S^2CR^3UM: A Solution to the In Silico Relevance, Reliability & Reproducibility Conundrum

Informaticists face daunting challenges with data management. Stony Brook Department of Biomedical Informatics created a quality control program to improve reliability, reproducibility, and relevance of data products

Sarah B. Putney, JD, MA[1]; Andrew White, PhD[2]; Janos Hajagos, PhD[3]; Jonas Almeida, PhD[3]; Joel H. Saltz, MD, PhD[3]; Mary. M. Saltz, MD[3]
[1]Suffolk Care Collaborative, Hauppauge, NY
[2]Rensselaer Polytechnic Institute of Technology, Troy, NY
[3]Department of Biomedical Informatics, Stony Brook School of Medicine, Stony Brook, NY

Stony Brook University

Stony Brook Medicine

Ensure Data Quality

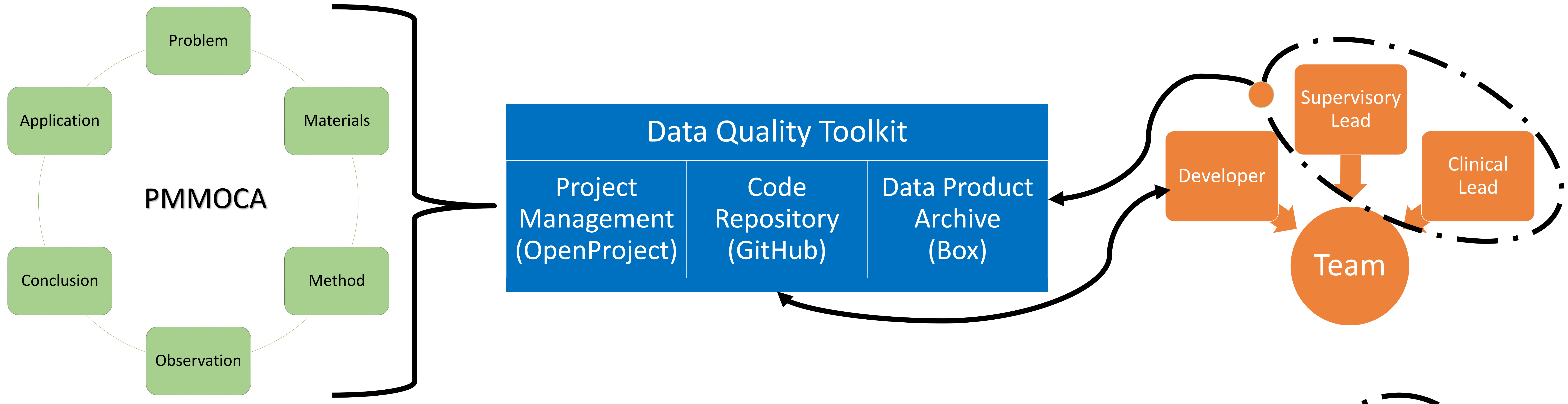
Transform Processes

Transform Culture

Reason and Rigor in Analysis (PMMOCA)

Tools and Infrastructure

Roles and Responsibilities



PMMOCA

Problem

Materials

Method

Observation

Conclusion

Application

Data Quality Toolkit

Project Management (OpenProject)

Code Repository (GitHub)

