before selection for cloud coins receipt, vendors being encouraged to accept PO's. All current vendors accept PO's now
- Business Associate Agreements (BAA's)
  - Needed for work under US Health Insurance Portability and Accountability Act (HIPAA)
  - BAA's generally between university and actual cloud provider (i.e. AWS rather than DLT) rather than reseller.
  - Concern about whether BAA required with reseller
  - Status: Resellers have **no access** to data in cloud provider. Single case where this has been an issue (UCSD/DLT), provider feels should be able to resolve

Some thoughts on financial Transactions

- NIH pilot pre-purchased services on behalf of investigators
  - Some providers had problems with this approach as the pre-purchase behaved as a debt on their balance sheets
  - Use of resellers resolved this issue
- More desirable transaction types (not available to US Government during pilot)
  - Use of escrow account with investigators given withdrawal rights up to amount of credit
  - Prepaid debit cards

ASCO CANCERLINQ™ Learning Intelligence Network for Quality

- European Open Science Cloud program interested in piloting Cloud Coins in upcoming interim report
- Obvious target would be genomic pipelines that have been containerized for easier movement across platforms



Silvana Muscella @silvanamuscella · Oct 28
An #EOSC in practice involving MS &#RIs credit coins considering #datacert #FAIR #GDPR @barendmons @gak1965 @geekeconomist @BurgelmanJean

1 1 2

Silvana Muscella
@silvanamuscella

Following

Replying to @silvanamuscella @gak1965 and 3 others

Really excited if we move this forward for interim report EOSC HLEG 👏👏👏

2:05 AM - 28 Oct 2017

Metadata for Interoperability

The year is 2045, and my grandchildren (as yet unborn) … find a letter dated 1995 and a CD-ROM (compact disk). The letter claims that the disk contains a document that provides the key to obtaining my fortune (as yet unearned). My grandchildren are understandably excited, but they have never seen a CD before - except in old movies - and even if they can somehow find a suitable disk drive, how will they run the software necessary to interpret the information on the disk? How can they read my obsolete digital document?

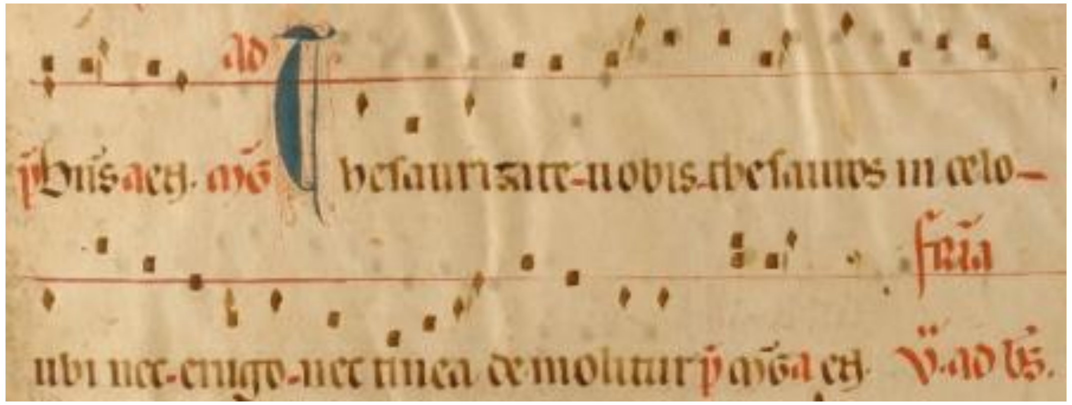
Jeff Rosenberg 1999

Data Loss in the absence of metadata

- SEER data
  - 1 | Male
  - 2 | Female
  - 3 | Other Hermaphrodite
  - 4 | Transsexual
  - 9 | Unknown

