Research Article

Bridging the Collaboration Gap: Real-time Identification of Clinical Specimens for Biomedical Research

Thomas J. S. Durant[1,2], Guannan Gong[2,3], Nathan Price[4], Wade L. Schulz[1,2]

[1]Department of Laboratory Medicine, Yale University School of Medicine, New Haven, CT, USA, [2]Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT, USA, [3]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University School of Medicine, New Haven, CT, USA, [4]Department of Information Technology, Yale New Haven Health, New Haven, CT, USA

Submitted: 10-Mar-2020 Revised: 17-Mar-2020 Accepted: 30-Mar-2020 Published: 20-May-2020

Abstract

Introduction: Biomedical and translational research often relies on the evaluation of patients or specimens that meet specific clinical or laboratory criteria. The typical approach used to identify biospecimens is a manual, retrospective process that exists outside the clinical workflow. This often makes biospecimen collection cost prohibitive and prevents the collection of analytes with short stability times. Emerging data architectures offer novel approaches to enhance specimen-identification practices. To this end, we present a new tool that can be deployed in a real-time environment to automate the identification and notification of available biospecimens for biomedical research. **Methods:** Real-time clinical and laboratory data from Cloverleaf (Infor, NY, NY) were acquired within our computational health platform, which is built on open-source applications. Study-specific filters were developed in NiFi (Apache Software Foundation, Wakefield, MA, USA) to identify the study-appropriate specimens in real time. Specimen metadata were stored in Elasticsearch (Elastic N. V., Mountain View, CA, USA) for visualization and automated alerting. **Results:** Between June 2018 and December 2018, we identified 2992 unique specimens belonging to 2815 unique patients, split between two different use cases. Based on laboratory policy for specimen retention and study-specific stability requirements, secure E-mail notifications were sent to investigators to automatically notify of availability. The assessment of throughput on commodity hardware demonstrates the ability to scale to approximately 2000 results per second. **Conclusion:** This work demonstrates that real-world clinical data can be analyzed in real time to increase the efficiency of biospecimen identification with minimal overhead for the clinical laboratory. Future work will integrate additional data types, including the analysis of unstructured data, to enable more complex cases and biospecimen identification.

Keywords: Biobanking, biomedical research, biospecimen science, clinical specimens, real-time identification, translational research

Introduction

In the era of precision medicine, human biospecimens are an important resource for basic, translational, and clinical research and are increasingly needed to advance our understanding of human physiology, disease, treatment response, and outcomes. The field of biobanking has undergone significant optimization efforts by national and international communities to improve and harmonize biospecimen curation to support this need.[1,2] However, the operationalization and maintenance of biobanks is resource-intensive and often cost prohibitive for many institutions. In addition, long-term biobanking may be suboptimal for some types of testing, such as for studies that rely on labile analytes.[3,4] As a result, comprehensive access to human biospecimens remains limited, and there is a persistent need for efficient solutions that can provide access to high quality and recently acquired human biospecimens.[5]

Human biospecimens can always be found in clinical laboratories, but access for the research is complicated by a series of technical, logistic, regulatory, and ethical challenges. Beyond the demands of delivering clinical results, laboratories lack efficient processes for biospecimen identification, human resources for sample

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_15_20

Address for correspondence: Dr. Wade L. Schulz, Department of Laboratory Medicine, 55 Park Street PS502A, New Haven, CT, USA.
E-mail: wade.schulz@yale.edu

How to cite this article: Durant TJ, Gong G, Price N, Schulz WL. Bridging the collaboration gap: Real-time identification of clinical specimens for biomedical research. J Pathol Inform 2020;11:14.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2020/11/1/14/284660

acquisition, and procedural infrastructure for biospecimen collection under the provisions of human interventional ethics committees. Despite these challenges, the clinical laboratory is a promising resource for the acquisition of biospecimens, and researchers are beginning to investigate curation methods that can integrate with existing clinical workflows and leverage EHR metadata for biospecimen identification and annotation.[6,7]