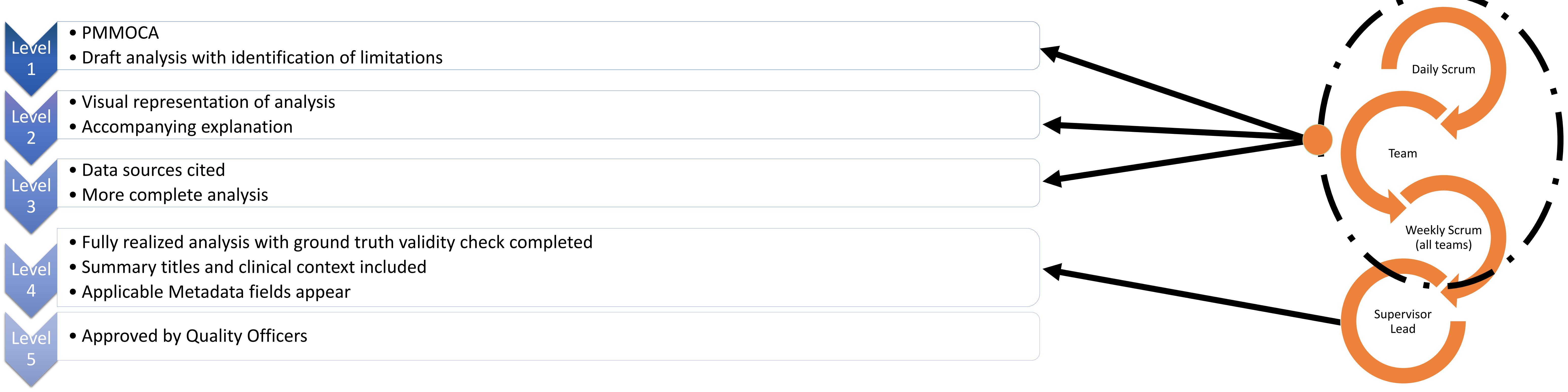
Data Product Archive (Box)

Supervisory Lead

Developer

Clinical Lead

Team



Level 1
- PMMOCA
- Draft analysis with identification of limitations

Level 2
- Visual representation of analysis
- Accompanying explanation

Level 3
- Data sources cited
- More complete analysis

Level 4
- Fully realized analysis with ground truth validity check completed
- Summary titles and clinical context included
- Applicable Metadata fields appear

Level 5
- Approved by Quality Officers

Daily Scrum

Team

Weekly Scrum (all teams)

Supervisor Lead

Developing the Data Product

Methods and Objectives:
We defined two synergistic goals for this project: a) to develop, implement, and iterate on a Toolkit for Quality improvement and ongoing quality control (QC) for data analytics; and b) to transform the culture to support a nimble and productive team. The program meta-process is organized around defined roles and responsibilities, and coordinated across time and technical systems, to deliver data products meeting scientific standards of relevance, reliability, and reproducibility. People & Communications: Roles and responsibilities were defined for the team and all developed tools and shared ideas for process improvements, based in part on The Checklist Manifesto1. Building on this foundation, the group drafted of a core set of checklists to guide workflow and to promote accuracy and consistency. Borrowing from the model of agile software development, a rapid cycle of work-review-correct/revise, a set of "scrums" (informal but focused meetings) began: thrice-weekly scrums where developers report on current work; more inclusive weekly scrums at which project priorities are set and data products in process are critiqued; and a weekly "super" scrum for program leaders. Program leaders reinforce expectations for attendance and participation. Systems for Data, Code & Data Products: A shared infrastructure was adopted to promote interoperability, retrieval, version control, reproducibility, etc. OpenProject (https://www.openproject.org/) is used for project management with projects and tasks defined using a simple structured format (Problem, Materials, Methods, Observations, Conclusions, Application.). GitHub (https://github.com/) is the designated repository to store software code for work in process and finalized. Analogously, BOX (https://www.box.com/) serves as the repository of data products at all stages of development, made searchable through a naming convention that included task #, completeness level, a 4-5 word description of the product, and file format. Links to storage of final products and code are placed in the OpenProject tracker. A 5 level grading system, with a checklist for each completeness level, was developed. The final and most complete level (5), includes criteria such as references to data provenance, a cross check for data veracity, the number of populations and subgroups analyzed, and geographic or service provider scope. Interactive tools for data visualization and exploration, such as integration of patient addresses to Google Street View (publically available with open source at http://sbu-bmi.github.io/dsrip/maps.html) were developed to help guide understanding of results. Standard templates are used to promote brand identity and consistency. Monthly project review meetings are conducted, engaging subject matter experts as needed, to ensure relevance and completeness of data products. Observations: Nine months after implementation review shows new data products to be easily retrieved, consistent, reproducible and well branded. Departmental improved quality allowed us to target the Stony Brook Medicine focus on Clostridium difficile (C. diff) testing. This effort encompassed dataset testing and validation to ensure accuracy of combined laboratory test, location and medication information. Insights gained in this dataset quality process allowed us to collaborate with Cerner to iteratively tune the ETL process and to optimize usefulness of the information obtained from our newly deployed Healthe EDW data warehouse. Challenges include consistent adherence to standards and checklists, but changes in workflow have made consistent QI a vital part of departmental culture.

