Clinical Data Warehouse at the HEGP hospital

Quality of medical data

Bastien Rance

Hôpital Européen Georges Pompidou

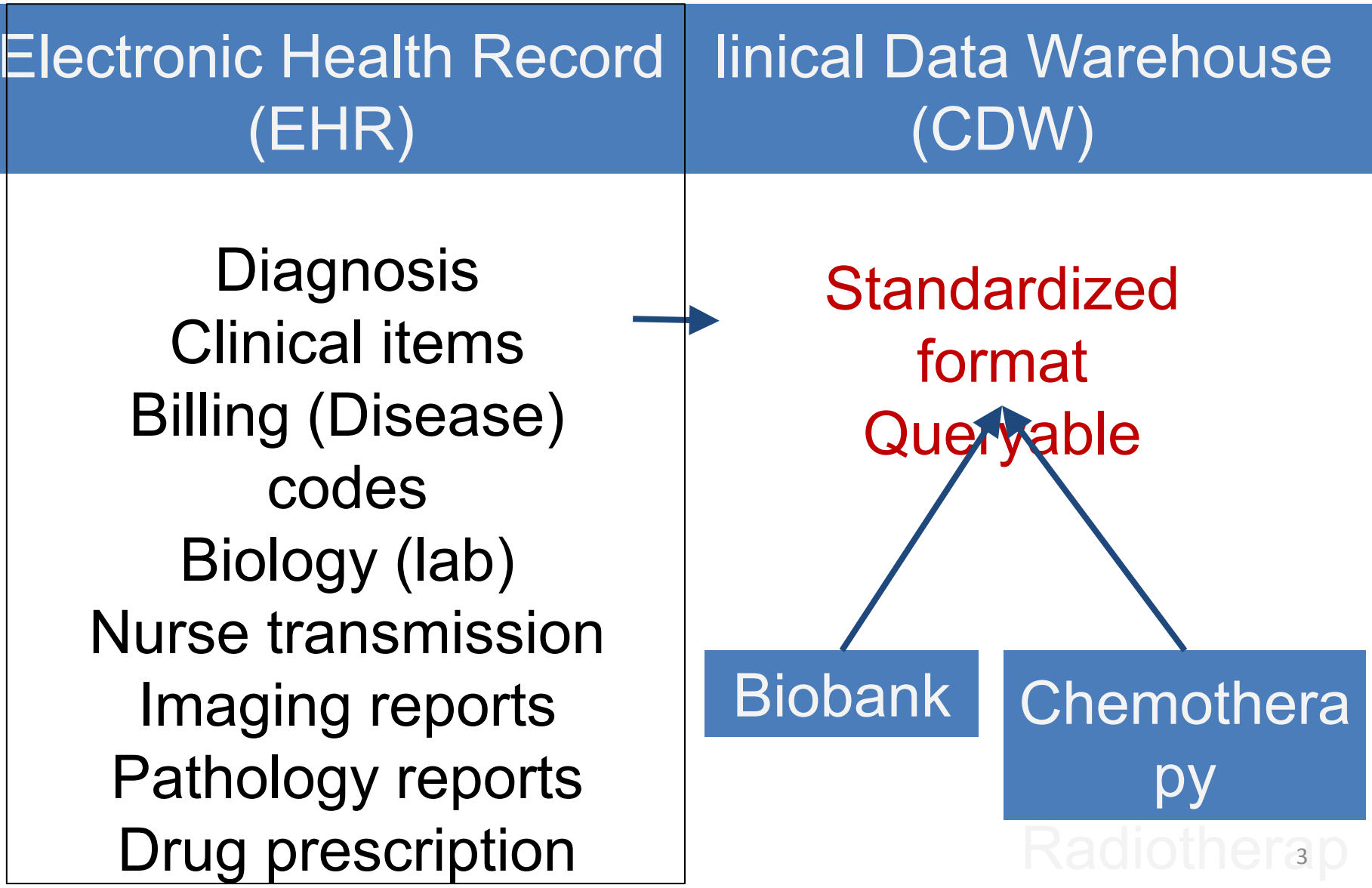
Opened in
2000

700 beds

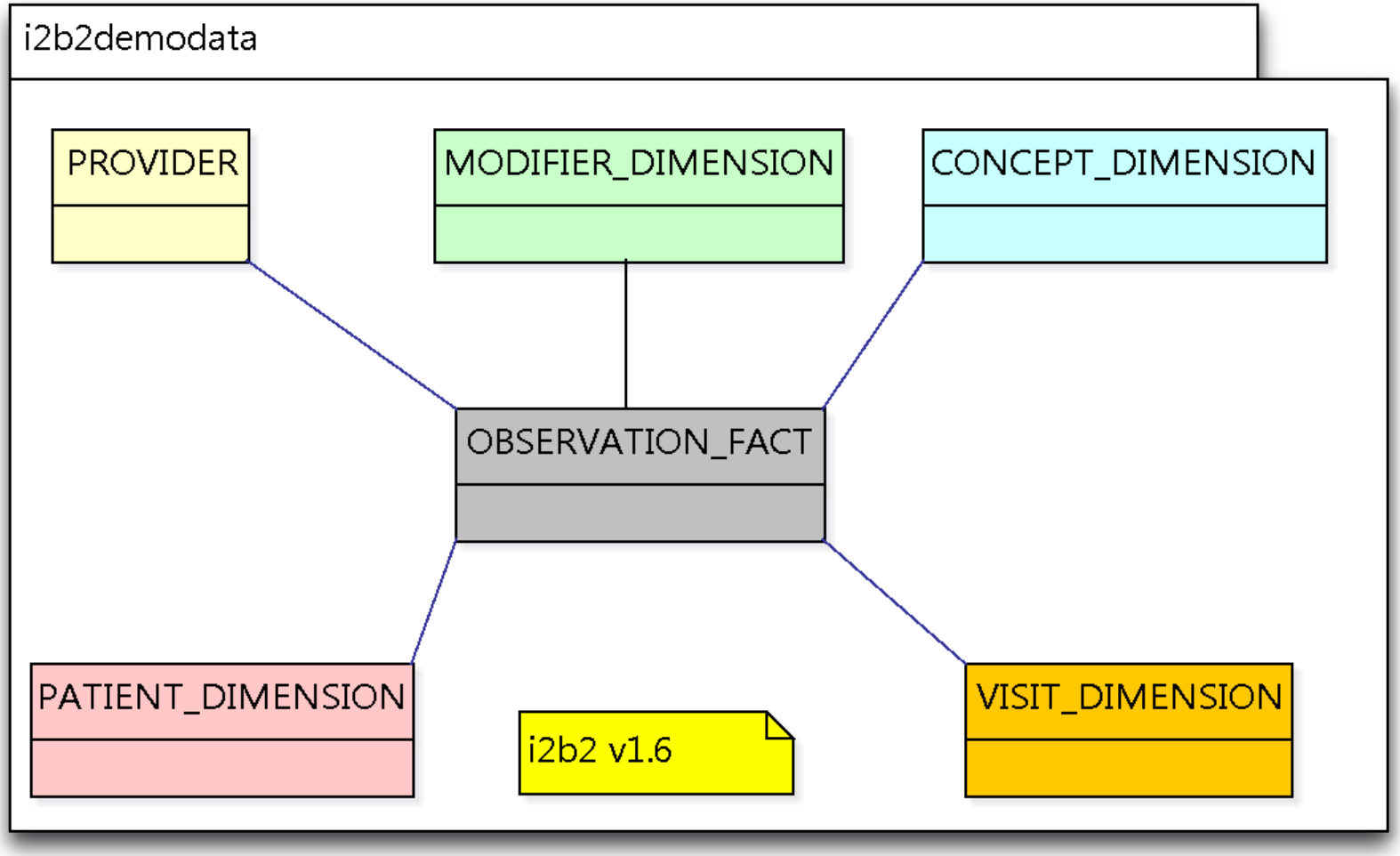
Speciality:
- **Oncology**
- **Cardiovascular diseases**
- **Emergency medicine**

HIMMS level 6
(http://www.himss.eu/node/1116)