One of the first-automated biospecimen identification systems was Crimson; an application used to identify the discarded blood samples accessioned into the clinical laboratory by querying the laboratory information system. Samples, which met predetermined inclusion criteria, were electronically reaccessioned into a deidentified research database that could be accessed by researchers with IRB approval.[8] While biobanks routinely link health information between specimen and participant postenrollment, such solutions demonstrate how EHR integration and associated metadata can be used for targeted and automated biospecimen selection. However, examples of this framework remain limited, both in the literature and in practice, which typically focus on retrospective specimen identification for long-term biobanking. With increased digitization of healthcare and modern data architectures that allow for real-time analysis of clinical data, biospecimens can be identified as samples that are processed through the clinical laboratory. This approach offers the benefit of increasing access to specimens of interest, including those with labile analytes, while not disrupting routine clinical workflows.[4]

In this report, we present Prism, a new tool built on open-source technology that can efficiently identify and notify the investigators of biospecimen availability in near real time. We describe the pipeline architecture and our experience with two IRB-approved pilot projects within our department (IRB Protocol IDs: Babesia – 2000023123; Diabetic biomarkers – 2000022266).

Methods

Prism platform architecture

We implemented a real-time pipeline, called Prism, that consists of three key components: real-time data acquisition, stream processing, and end-user alerting, to support case and specimen identification [Figure 1]. Parameterized NiFi processors were used to filter and identify the clinical specimens based on study-specific inclusion criteria extracted from corresponding laboratory result metadata. NiFi is an open-source application designed for stream processing. Specimens that met inclusion criteria were indexed within Elastic search (Elastic NV; Mountain View, CA, United States). Alerting was done through Watcher (Elastic NV; Mountain View, CA, United States) and secure E-mail, with a reporting dashboard built in Kibana (Elastic NV; Mountain View, CA, United States).

This framework was deployed within our organization's computational health-care platform, Baikal, which has been previously described [Supplemental Figure 1].[9] The Baikal platform is built on open-source technology and provides a mechanism to manage and analyze high-volume and high-frequency clinical data in real-time, including laboratory results.

Throughput assessment

Scalability and computing resource needs for Prism were estimated through the deployment in a standalone workstation environment with a single CPU with 6 cores (Intel Core i7-6850K CPU @ 3.60GHz) and 256 GiB of memory. Apache NiFi was deployed within a Docker (version 19.03.2, build 6a30dfc; Docker, Inc., San Francisco, CA, United States) container under the Ubuntu (Version: 16.04.6 LTS (Xenial); Canonical Ltd; London, UK) operating system. We ran a modified version of the Prism dataflow using file-based record I/O instead of streaming data from a network interface. Data for the assessment were obtained by randomly selecting data from our production Health-Level 7 (HL7) feed that were assembled into three data sets of increasing size. These data sets contained 1×10^5, 1×10^6, and 1×10^7 Java Script Object Notation-transformed HL7 ORU messages, resulting in 0.75 GB, 7.27 GB, and 72.8 GB of data, respectively. Two series of five trials were performed with the 1×10^6 record data set. In the first series, all five trials were run consecutively. In the second series, Docker was restarted between each trial to assess for any possible performance impacts in long-running containers. Throughput was measured using built-in NiFi monitoring tools to assess record count and throughput.

Results

Babesia specimen identification

Babesia is a tick-borne hemoprotozoan, which infects human erythrocytes and can be life-threatening for patients who are asplenic, immunocompromised, or elderly. The gold standard for the laboratory diagnosis is microscopic analysis of peripheral blood smear. For research into the automation of digital microscopic analysis using the computer vision, the researcher needed peripheral blood smears, which were identified as containing Babesia. Incoming HL7 messages corresponding to a Babesia result record with a "Positive" result values were flagged and sent to the Prism index in Elasticsearch [Figure 2]. Researchers were securely notified of all "Positive" Babesia specimens identified every 4 h.

Specimen identification for positive Babesia specimens went live in May 2018. In a collection period of 16 months (June 2018–September 2019), Prism identified 131 unique lavender-top tubes, belonging to 44 unique patients, which were identified as positive for Babesia by manual light microscopy. The collection period for this project was extended beyond the anticipated time requirement as Babesia exhibits a strong seasonal prevalence, and positive specimen rates dropped over the colder months [Figure 3].