Conclusion:
We conclude that by changing culture around data quality and by using a Toolkit for Quality, scientific integrity and in silico quality control can converge in a manageable, affordable, and productive workflow in an academic setting.

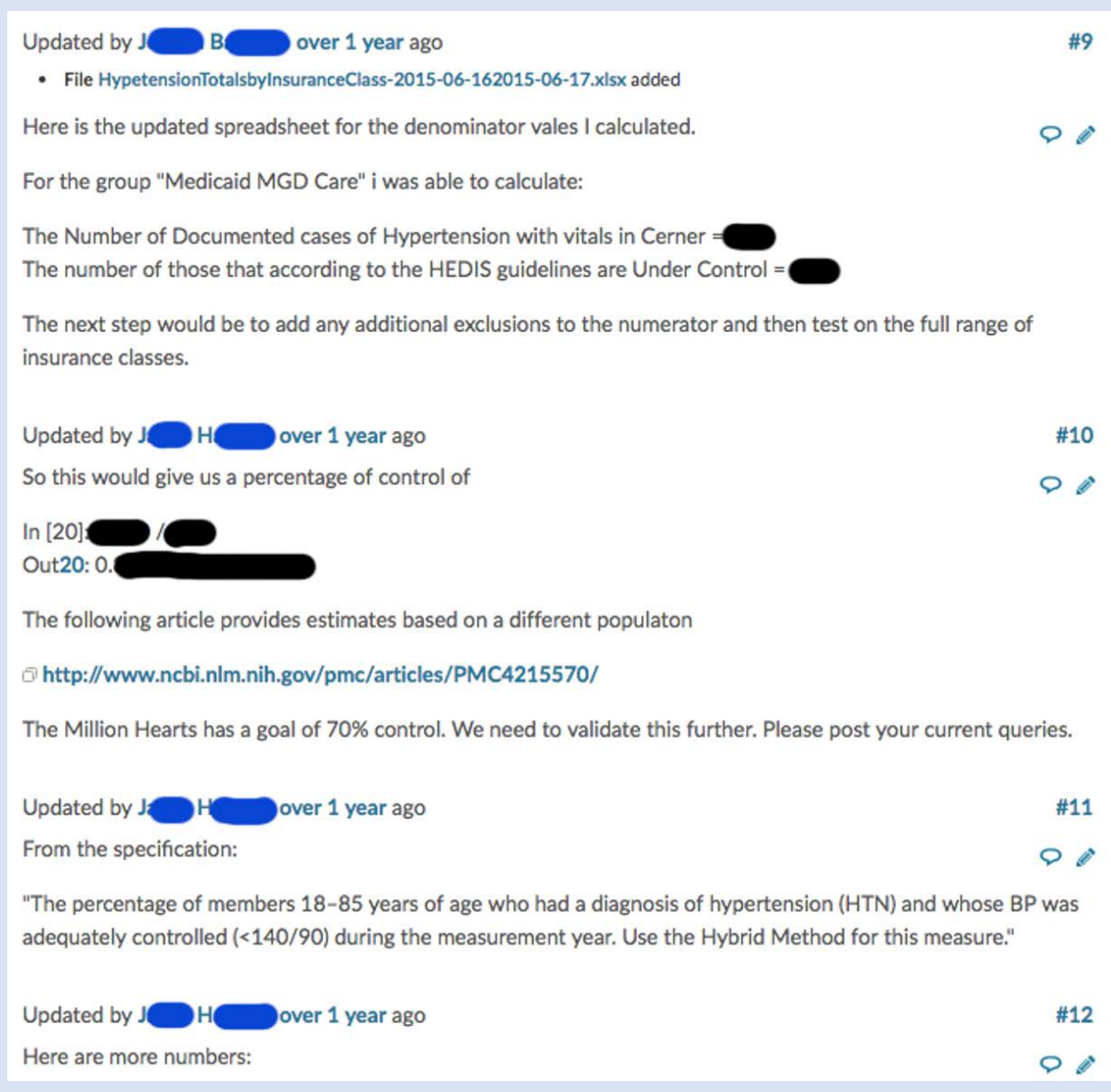
References:

[1]Gawande A. The checklist manifesto. New York: Metropolitan Books; 2010.