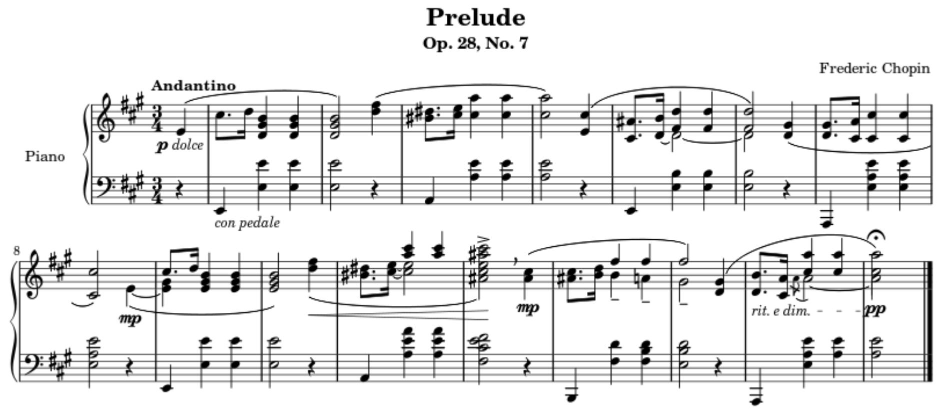
- ECOG
  - 121102 | Other sex
  - 121104 | Ambiguous sex
  - F | Female
  - FC | Female changed to male
  - FP | Female pseudohermaphrodite
  - H | Hermaphrodite
  - M | Male
  - MC | Male changed to female
  - MP | Male pseudohermaphrodite
  - O | Undetermined sex
  - U | Unknown sex

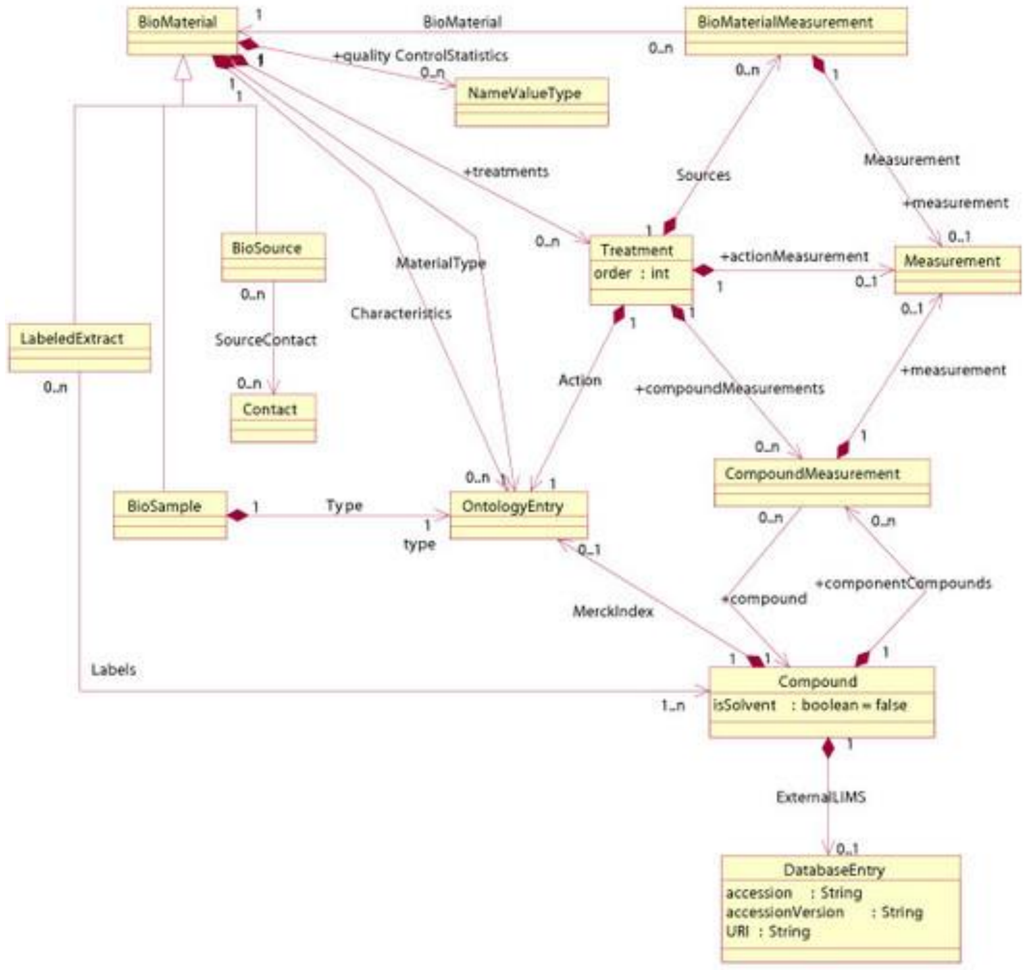
The Importance of Standards



AD MISSAM. INTROITUS
S I-TI-VIT
a-nima me-a
ad De-um fontem
vivum quando ve:
ni-am et appa-re.
bo ante fa-ci-em
Dei. Q ue-mad.
modum desiderat