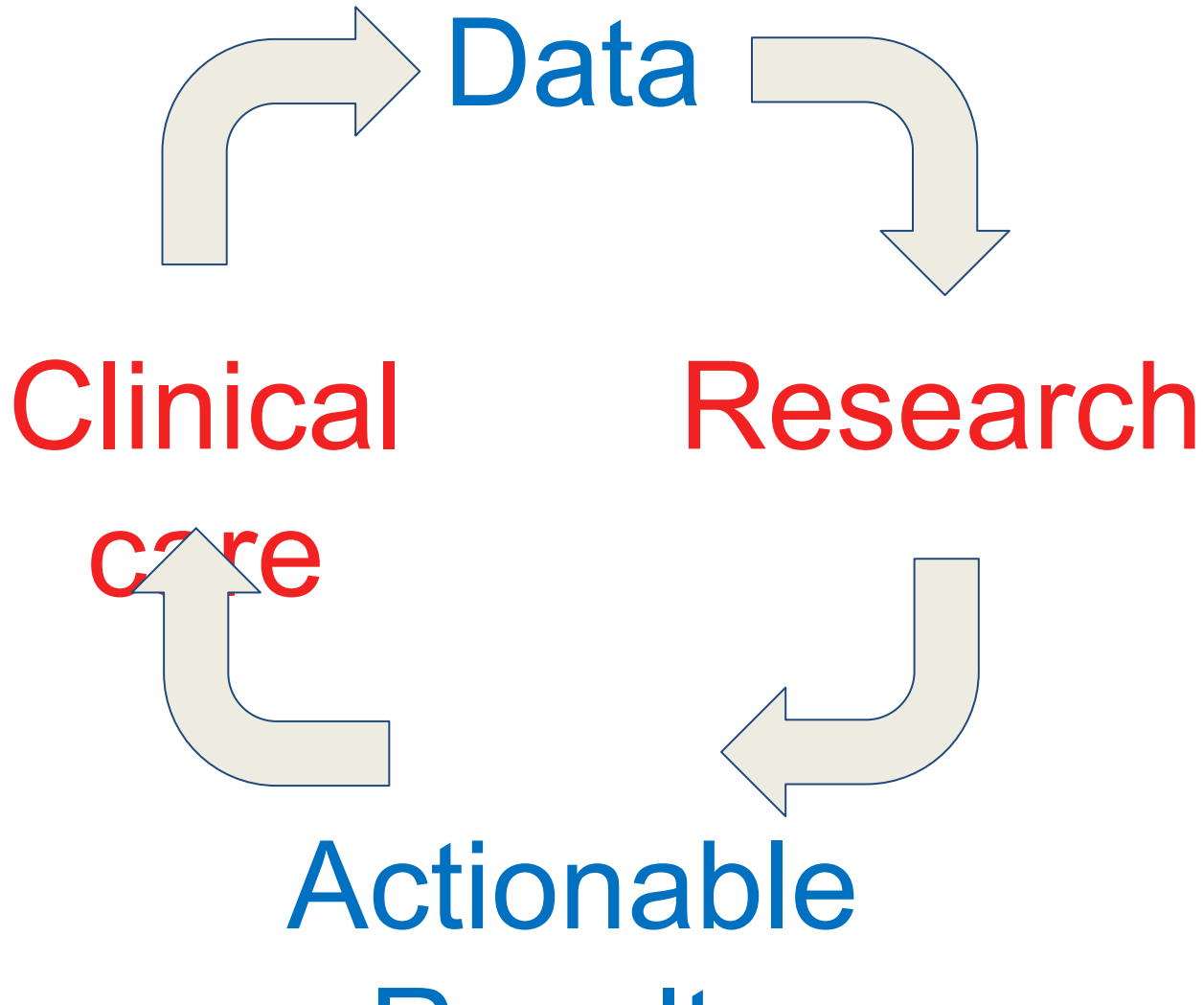
Clinical Data Warehouse



i2b2 – Informatics for integrating the biology and the bedside



Translational Research



Types of data

Healthcare data

Real life
Used for clinical decision making

Collected during the patient stay by the clinician, the intern, the nurse…

Repeated over time

Integrated in the Clinical Data Warehouse

Privacy issue+++

Clinical Research data

Controlled population
Used for clinical studies

Collected during the patient stay or dedicated meeting by the clinician, the nurse, clinical research technician

Controlled by Clinical Research Assistants

Controlled by Data Managers

Controlled by Statisticians

Often published

Secondary Use of Healthcare Data

A few success stories worldwide

Evidence-Based Medicine in the EMR Era

J.Frankovich, C.A. Longhurst, S M. Sutherland.

Stanford, NEJM Nov. 9th, 2011

Results of Electronic Search of Patient Medical Records (for a Cohort of 98 Pediatric Patients with Lupus) Focused on Risk Factors for Thrombosis Relevant to Our 13-Year-Old Patient with Systemic Lupus Erythematosus.*

Outcome or Risk Factor	Keywords Used to Conduct Expedited Electronic Search	Prevalence of Thrombosis	Relative Risk (95% CI)
		no./total no (%)	
Outcome — thrombosis	"Thrombus," "Thrombosis," "Blood clot"	10/98 (10)	Not applicable
Thrombosis risk factor			
Heavy proteinuria (>2.5 g per deciliter)			
Present at any time	"Nephrosis," "Nephrotic," "Proteinuria"	8/36 (22)	7.8 (1.7–50)
Present >60 days	"Urine protein"	7/23 (30)	14.7 (3.3–96)
Pancreatitis	"Pancreatitis," "Lipase"	5/8 (63)	11.8 (3.8–27)
Antiphospholipid antibodies	"Aspirin"	6/51 (12)	1.0 (0.3–3.7)

Evidence-Based Medicine in the EMR Era

J.Frankovich, C.A. Longhurst, S M. Sutherland.

Stanford, NEJM Nov. 9th, 2011

Results of Electronic Search of Patient Medical Records (for a Cohort of 98 Pediatric Patients with Lupus) Focused on Risk Factors for Thrombosis Relevant to Our 13-Year-Old Patient with Systemic Lupus Erythematosus.*

Outcome or Risk Factor	Keywords Used to Conduct Expedited Electronic Search	Prevalence of Thrombosis — no./total no. (%)	Relative Risk (95% CI)
Outcome — thrombosis	"Thrombus," "Thrombosis,"	10/98 (10)	Not applicable
Thrombosis risk factor			
Heavy proteinuria (>2.5 g per deciliter)			
Present at any time	"Nephrosis," "Nephrotic," "Proteinuria"	8/36 (22)	7.8 (1.7–50)
Present >60 days	"Urine protein"	7/23 (30)	4.7 (3.3–96)
Pancreatitis	"Pancreatitis," "Lipase"	5/8 (63)	11.8 (3.8–27)
Antiphospholipid antibodies	"Aspirin"	6/51 (12)	1.0 (0.3–3.7)

"we made the decision on the basis of the best data available"

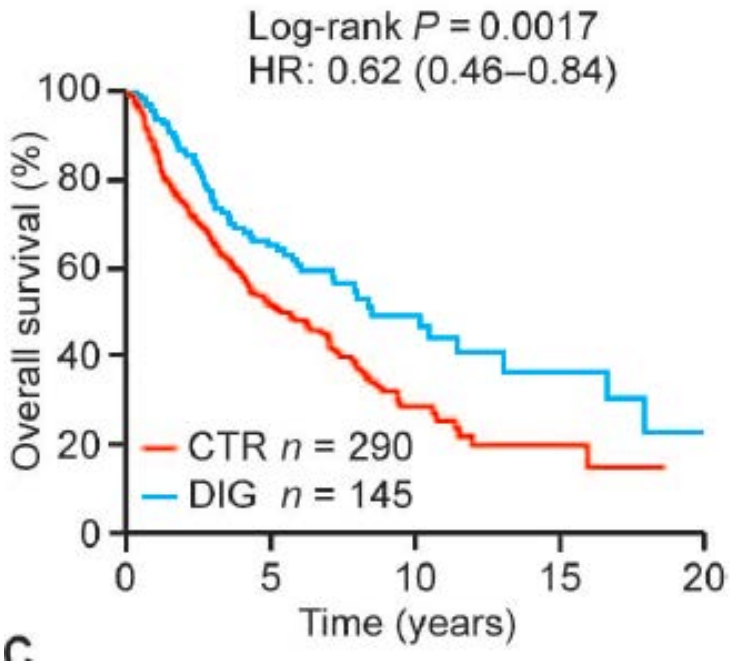
"in the light of experience as guided by intelligence."

CANCER

Cardiac Glycosides Exert Anticancer Effects by Inducing Immunogenic Cell Death

Laurie Menger,[1,2,3] Erika Vacchelli,[1,2,3] Sandy Adjemian,[1,2,3] Isabelle Martins,[1,2,3] Yuting Ma,[1,2,3] Shensi Shen,[1,2,3] Takahiro Yamazaki,[2,3,4] Abdul Qader Sukkurwala,[1,2,3] Mickaël Michaud,[1,2,3] Grégoire Mignot,[5,6] Frederic Schlemmer,[1,2,3] Eric Sulpice,[7] Clara Locher,[2,3,4] Xavier Gidrol,[7] François Ghiringhelli,[5,6] Nazanine Modjtahedi,[1,2,3] Lorenzo Galluzzi,[2,8] Fabrice André,[2,9] Laurence Zitvogel,[2,3,4] Oliver Kepp,[1,2,3]*† Guido Kroemer[1,8,10,11,12]*†

Cell > Mice > **Retrospective data**



Text-based research algorithm to identify all carcinoma patients who received **digitalin** during conventional carcinoma therapies between 1981 and 2009

Compared the overall survival of:
145 patients treated with CGs
290 patients who did not receive CGs.

Financial benefits

AMIA Annu Symp Proc. 2006; 2006: 1044.

Calculating the Benefits of a Research Patient Data Repository

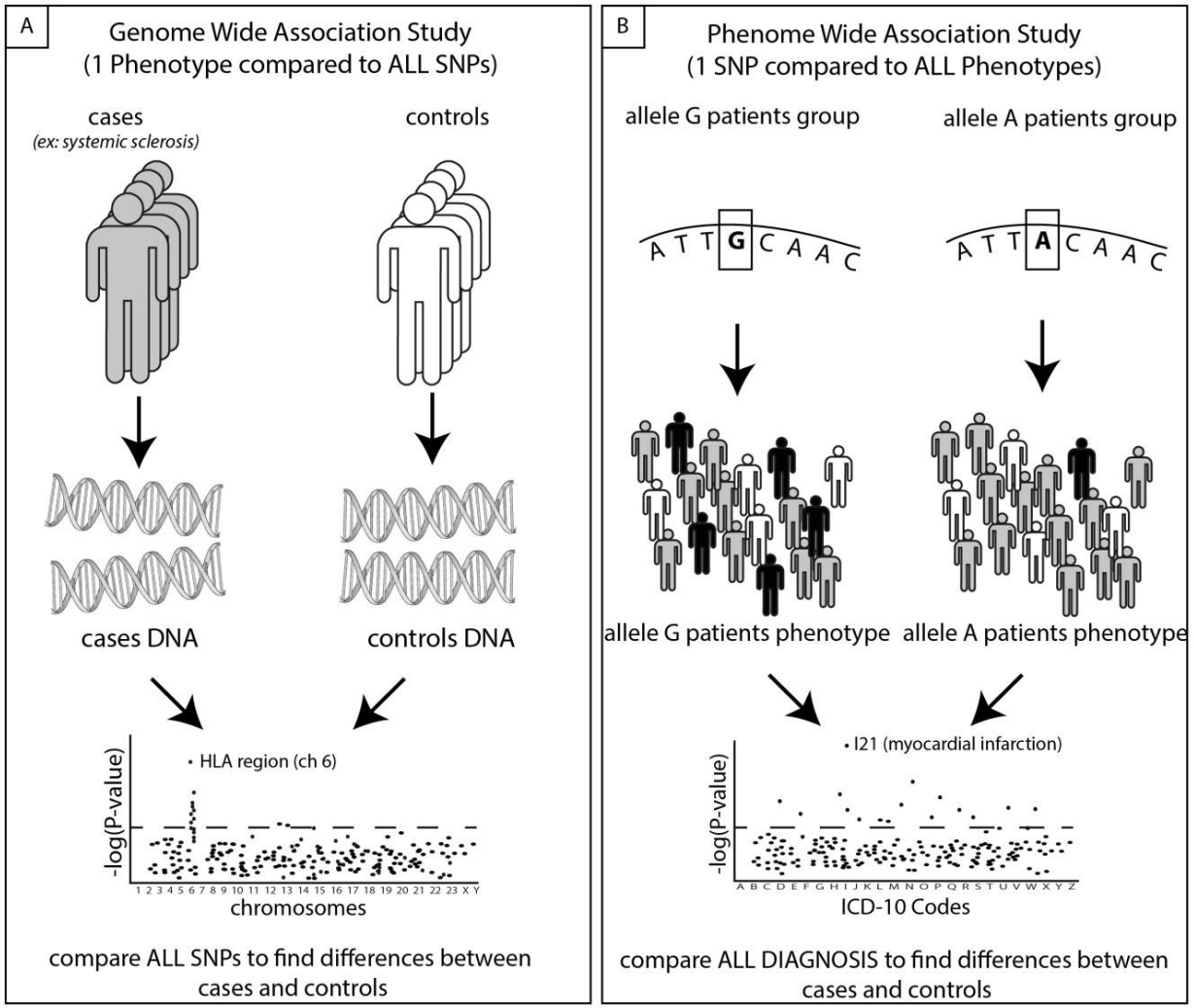
Ruth Nalichowski, MBA, Diane Keogh, Henry C. Chueh, MD, MS, and Shawn N. Murphy, MD, Ph.D.