Data Product Levels

Data Level	Action	Method	Result
Level 1 – A	Early query and analysis results (i.e. tables, SQL code, pointing out issues) in progress	Informal sharing with others at daily scrums; feedback from lead with revisions	Not archived by BMI; clarification from BMI CISO where to be stored
Level 2 – B	Gelling ideas and initial visualizations (i.e. graphs, heatmaps, charts)	BMI meetings internal presentations at weekly scrums	Not archived by BMI; clarification from BMI CISO where to be stored
Level 3 – Γ	Probably final data products awaiting Level 4 certification	BMI meetings and presentations at weekly Data Analytics - submit for review and approval to Quality Officers	Not archived by BMI; clarification from BMI CISO where to be stored
Level 4 – Δ	Data products which have undergone "sanity cross-check"	May be provisionally shared	Archived as level 4 data product and marked with a Δ
Level 5 - Ω	Data products approved by Quality Officers	Final data product	Archived as level 5 data product and marked with a Ω

The Iterative Process



Updated by J[redacted] B[redacted] over 1 year ago #9
- File HypetensionTotalsbyInsuranceClass-2015-06-162015-06-17.xlsx added

Here is the updated spreadsheet for the denominator vales I calculated.

For the group "Medicaid MGD Care" i was able to calculate:

The Number of Documented cases of Hypertension with vitals in Cerner =[redacted]
The number of those that according to the HEDIS guidelines are Under Control =[redacted]

The next step would be to add any additional exclusions to the numerator and then test on the full range of insurance classes.

Updated by J[redacted] H[redacted] over 1 year ago #10

So this would give us a percentage of control of

In [20]:[redacted] /[redacted]
Out20: 0.[redacted]

The following article provides estimates based on a different populaton

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215570/

The Million Hearts has a goal of 70% control. We need to validate this further. Please post your current queries.

Updated by J[redacted] H[redacted] over 1 year ago #11

From the specification:

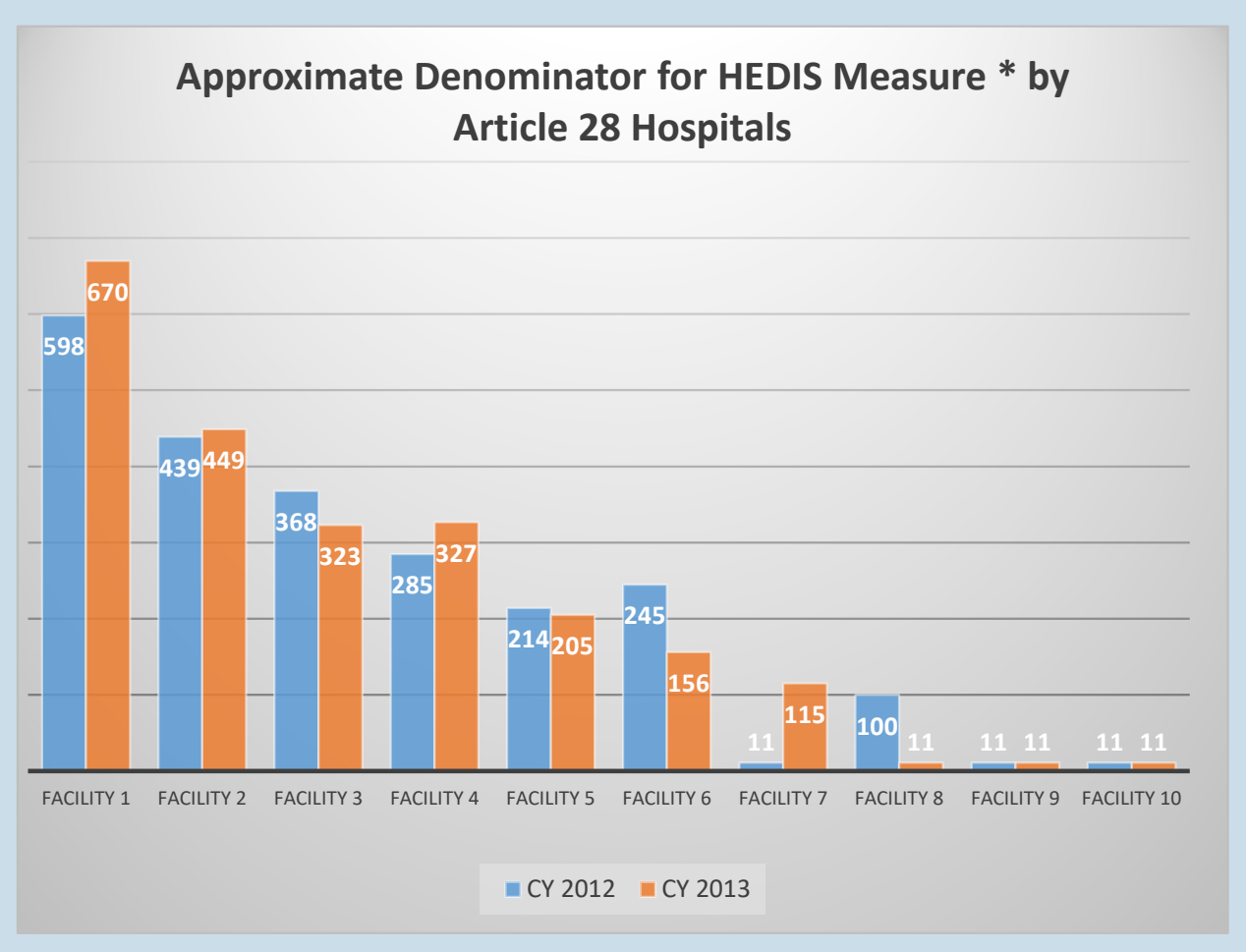
"The percentage of members 18-85 years of age who had a diagnosis of hypertension (HTN) and whose BP was adequately controlled (<140/90) during the measurement year. Use the Hybrid Method for this measure."

Updated by J[redacted] H[redacted] over 1 year ago #12

Here are more numbers:

The Final Data Product

Conditions used to define the total population of Suffolk County PPS members used to compute the HEDIS Measure* by Article 28 Hospitals in <u>Suffolk County</u>



Approximate Denominator for HEDIS Measure * by Article 28 Hospitals

Version/Date	Version 3; Date Produced
DSRIP Project:	3.a.1
Creator:	Developer Q
Data Sources:	X
Reviewers:	Clinical liaison, Developer Q
Date Range:	CY2012-2013
Codes Used: ICD-9	Dx: applicable ICD-9 Codes
Service Site:	Article 28 Hospitals NYS
Insurance Type:	Medicaid
N=total discharges with primary Dx above	5085
Cross Check	*** CY 2013 billed claims data
Text Explanation:	Hotspot map may help illustrate geographic distribution of encounters. Note: includes all patients treated in Suffolk County regardless of where they live.

Ω

Creating the Tools

Checklists for quality

Checklist to Clarify a Task
- ❑ Who is the data product requestor?
- ❑ What is requested?
- ❑ Why is this requested?
- ❑ When is the product due to requestor/customer?
- ❑ Which Open Project Task(s) refer(s) to this project
- ❑ How will the data product be formatted?
- ❑ Do we have access to the data needed to create the product?

User: Manager setting up data product request in Open project; Read-Do/Do-confirm

Checklist of 5 Quality Principles
Every BMI Data Product must be:
- ❑ Consumer facing: saved in institutional Box.com as a dataset or de-referenceable link
- ❑ Discoverable: someone other than product developer or technical lead has recognized tagging
- ❑ Contextualized: by manager with at least one Open Project task
- ❑ Shared: with at least one customer/stakeholder, preferably in person, and revised per feedback
- ❑ Reproducible: work code saved in GitHub

Users: Everyone!