Diabetic biomarker specimen identification

The development of type 2 diabetes can be prevented or delayed in prediabetic individuals with lifestyle modifications such as

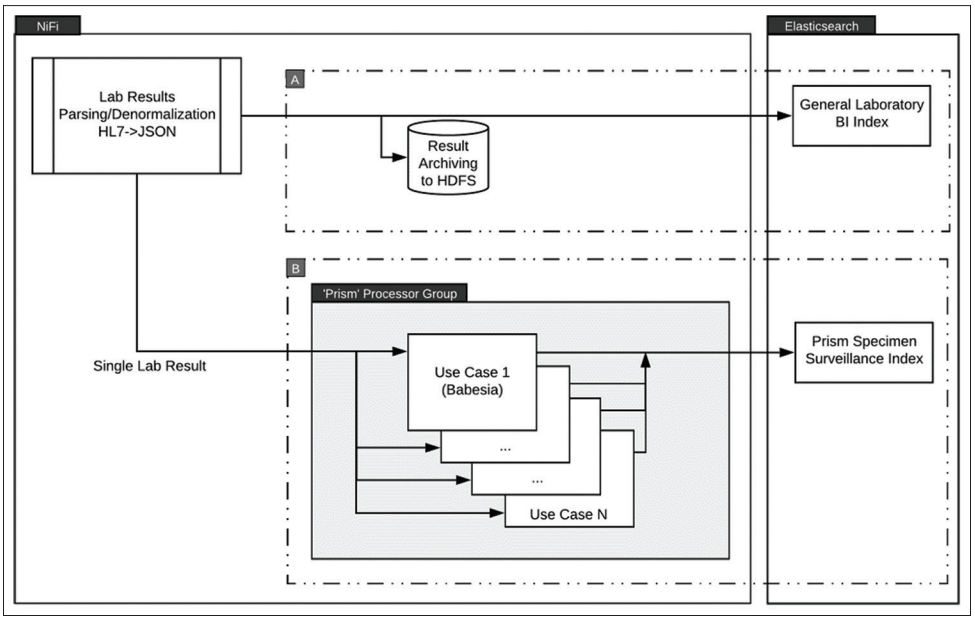


Figure 1: Dataflow diagram for laboratory results in using NiFi and Elasticsearch. (A) Existing laboratory result dataflow. (B) "Prism" specimen identification dataflow. BI: Business Intelligence, HDFS: Hadoop-Distributed File System, HL7: Health-Level 7, JSON: Java Script Object Notation

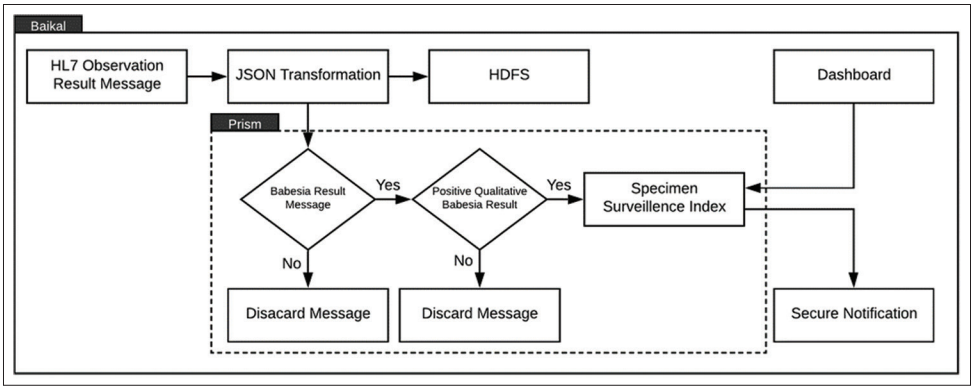


Figure 2: Laboratory result monitoring for positive Babesia specimens. Incoming HL7 observations are transformed to denormalized JSON documents and stored to HDFS. Prism dataflow ingests streaming JSON result records and filters "Positive" Babesia results to the "Prism" Specimen Surveillance Index from which secure notifications of positive Babesia results are generated. HDFS: Hadoop-Distributed File System, HL7: Health-Level 7, JSON: Java Script Object Notation

	Collection Date		Unique Count	
Use Case	**Start**	**End**	**Specimens**	**Patients**
Babesia	2018-06-13	2019-09-22	131	44
HbA1c-LGT	2018-12-07	2019-03-20	2,861	2,771