2006 study

$7 million saved on patient recruitment

Between $94-136 million in related funding

Phenome Wide Association Studies (PheWAS)



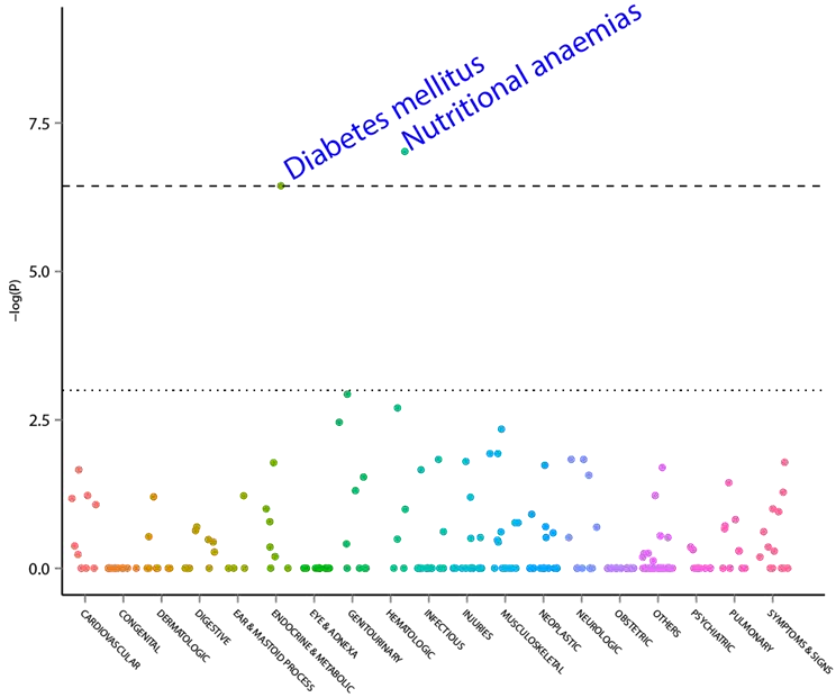
McCarthy et al, Nature Reviews Genetics, 2008 / Denny et al, Bioinformatics 2010

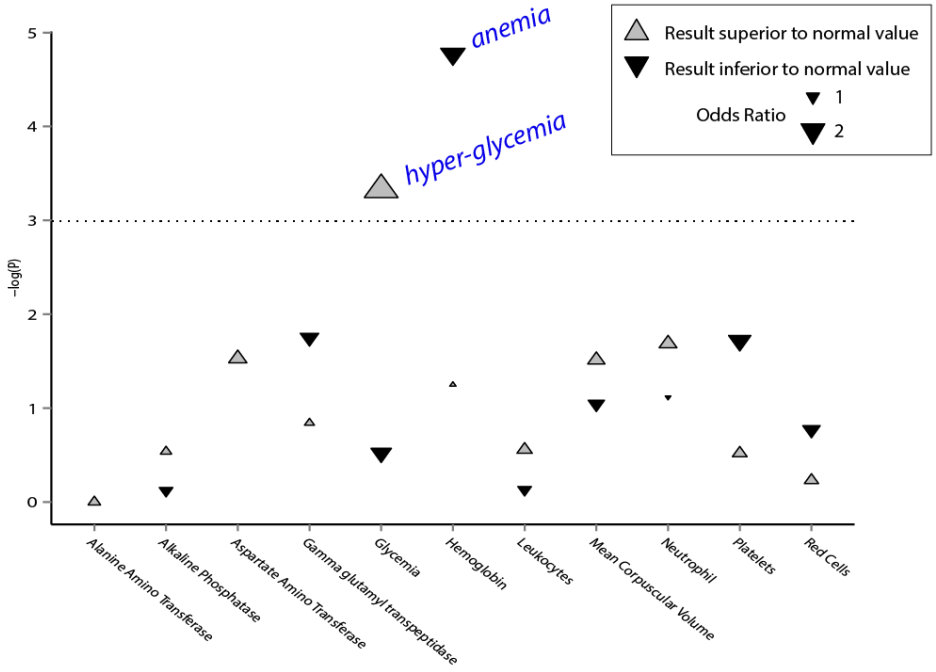
OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics

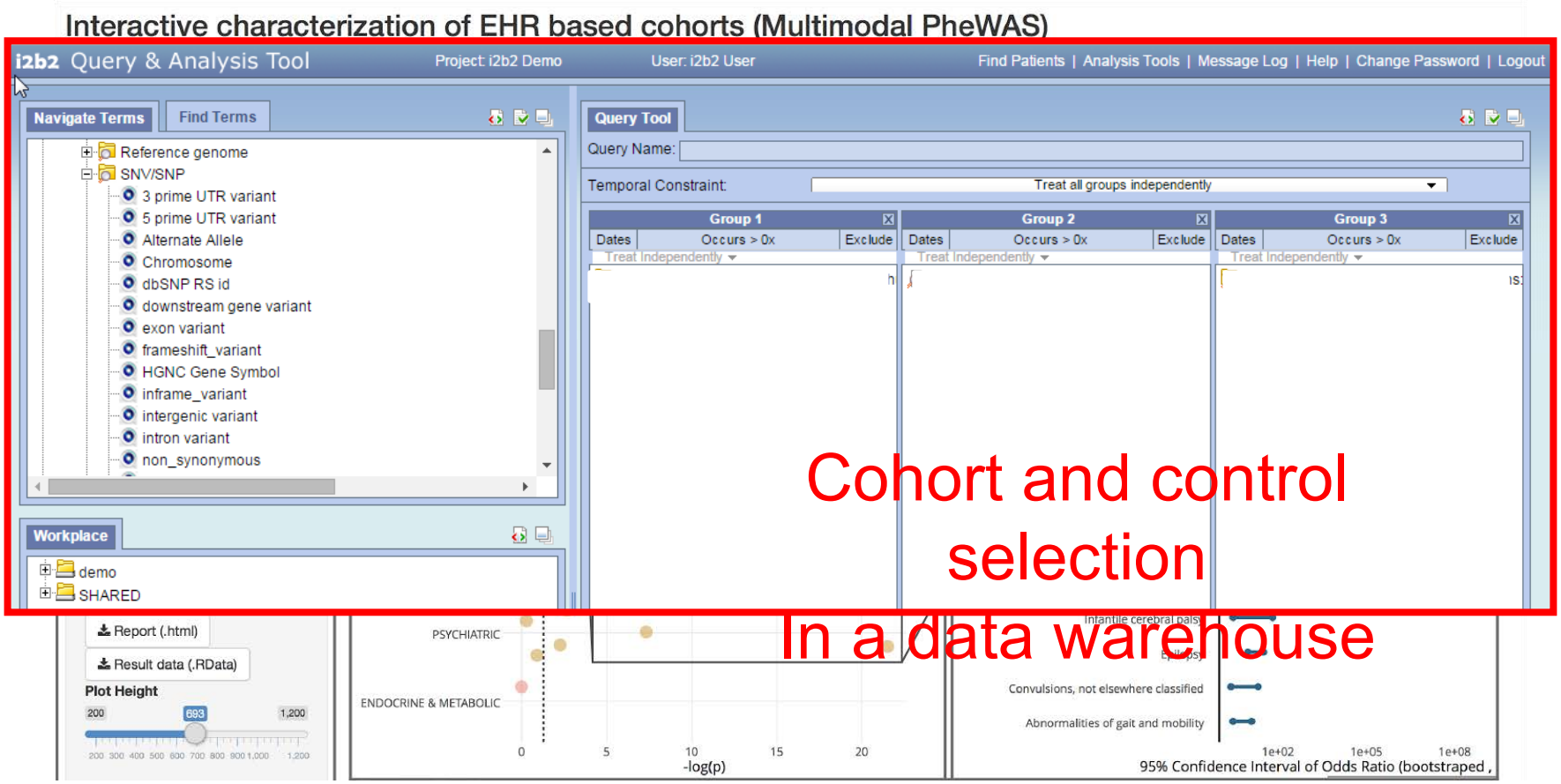
Antoine Neuraz[1,2], Laurent Chouchana[3], Georgia Malamut[4], Christine Le Beller[5], Denis Roche[6], Philippe Beaune[3,6], Patrice Degoulet[1,2], Anita Burgun[1,2], Marie-Anne Loriot[3,6], Paul Avillach[1,2]*