Checklist to guide Daily Scrum Task Reports
- ❑ Manager: verify Checklist to Clarify a Task run in Open Project
- ❑ DP Creator: State task # requested and verify it equals task done
- ❑ DP Creator: report/discuss/request guidance
  - ❑ Sanity checks
  - ❑ Conservation principle
  - ❑ 5 Quality Principles
  - ❑ hours-report how many and if entered
- ❑ Manager and DP creator: agree on next steps and who will enter hours

Users: Manager leading Daily Scrum-Read/Do; Data Product Creator presenting a task at Daily Scrum – Read/Do

Checklist to Prepare Data Product for Weekly Scrum
- ❑ Has the Checklist of 5 Quality Principles been done?
- ❑ Can you account for every line of data per conversation principle?
- ❑ Do all applicable Metadata Fields appear on data product?
- ❑ What feedback from stakeholders does this need?

User: Data Product Creator before presenting at Weekly Scrum; Read-Do/Do-Confirm

Checklist for level 4 QC Check
- ❑ Applicable Metadata fields appear
- ❑ Checklist for Sharing Outside BMI is complete in Google Docs

User: Quality Control Officers as Read-Do/Do-Confirm to determine Level 4 Quality

Checklist for Preparing Data Product for Sharing Outside BMI
- ❑ Complies with Data Privacy/Security policies
- ❑ Stored in Box.com as dataset or via link
- ❑ Sufficiently tagged (bottom-up and top-down ontologies OK)

User: Data Product Creator; Do-Confirm to ensure that product is consumer-facing (can be accessed, shared, and taken further by stakeholder outside BMI)

Checklist for Our Checklists
- ❑ Is it short?
- ❑ Does it make clear:
- ❑ Purpose?
- ❑ User(s)?
- ❑ How to use it? (verbal, read-do, do-confirm, etc.)
- ❑ Version date?
- ❑ Does it detect issues when they can still be resolved
- ❑ Has it been tried in real scenarios
- ❑ Has it been revised in response to repeated trials?

User: BMI Personnel developing, using. Or refining Checklist for BMI QC purposes; Read-Do/Do-Confirm

PMMOCA

PM-MOCA – Problem, Materials, Methods, Observations, Conclusions, Applications
Example: 30-day readmission for Heart Failure patients
Problem
- What are the numbers associated with the 30-day readmission of the SBUH patients with primary diagnosis of Heart Failure during CY2011-13?
- This task serves DSRIP Project 2.a.iv "Care transitions intervention to reduce 30-day readmissions for chronic disease."
- Over-arching DSRIP goal is to reduce unnecessary hospitalizations in the Medicaid and uninsured population by 25% over 5 years.
- Understanding 30-day readmissions is key to reducing unnecessary hospitalizations

Materials
- Patients with one of the following ICD-9 codes as the primary discharge diagnosis:
  - 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.0, 428.1, 428.20, 428.21, 428.22, 428.23, 428.30, 428.31, 428.32, 428.33, 428.40, 428.41, 428.42, 428.43, 428.9
- Data source: merged SBUH SMS and Cerner data from CY2011-CY2013

Methods
Calculations based upon:
    Top 15 primary diagnoses (by # of encounters)
        Time of admission
        Time of discharge
    Unplanned 30-day readmission rates by payer
    Unplanned 30-day readmission rates by dispositions
    Etc…

Observations
- Patients coming emergently to the hospital and being discharged to home represent > 1,000 encounters with > 20% readmission rate
- Cardiology patients being discharged to home represent ~ 1,500 encounters with subsequent unplanned 30-day readmissions.