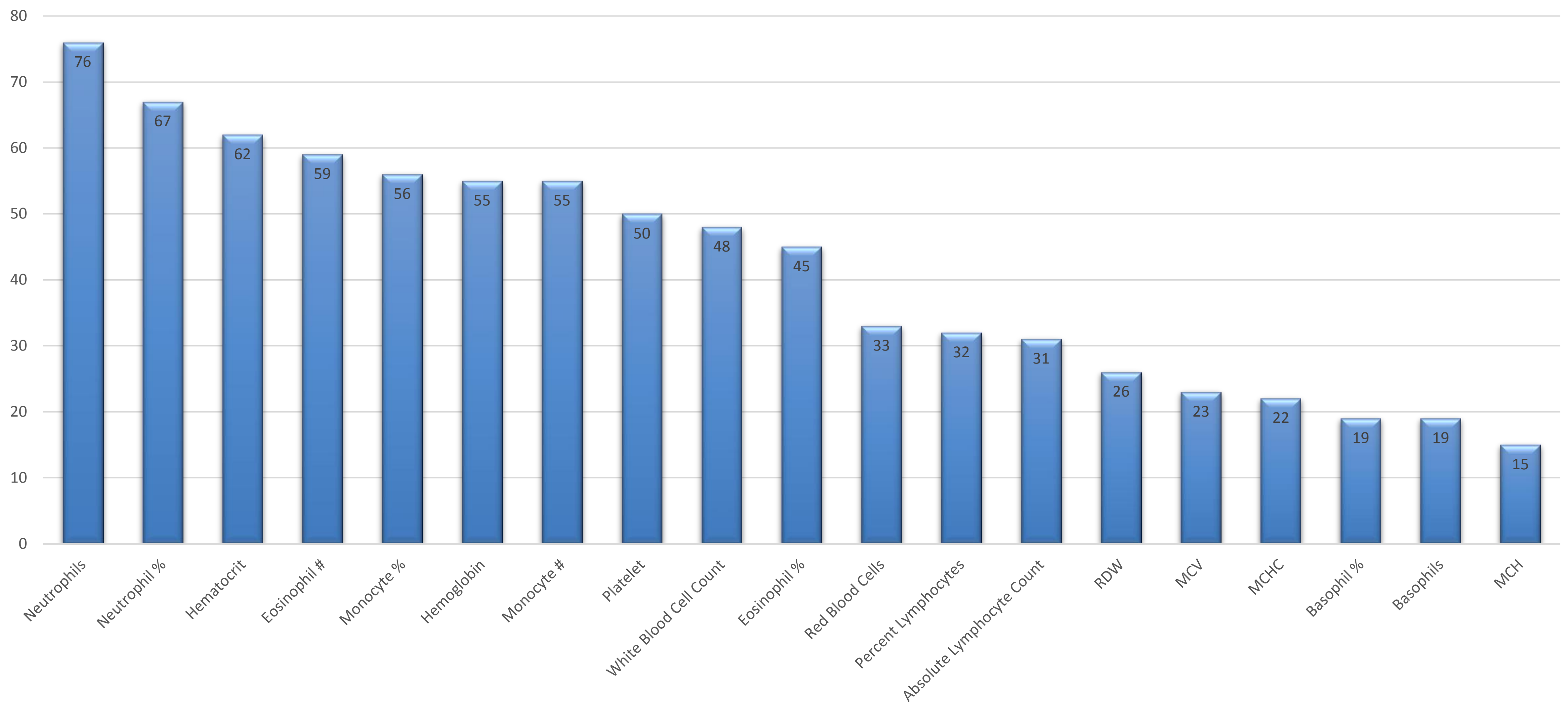
Good Standards vs. Bad Standards





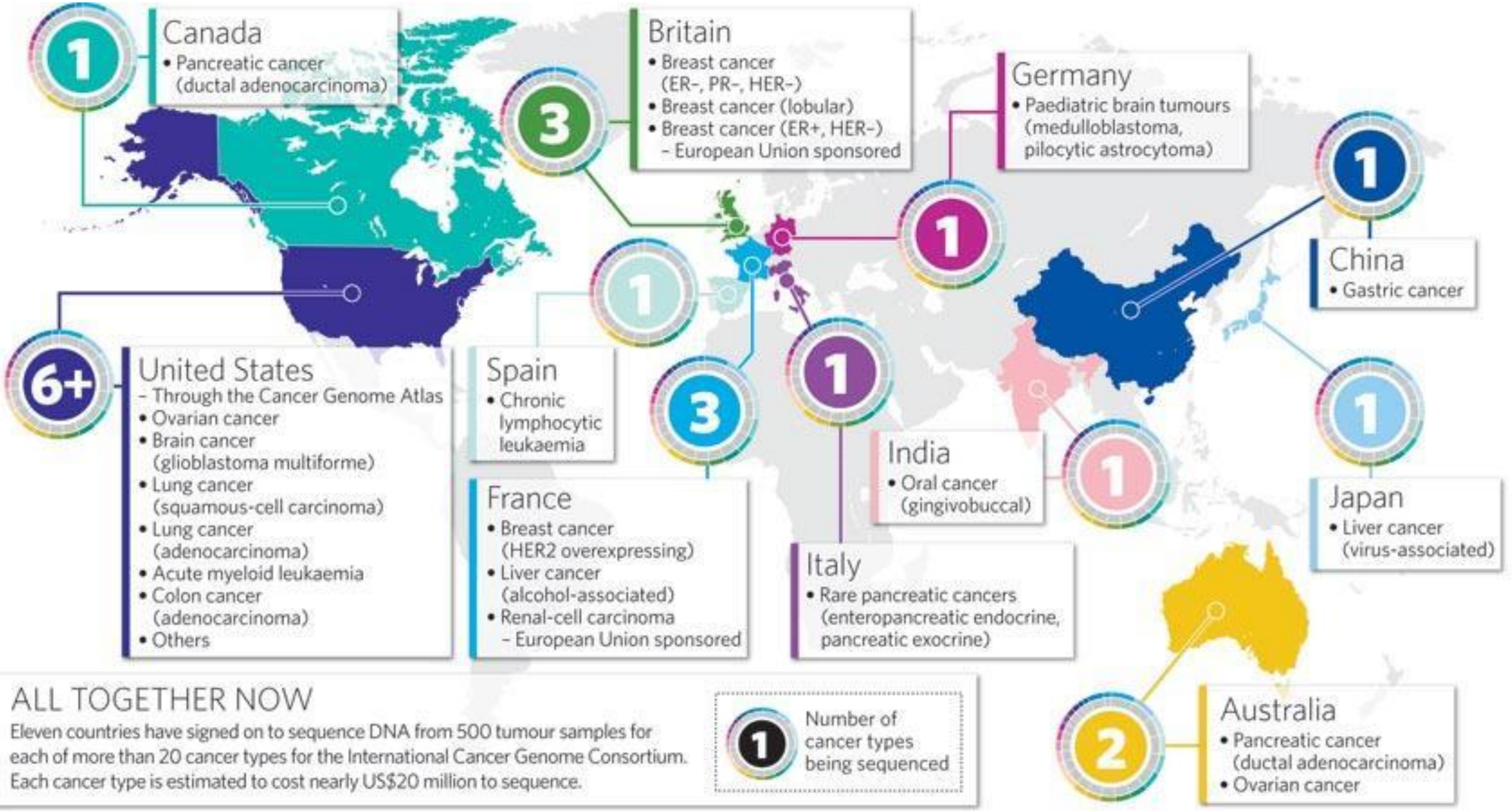
Pre- vs post-coordination

Distinct Lab Names in EMRs per Codified Lab Name
(n = 30 practices)