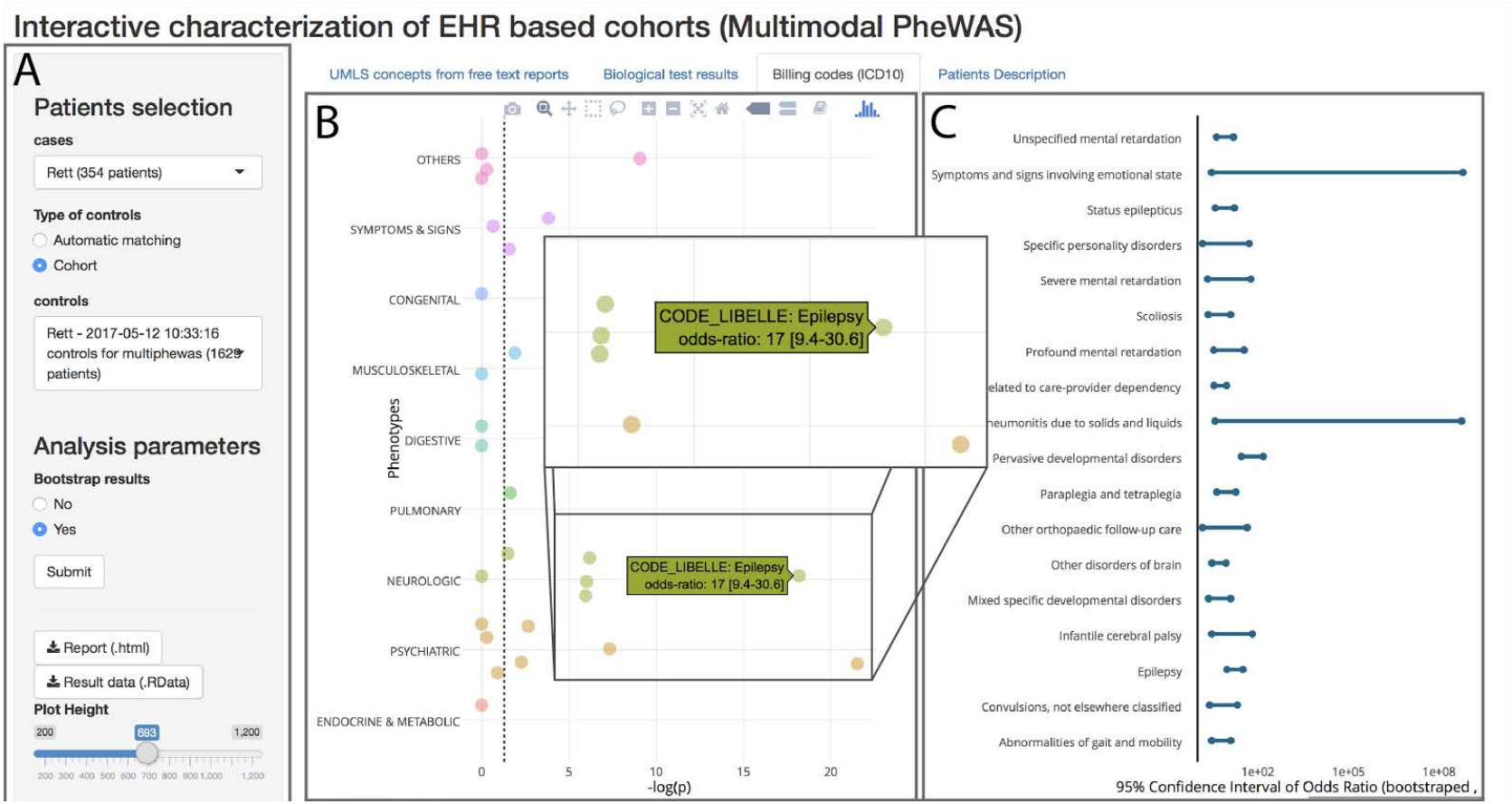
Phenotype	Low activity	Intermediate activity	Normal activity	Very High activity
Thiopurine	10 % dose	30 – 70 % dose	100 % dose	> 100 % Dose ?

PheWAS on-demand



Automated and interactive characterization of clinical data warehouses based cohorts: an open-source web application for multimodal phenome-wide association studies. Neuraz *et al*. *submitted* 2017

PheWAS on-demand



Automated and interactive characterization of clinical data warehouses based cohorts: an open-source web application for multimodal phenome-wide association studies. Neuraz *et al*. *submitted* 2017

Multi-omics analysis

NATURE COMMUNICATIONS | ARTICLE OPEN

Multi-omics analysis defines core genomic alterations in pheochromocytomas and paragangliomas

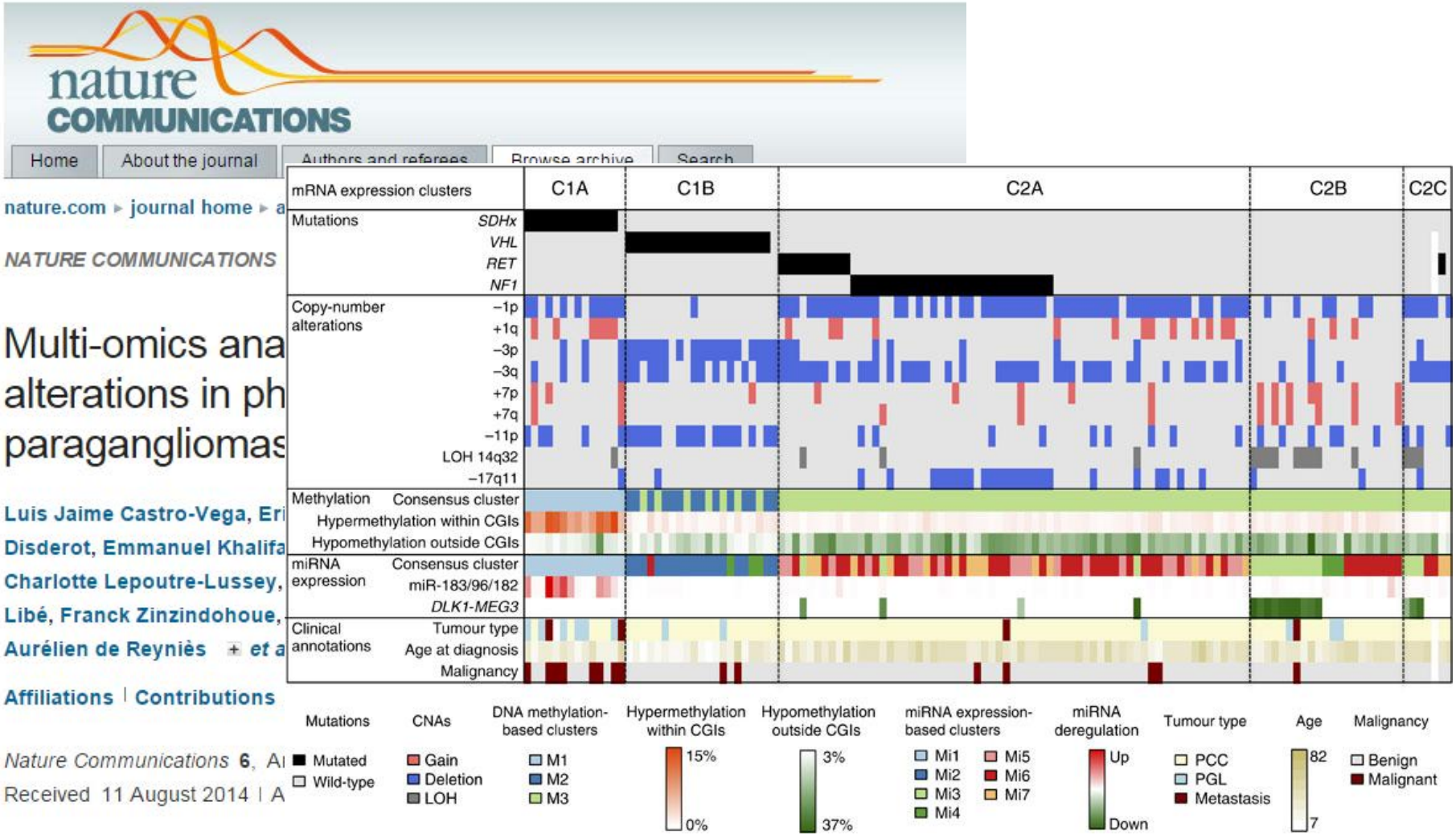
Luis Jaime Castro-Vega, Eric Letouzé, Nelly Burnichon, Alexandre Buffet, Pierre-Hélie Disderot, Emmanuel Khalifa, Céline Loriot, Nabila Elarouci, Aurélie Morin, Mélanie Menara, Charlotte Lepoutre-Lussey, Cécile Badoual, Mathilde Sibony, Bertrand Dousset, Rossella Libé, Franck Zinzindohoue, Pierre François Plouin, Jérôme Bertherat, Laurence Amar, Aurélien de Reyniès + *et al.*

Affiliations | Contributions | Corresponding author

Nature Communications **6**, Article number: 6044 | doi:10.1038/ncomms7044
Received 11 August 2014 | Accepted 05 December 2014 | Published 27 January 2015