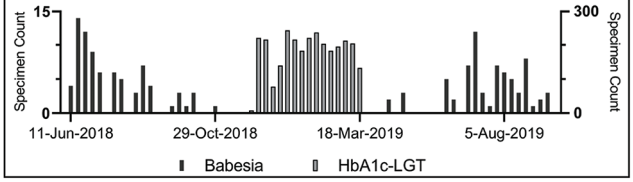


Figure 3: Total number of unique patients and specimens identified for Babesia and A1C-LGT specific use cases within the specified date range. Columns represent count of unique specimens per week. Right Y-axis: HbA1c-LGT, Left Y-axis: Babesia. HbA1c: Hemoglobin A1c, LGT: Light-green top

dietary changes or increased physical activity. Accordingly, there is a need to identify the biomarkers to guide preventative interventions.[10,11] To identify possible biomarkers, a researcher at our institution was interested in obtaining blood specimens from patients with and without diabetes, with borderline cases excluded, as a prelude to a larger prospective biomarker study. The deidentified samples would undergo metabolomic analysis by the liquid chromatography-mass spectrometry to identify the metabolites that were significantly changed between the two groups as candidate biomarkers.

Hemoglobin A1C values < 5.7 and > 6.5 were used to delineate between diabetic and nondiabetic patients, with additional inclusion criteria of outpatient specimen collection and patient age range 18–70 years. Of note, the preferred collection container for Hgb A1c at our institution is a lavender-top tube,

which does not contain gel-separation barriers. In an effort to optimize biomarker recovery, plasma from light-green-top tubes was requested for this study. Accordingly, LGT tubes were flagged when a paired sample with an Hgb A1c within the appropriate range was found within the past 7 days [Figure 4]. Researchers were securely notified every morning by the E-mail of all matching LGT specimens present in the Prism index reported within the past 24 h.

Specimen identification for diabetic biomarker discovery went live in December 2018. In a collection period of 4 months (December 2018 and March 2019), Prism identified 2,861 unique LGT specimens from 2,771 unique patients [Figure 3].

Throughput assessment

We assessed the processing throughput to ensure the pipeline could scale to large environments and consistently manage high-volume data. Our institution's computational health platform processes approximately 350,000 discrete HL7 ORU messages per day. Accordingly, we evaluated processing time across five trials and observed an average execution time of approximately 8 min for one million records, which represents slightly <3 days of laboratory result volume. Processing time was observed to be linear over two orders of magnitude in dataset size [Figure 5a], and the average total execution time to process one million messages differed by 2% between runs with (494 s) and without (483 s) Docker container restart [Figure 5b].

Discussion

In this report, we describe a novel data analysis pipeline, called Prism that can be used to improve the efficiency of biospecimen collection. This workflow has been deployed to identify the biospecimens in near real-time for two biomedical research use cases. We demonstrated that this solution is highly scalable to meet the needs of even large academic centers and reference laboratories. We also found, consistent with our prior work, that virtualization of this workflow within a microservices environment does not introduce a performance penalty.[12]

In 2000, it was estimated that 300 million human biospecimens were preserved in the United States, with a projected 7% annual growth rate.[13] However, researchers continue to report difficulty in obtaining specimens for biomedical research and express underlying concerns in the validity of their results when using specimens subjected to long-term storage conditions.[5] In addition, while many biospecimens are being stored, a large proportion is expected to remain unused, and there is increasing concern that untargeted collection of biospecimens consumes resources that could be better allocated.[14-16] Accordingly, as institutions seek to expand biomedical research efforts, particularly in the era of personalized medicine, novel approaches for improving access to high-quality human biospecimens should be evaluated.

The quality of biomedical research is dependent on the integrity of biospecimens and as with clinical testing, analyte recovery is subject to a significant number of preanalytical considerations.[4] While biobanking procedures have seen significant optimization in recent years, poor reproducibility of studies that use biospecimens has been thought to be caused, in part, by the variable quality and inadequate documentation of biospecimen metadata.[5] To this end, biobanks are beginning to emulate testing procedures found in the clinical laboratory to optimize the analyte recovery and test reproducibility.[17,18] Tools that can identify the samples accessioned to the clinical laboratory, such as Prism, would align with these efforts by identifying the specimens that have been collected and processed under clinical conditions.