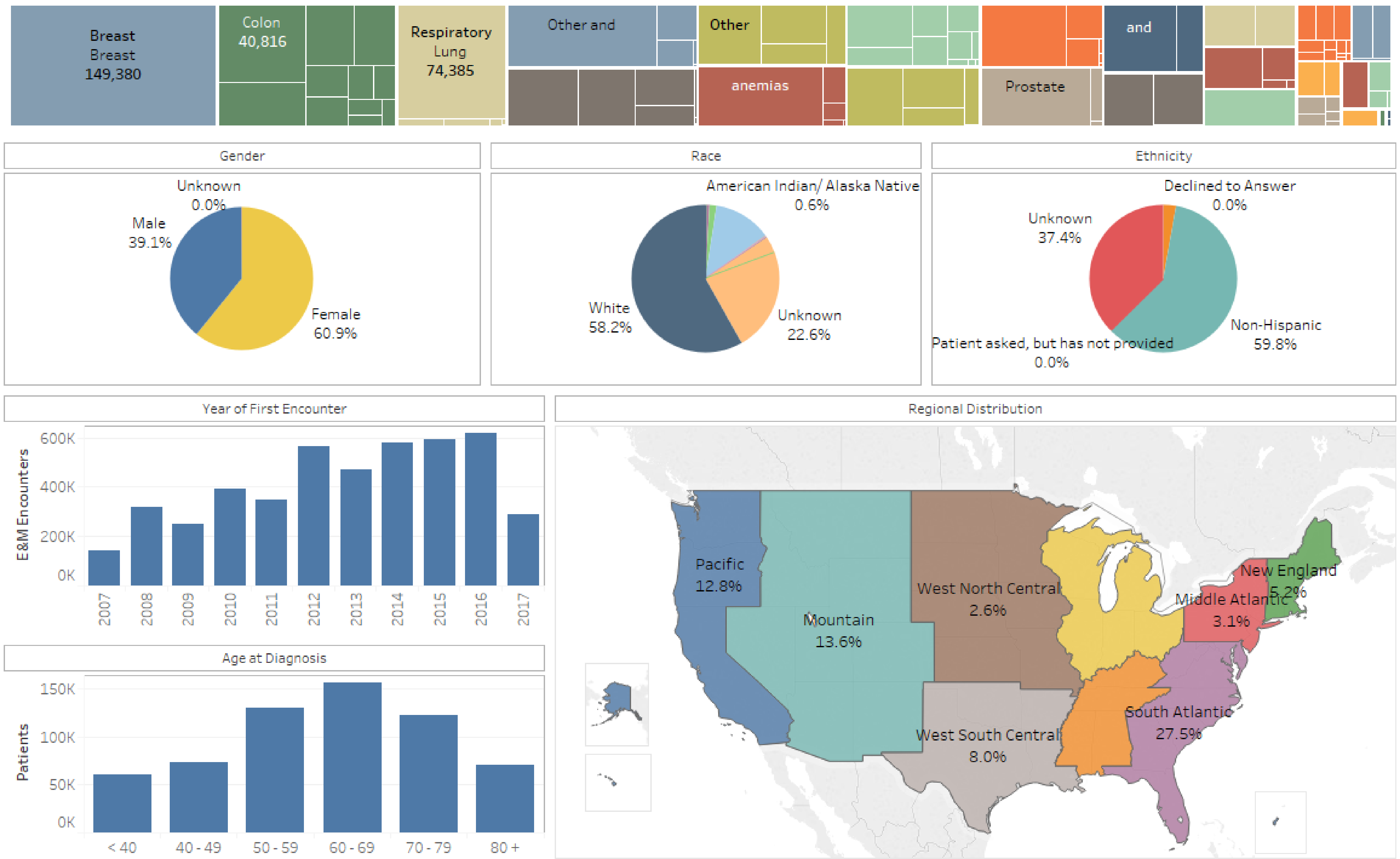
Pre- vs post-coordination

- Many drugs exist in many formulations and have many names
- A small subset of the names for paracetamol (a.k.a. acetominophin)
  - Brand names:
    - Aceta, Actimin, Anacin-3, Apacet, Aspirin Free Anacin, Atasol, Banesin, Ben-u-ron, Biogesic, Crocin, Dafalgan, Dapa, Dolo, Datril, Extra-Strength, Efferalgan, DayQuil, Depon & Depon Maximum, Feverall, Few Drops, Fibi, Fibi plus, Genapap, Genebs, Lekadol, LemSip, Liquiprin, Lupocet, Milidon, Neopap, Ny-Quil, Oraphen-PD, Panado, Panadol, Panadrex, Panamax, Paracet, Parol, Panodil, Paratabs, Paralen, Phenaphen, Plicet, PyongSu Cetamol, Redutemp, Snaplets-FR, Suppap, Tachipirina, Tamen, Tapanol, Tempra, Tipol, Tylenol, Uphamol, Valorin, Xcel
  - In other countries:
    - Acamol, Ace +, Acetalgin, Adol, Aldolor, Alvedon, APAP, Apiretal, Apiretal Flas, Atamel, Atasol, Benuron, Biogesic, Biogesic-Kiddelets, Buscapina, Calpol, Cemol, Coldex, Cotibin, Crocin, Dafalgan, Daleron, Dawa ya magi, Depon, Dexamol, Dexamol Plus, Dolex, Dolo, Dogesic, Dolipane, Dolprone, Efferalgen, Europain, Febrectal, Febrex Plus, Febricet, Febridol, Geolcatil, Gripin, Gesic Hexdex, Hedanol, Herron, Influbene, Kafa, Kitadol, Lekadol, Lupocet, Metacin, Mexalen, Milidon, Minoset, Momentum, Napa, Neo-Kiddielets, Pacimol, Pacol, Parol, Panado, Panadol, Panamax, Pand, Panodil, Para, Paracet, Paracitol, Paralen, Paramed, Paramol, Paol, Perdolan, Perfalgan, Pinex, Pyrenol, Plicet, Reliv, Rokamol, Rubophen, Sara, Scanol, Tachipirin, Tafirol, Tapson, Termalgin, Tempra, Tipol, Treuphadol, Thomapyrin, Uphamol, Vermidon, Vitamol, Xumadol, Zolben