http://www.nature.com/ncomms/2015/150127/ncomms7044/full/ncomms7044.html

Multi-omics analysis



http://www.nature.com/ncomms/2015/150127/ncomms7044/full/ncomms7044.html

CARPEM: Cancer Research Project

http://www.carpem.fr/

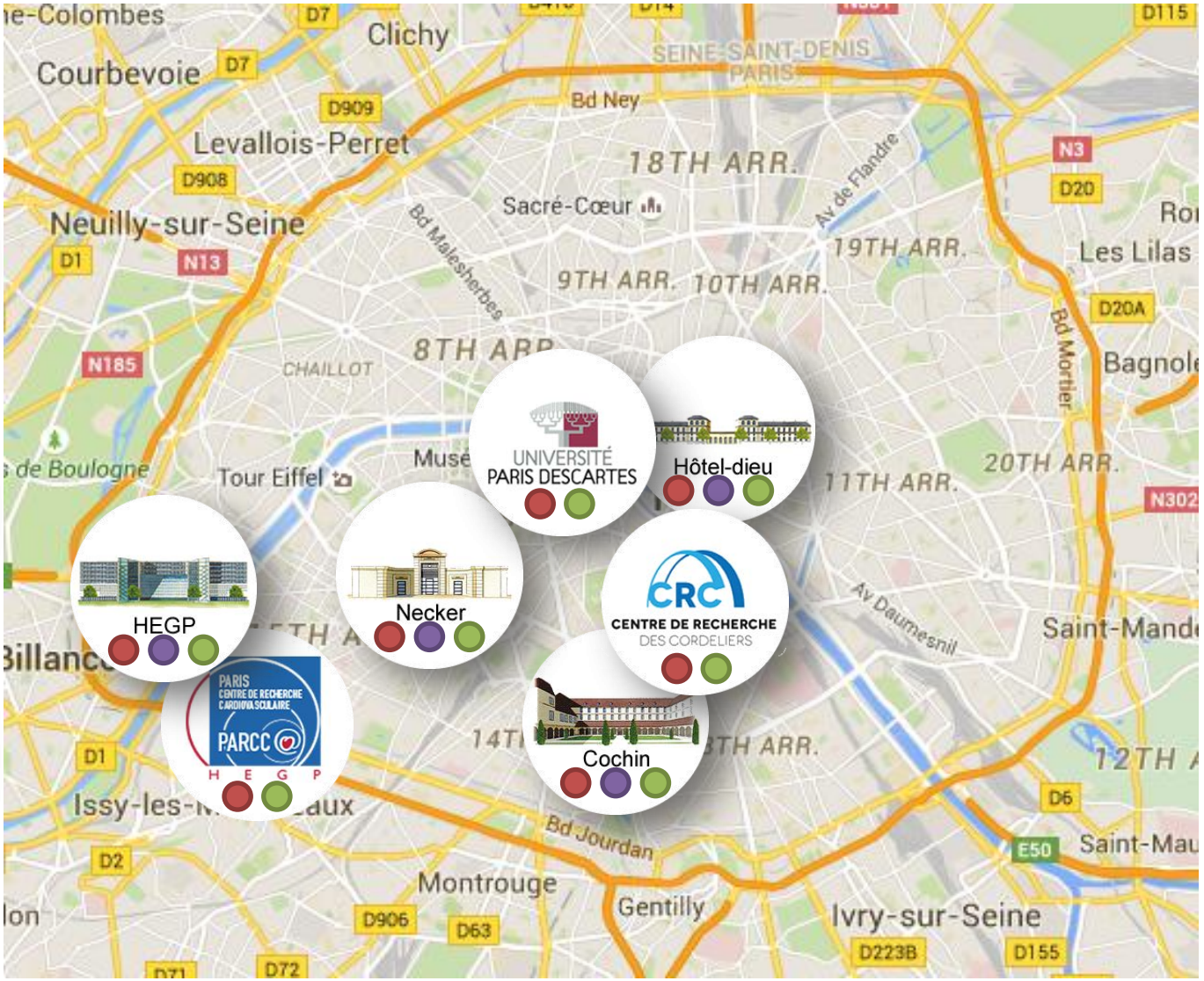
The CARPEM Program

Cancer Research and Personalized Medicine

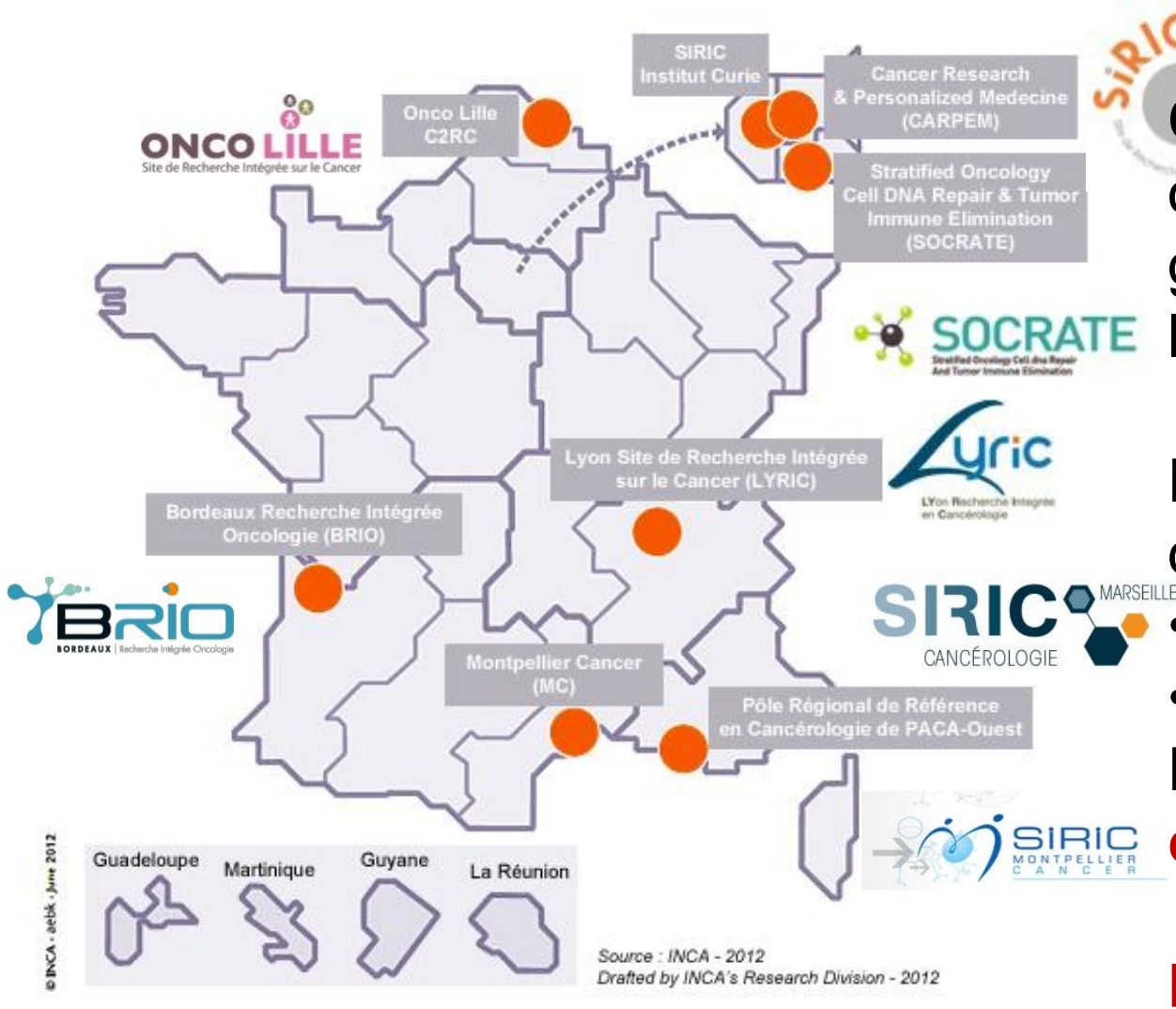
Clinical Data

Biobank Data

NGS Data



CARPEM in the French landscape



CARPEM is member of the OSIRIS working group on data-sharing lead by INCa

First objectives, define:

- 100 clinical items
- 100 omics items

For a **uniform collection** in France

Regional working group on data-

Quality of data / Quality of care / Administrative quality indicators

Haute Autorité à la Santé

Quality of data / Quality of care / Administrative quality indicators

Haute Autorité à la Santé

HAS

HAUTE AUTORITÉ DE SANTÉ

Contribuer à la régulation par la qualité et l'efficience

Admistrative : U.S. example : “meaningful use” (stage 1, 2)

	Stage 1	Stage 2
Objective	Record and chart changes in vital signs: • Height • Weight • Blood pressure • Calculate and display BMI • Plot and display growth charts for children 2-20 years, including BMI	Record and chart changes in vital signs: • Height • Weight • Blood pressure ***(age 3 and over)*** • Calculate and display BMI • Plot and display growth charts for patients ***0-20 years***, including BMI
Measure	For more than 50% of all unique patients age 2 and over seen by the EP, blood pressure, height and weight are recorded as structured data	More than ***80%*** of all unique patients seen by the EP have blood pressure ***(for patients age 3 and over only)*** and height and weight ***(for all ages)*** recorded as structured data

Imprecision & correction

Secondary use often different from the primary collection cause

E.g.
Diagnostic codes:

- ICD10 - I10 Hypertension
- ICD10 – C50 Breast cancer

Medical forms (human collection)
typo
impression of the measure
Vital signs (machine)
impression of the measure

Missing data

Health data

Open world assumption

Statistical approach to missingness:
Treatment: Insulin
Diagnostic code: [empty]