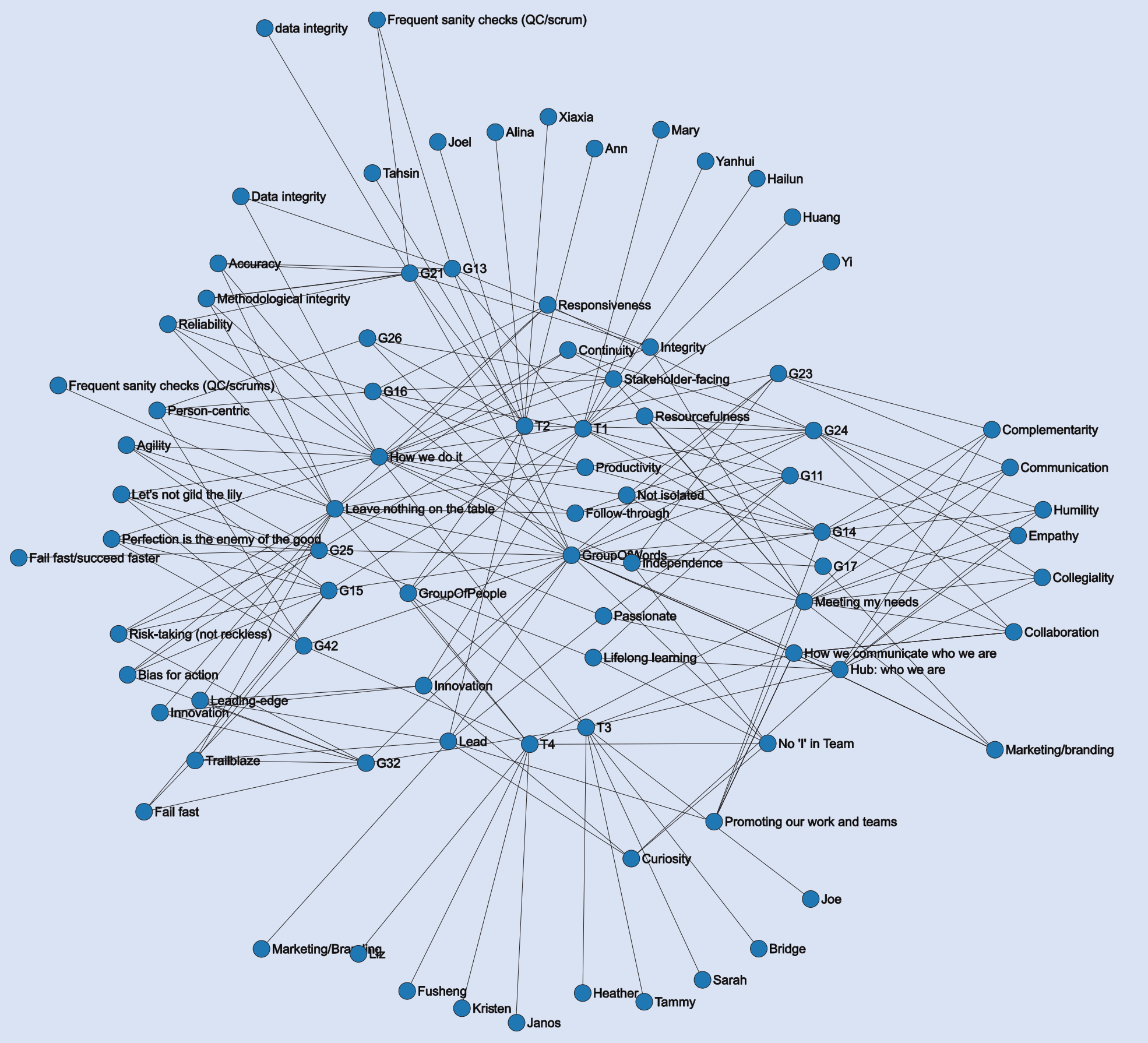
Conclusions
- Cardiology has the highest readmission rate as well as the largest number of encounters
- CHF Patients readmitted within 30 days almost all come back because of symptomatic HF

Application
- Patients coming emergently to the hospital and being discharged to home represent > 1,000 encounters with > 20% readmission rate and targeting this population with increased support post discharge may positively impact outcomes
- UHC Service Lines of GI, Medicine, and Cardiology are areas where an improvement could impact overall hospital performance