1
Canada
• Pancreatic cancer (ductal adenocarcinoma)
3
Britain
• Breast cancer (ER-, PR-, HER-)
• Breast cancer (lobular)
• Breast cancer (ER+, HER-) - European Union sponsored
Germany
• Paediatric brain tumours (medulloblastoma, pilocytic astrocytoma)
1
1
China
• Gastric cancer
1
6+
United States
- Through the Cancer Genome Atlas
• Ovarian cancer
• Brain cancer (glioblastoma multiforme)
• Lung cancer (squamous-cell carcinoma)
• Lung cancer (adenocarcinoma)
• Acute myeloid leukaemia
• Colon cancer (adenocarcinoma)
• Others
Spain
• Chronic lymphocytic leukaemia
3
1
France
• Breast cancer (HER2 overexpressing)
• Liver cancer (alcohol-associated)
• Renal-cell carcinoma - European Union sponsored
India
• Oral cancer (gingivobuccal)
1
1
Japan
• Liver cancer (virus-associated)
Italy
• Rare pancreatic cancers (enteropancreatic endocrine, pancreatic exocrine)
ALL TOGETHER NOW
Eleven countries have signed on to sequence DNA from 500 tumour samples for each of more than 20 cancer types for the International Cancer Genome Consortium. Each cancer type is estimated to cost nearly US$20 million to sequence.
1
Number of cancer types being sequenced
2
Australia
• Pancreatic cancer (ductal adenocarcinoma)
• Ovarian cancer