Should be Diabetes

Clinical Research data

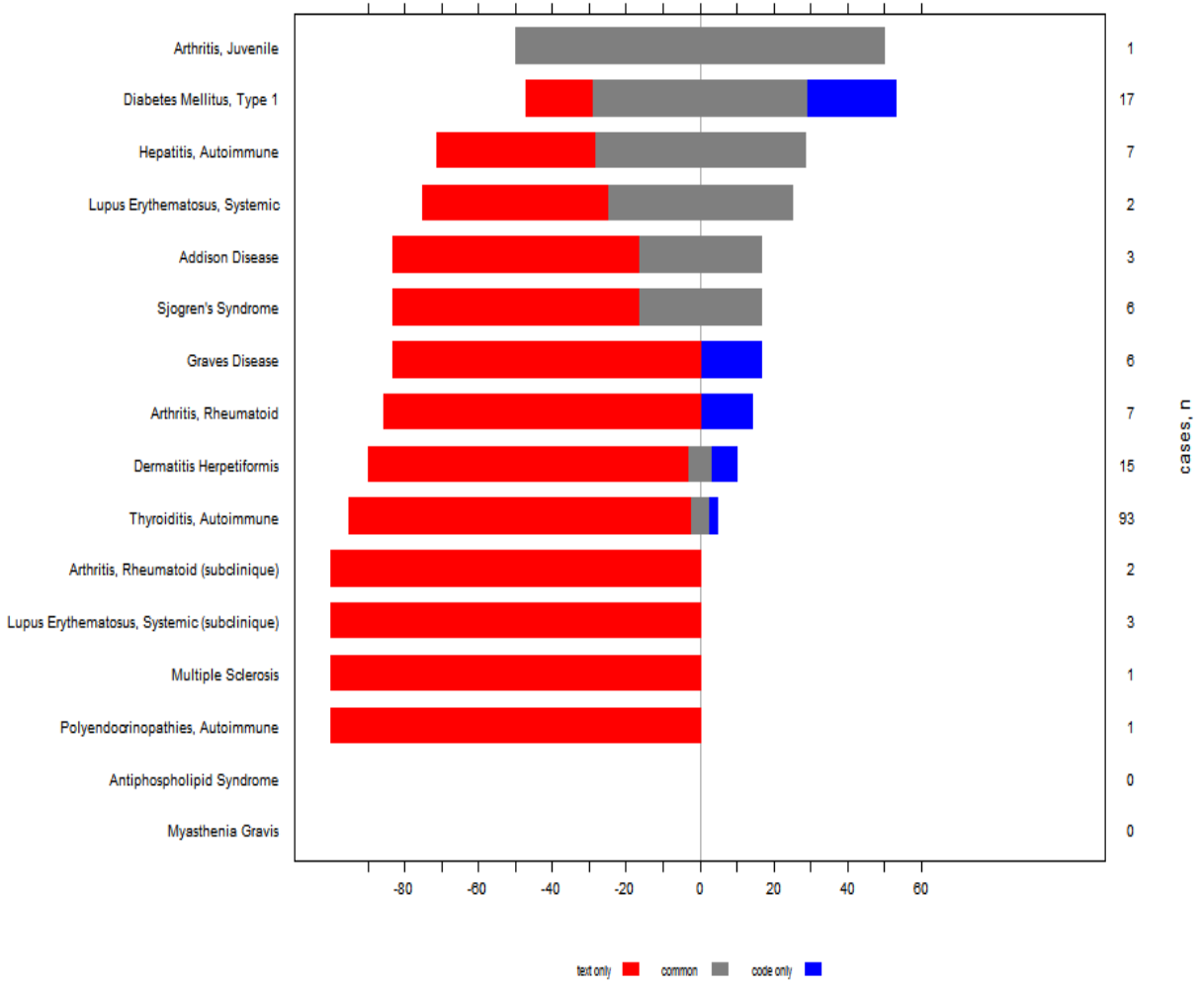
Often close world assumption

Types of missing data:
Not Applicable
Not Realized
Missing Data

The importance of dirty data

Hidden treasures

Case study: Autoimmune comorbidities of the Celiac Disease



80%

of the information is present only in free-text

(and not in structured data)

Semi-structured texts

Metastatic Renal Clear Cell Carcinoma patients

RECIST follow-up

Semi-structured text report

* Les LESIONS CIBLES sont définies de la manière suivante:

Au niveau du poumon:
- Cible 1: Nodule du lobe inférieur gauche de **14 mm** de plus grand axe.

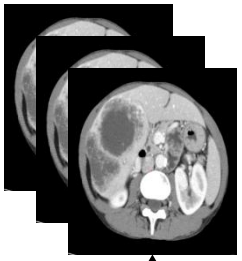
Au niveau du médiastin:
- Cible 2: Adénomégalie de la loge de Baréty de **46 mm** de plus grand axe.
- Cible 3: Adénomégalie de la fenêtre aortopulmonaire de **35 mm** de plus grand axe.

[…]

CONCLUSION

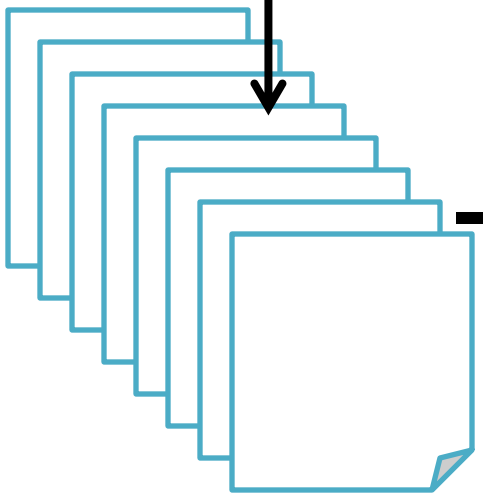
1) La somme des plus grandes longueurs pour le scanner cycle 3 est donc mesurée à **14+46+35+43+34+26 = 198 mm**. Par rapport au **scanner de référence du 21/02/2004** dont la somme est mesurée à **209 mm**, l'évolution est de **-5%**.
L'évolution des cibles mesurables est donc stable **(SD)**.
2) Absence d'évolution non-équivoque des lésions non-cibles **(SD)**.
3) Absence de nouvelle lésion non cible **(No)**.
4) La réponse globale est **(SD-SD-No)** soit **SD**.
Stabilité de l'atélectasie lobaire supérieure droite secondaire à l'obstruction quasi-complète de la bronche lobaire par l'adénopathie.

Leveraging semi-structured text



PACS
Radiology image archives

Clinical Data Warehouse



RECIST extractor

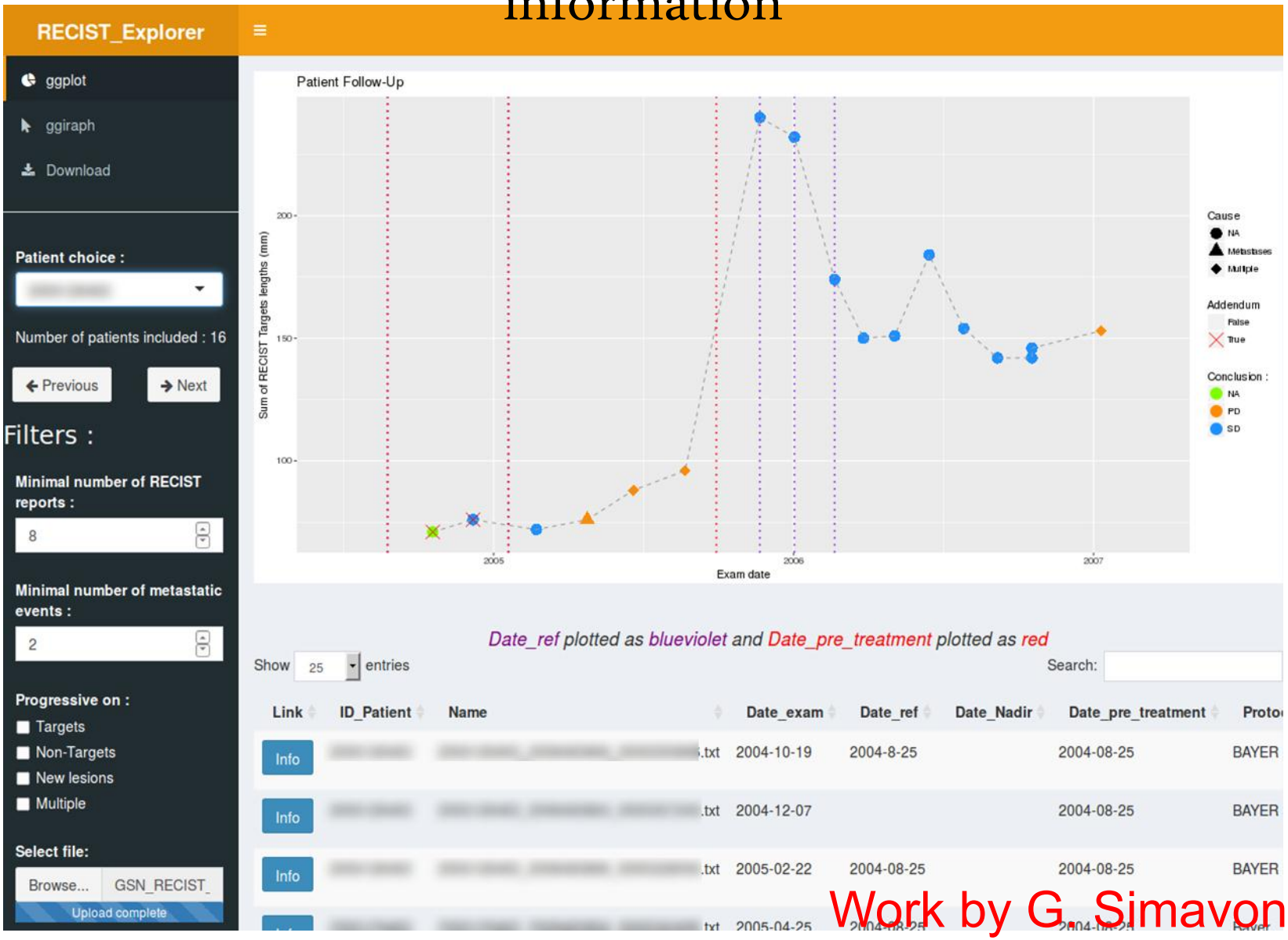
RECIST Explorer

5,000+ Semi-structured Radiology

Simple Natural Language Processing

Queriable Dynamic

RECIST Explorer –From text to structured-information

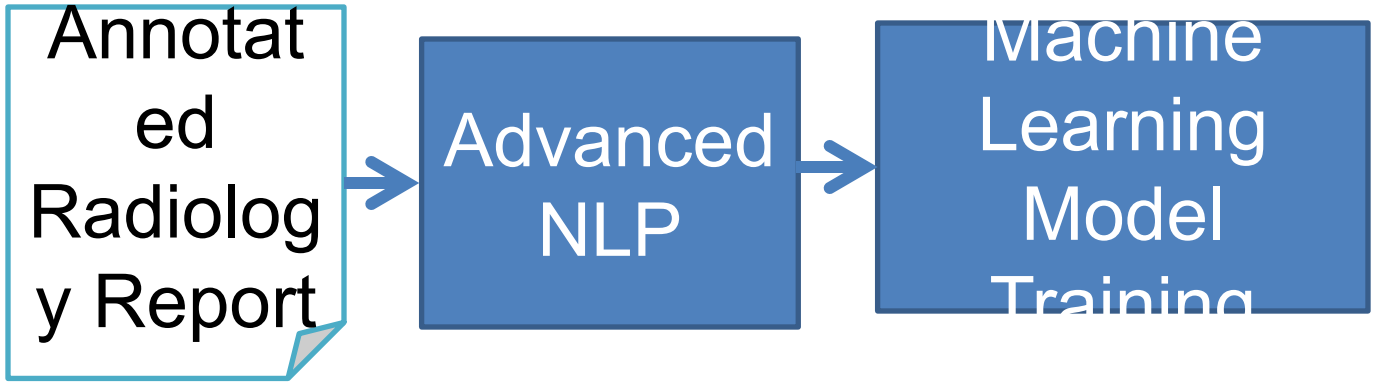


Work by G. Simavonian, MS

Mining Clinical narratives

Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings

Anne-Dominique Pham[†] ✉, Aurélie Névéol[†], Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello and Anita Burgun