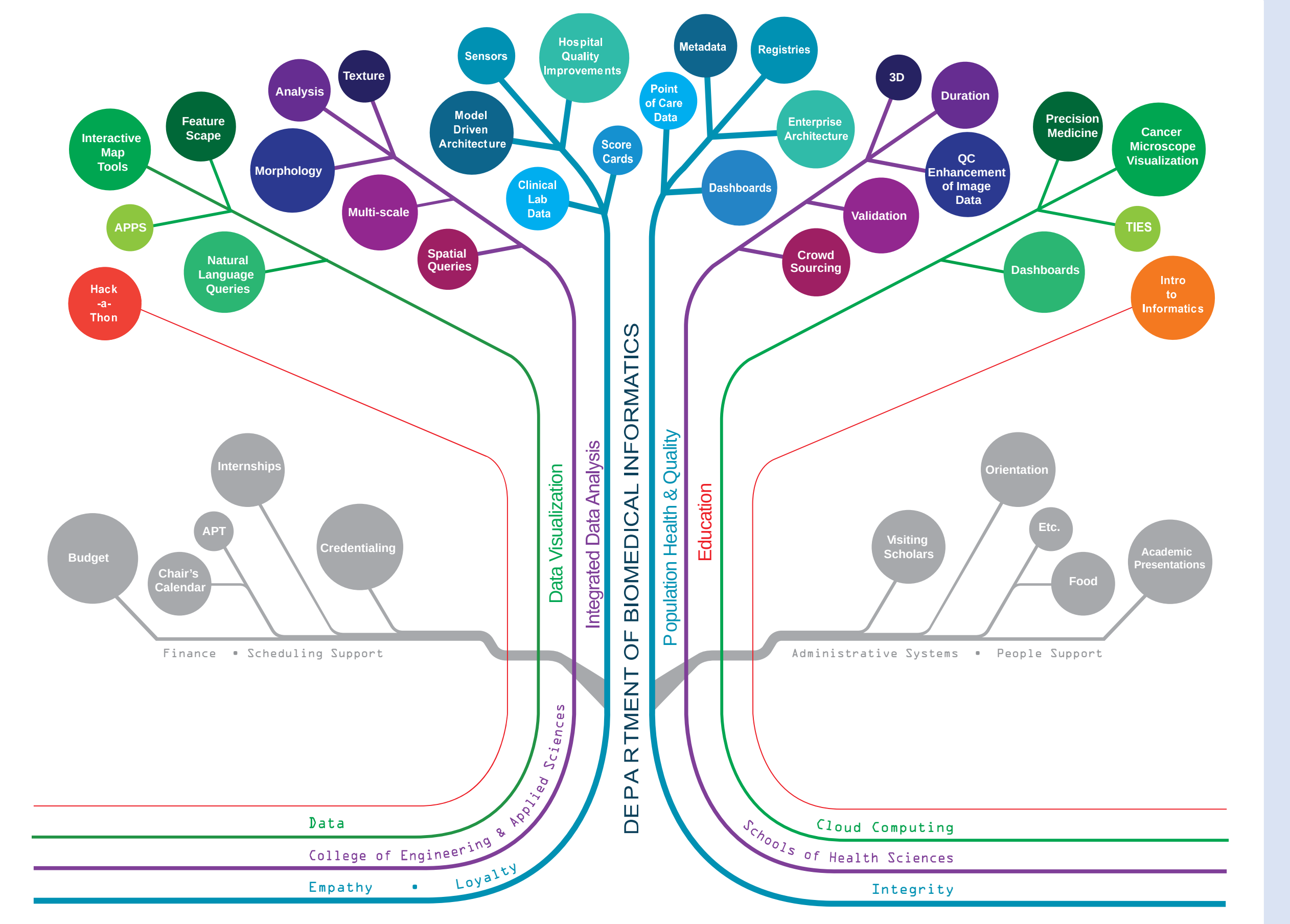
Building the Team

Departmental Retreat

The mission of the Department of Biomedical Informatics of Stony Brook University is to advanced biomedical knowledge through innovative data science and education



Interactive version at: https://bmi.stonybrookmedicine.edu/wordcloud/



Poster created by Erich Bremer