CancerLinQ Patient Demographics

Patient Count: 710,754 Practice Count: 38 E&M Encounters: 4,900,216



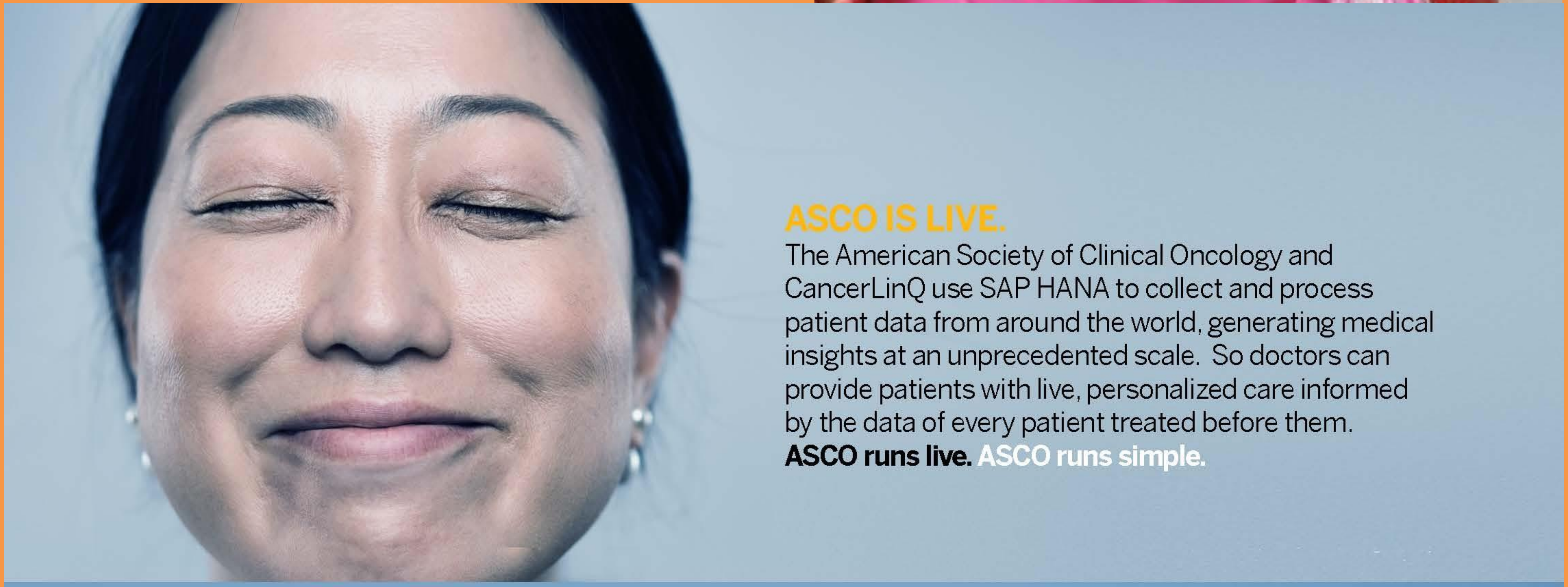
DISEASES
- ✓ Benign Hematology
- ✓ Cancer

SYSTEMS
- ✓ Aplastic anemias
- ✓ Bone and Connective ..
- ✓ Brain/CNS
- ✓ Breast
- ✓ Coagulation Defects
- ✓ Digestive
- ✓ Female Genital
- ✓ Head and Neck
- ✓ Hemolytic anemias
- ✓ Illdefined
- ✓ Leukemia
- ✓ Lymphoma
- ✓ Male Genital
- ✓ Melanoma/Skin
- ✓ Mesothelial/Soft Tiss..
- ✓ Multiple Myeloma
- ✓ Neuroendocrine
- ✓ Nutritional anemias
- ✓ Occular
- ✓ Other disorders of bl..
- ✓ Other Specified Type
- ✓ Respiratory
- ✓ Thyroid/Endocrine
- ✓ Unspecified

ANATOMIC SITE
- ✓ Acquired hemolytic a..
- ✓ Acquired pure red cel..
- ✓ Acute posthemorrha..
- ✓ Anemia in chronic dis..
- ✓ Appendix
- ✓ Bladder
- ✓ Bone
- ✓ Brain/CNS
- ✓ Breast
- ✓ Carcinoid Tumor
- ✓ Cervix

CancerLinQ

Committed to Improving Lives in the Fight Against Cancer



ASCO IS LIVE.

The American Society of Clinical Oncology and CancerLinQ use SAP HANA to collect and process patient data from around the world, generating medical insights at an unprecedented scale. So doctors can provide patients with live, personalized care informed by the data of every patient treated before them.

ASCO runs live. ASCO runs simple.