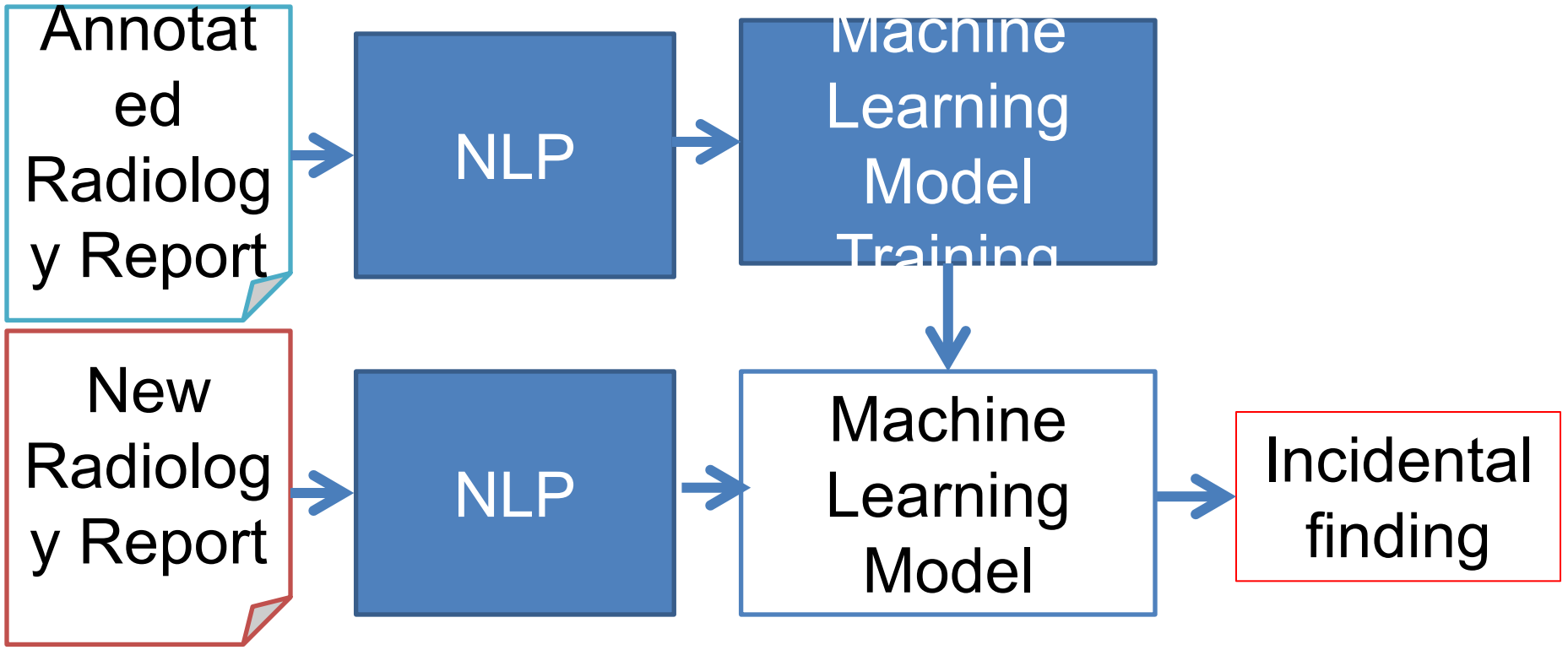


Pham *et al.* 2015 BMC Bioinfomatics

Mining Clinical narratives

Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings

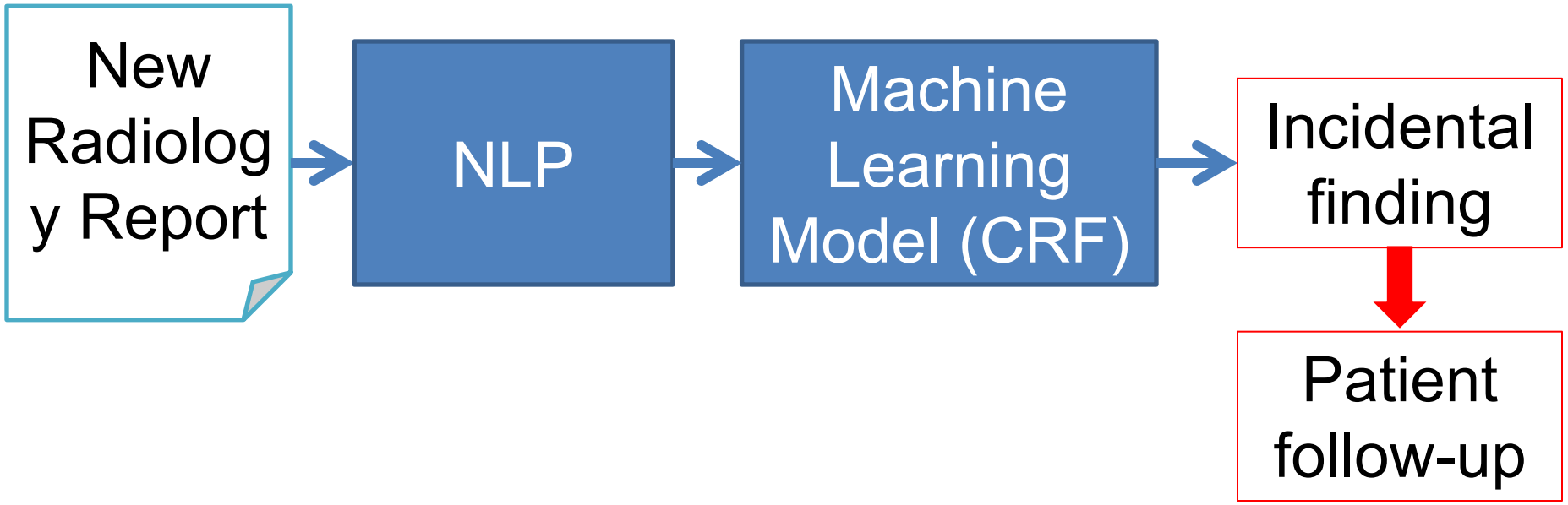
Anne-Dominique Pham[†] ✉, Aurélie Névéol[†], Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello and Anita Burgun



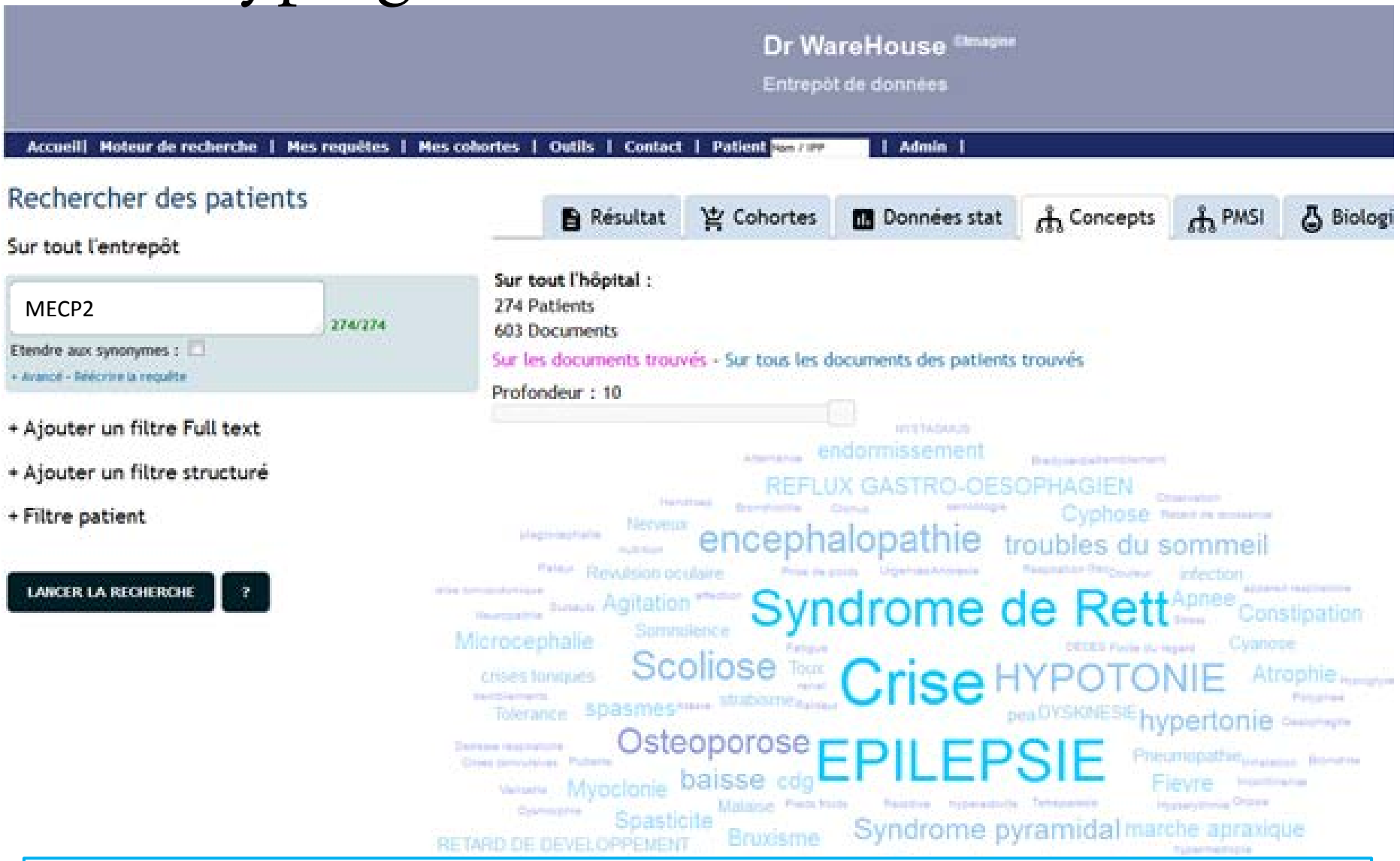
Mining Clinical narratives

Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings

Anne-Dominique Pham[†] ✉, Aurélie Névéol[†], Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello and Anita Burgun



Phenotyping

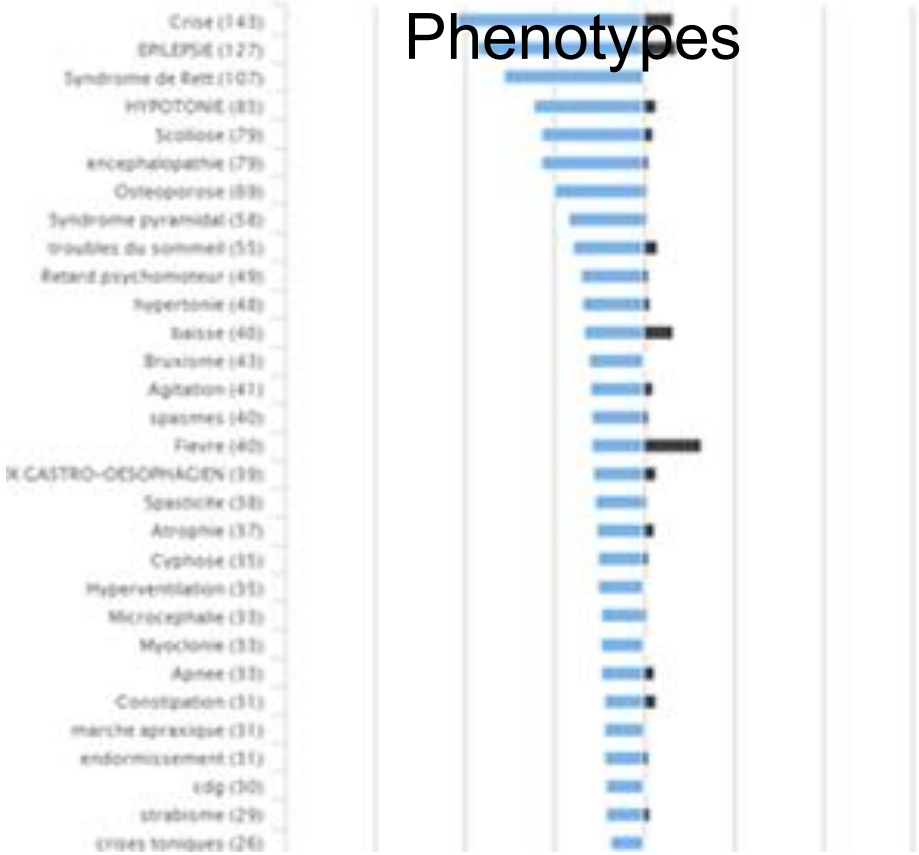


Discovering Phenotype Associations in Clinical Data Warehouse Using Free-text. Garcelon et al. *Submitted*

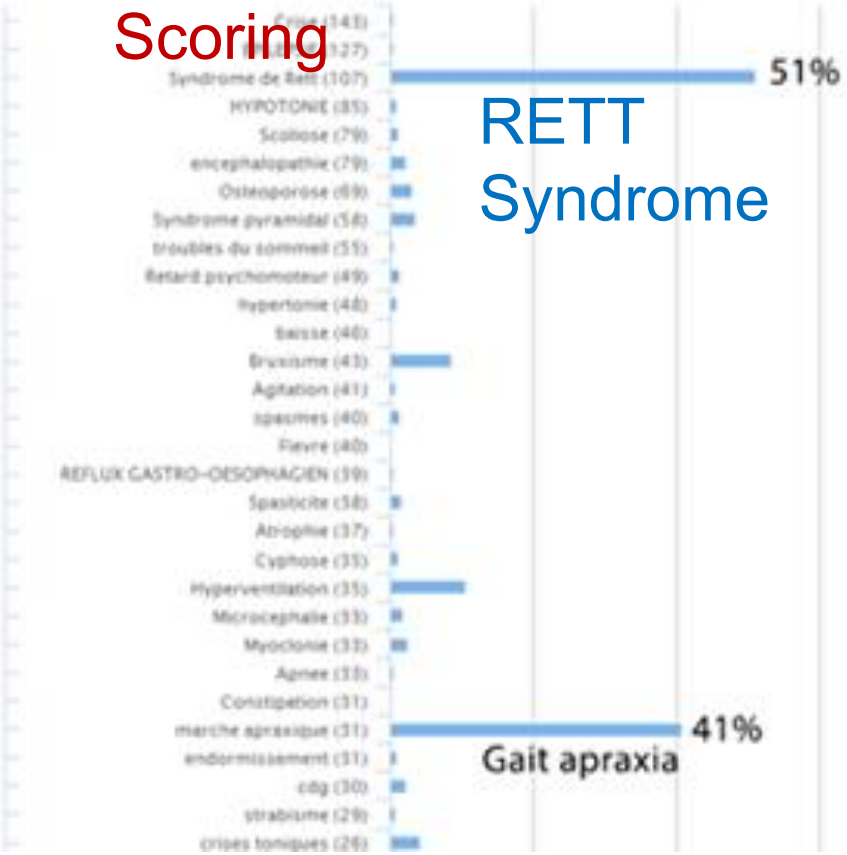
Phenotyping

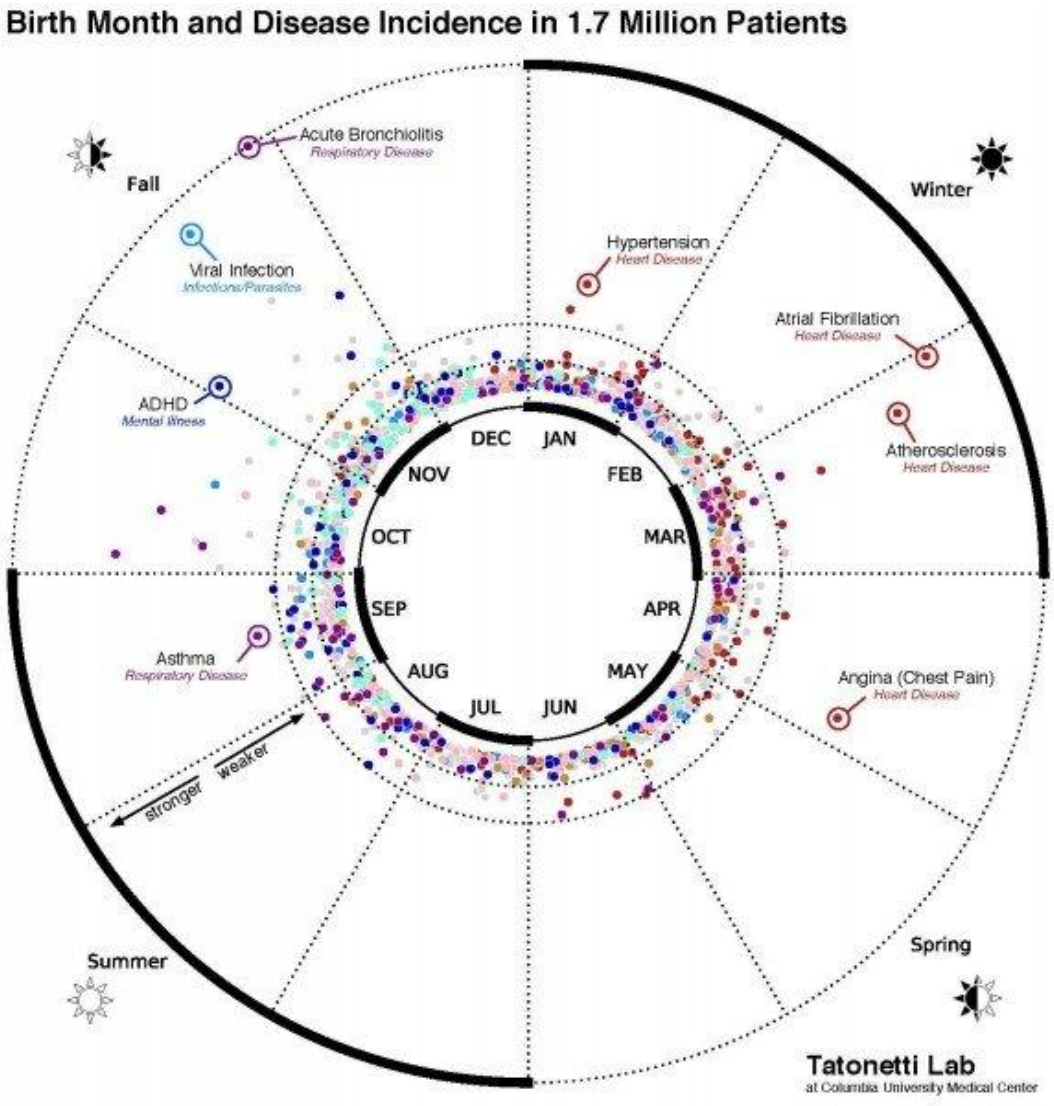
Query: MECP2

Frequently Associated Phenotypes



Phenotype Specificity Scoring





Boland MR *et al*. Birth month affects lifetime disease risk: a phenome-wide method. JAMIA 2015

Value

SCIENTIFIC DATA a natureresearch journal

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

SCIENTIFIC DATA a natureresearch journal

Altmetric: 534 Views: 42,582 Citations: 86 More detail »

Comment | OPEN

The FAIR Guiding Principles for scientific data man

Mark D. Wilkinson, Mi

INSTITUT NATIONAL DU CANCER | ACCÉLÉRONS LES PROGRÈS FACE AUX CANCERS

SUIVEZ-NOUS

INSTITUT NATIONAL DU CANCER | PLAN CANCER | EXPERTISES ET PUBLICATIONS | COMPRENDRE, PRÉVENIR, DÉPISTER | PATIENTS ET PROCHES | PROFESSIONNELS D SANTÉ

Accueil ▸ Actualités et événements ▸ Actualités

▸ Le groupe inter-SIRIC OSIRIS propose une liste d'items cliniques et omiques clés pour le partage de données et la description des essais de médecin...

Tweeter Partager Partager

Actualités

Le groupe inter-SIRIC OSIRIS propose une liste d'items cliniques et omiques clés pour le partage de données et la description des essais de médecine moléculaire

11/05/2017

Contact

Bastien Rance

HEGP, AP-HP | INSERM

bastien.rance@aphp.fr

Actions on Biomedical Data implies

Philip E. Bourne, NIH Associate Director for Data Science

- Insuring data quality and hence trust
- Making data sustainable
- Making data open and accessible
- Making data findable
- Providing suitable metadata and annotation
- Making data queryable
- Making data analyzable
- Presenting data as to maximize its value
- Rewarding good data practices

Boundaries on Biomedical Data implies

Philip E. Bourne, NIH Associate Director for Data Science

- Working across biological scales
- Working across biomedical disciplines
- Working across basic and clinical research and practice
- Working across institutional boundaries
- Working across public and private sectors
- Working across national and international borders
- Working across funding agencies