Despite ongoing adoption of clinical procedures in biospecimen science, the collection and processing of labile analytes remain challenging, and some components may require unique processing protocols.[17,19] Proteomic and molecular analytes are particularly sensitive to specimen transport delays, matrix effects, and optimal-storage environments.[20] Accordingly, some components of interest may require sample processing techniques that exist outside routine clinical workflows. In this setting, real-time streaming analytics could also be envisioned to identify patients which match study-specific inclusion criteria to guide targeted subject enrollment and subsequent collection.

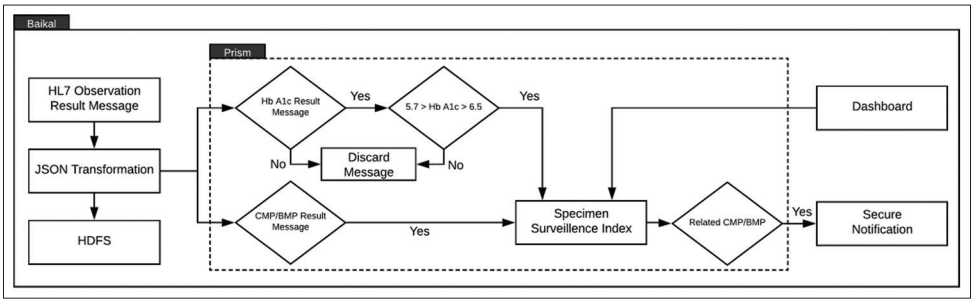


Figure 4: Laboratory result processing diagram for diabetic biomarker monitoring. Incoming HL7 observations are transformed to denormalized Java Script Object Notation documents and stored to HDFS. The Prism dataflow ingests streaming Java Script Object Notation result records and filters hemoglobin A1C results in the "Normal" (<5.7) and "Diabetic" (>6.5) cohorts to the "Prism" Specimen Surveillance Index in Elastic. Results from CMP/BMP panels (light-green top specimens) are sent to the Prism index. Secure notifications are sent for A1C specimen IDs with related light-green specimen info. BMP: Basic Metabolic Panel, CMP: Comprehensive Metabolic Panel, HbA1c: Hemoglobin A1c; HDFS: Hadoop-Distributed File System, HL7: Health-Level 7, JSON: Java Script Object Notation

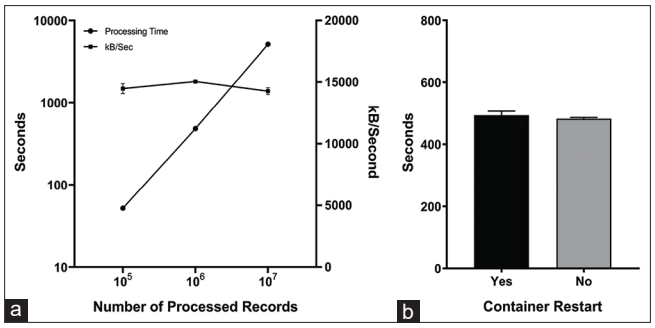


Figure 5: (a) Average number of seconds (Left Y-axis) required to process the laboratory JSON result records and average processing rate (kB/sec) across five consecutive trials, without container restart, for JSON record datasets increasing in size by factor of 10. JSON: Java Script Object Notation, kB: Kilobyte. (b) Average number of seconds required to process 10^6 Java Script Object Notation result records across five consecutive trials, with and without Docker container restart between trials

In addition to specimen identification and collection, annotation with patient metadata remains an important and challenging facet of contemporary biobanking. Large scale biobanks such as the U. K. Biobank rely on a combination of data sources for curating specimen metadata including participant enrollment surveys, physical measures (e.g., blood pressure and spirometry), and linkage to digital health information.[21,22] Indeed, while the majority of national biobanking resources capture data from both inpatient and outpatient medical records, there is also interest in capturing data that is not stored in the EHR.[21,23,24] As digital health information continues to expand, health-care systems are increasingly working to develop clinically integrated data-management tools for the centralization of disparate data resources.[9] Deployment of automated specimen identification tools in these frameworks may facilitate correlation with these data and would align with national efforts to do so.

It should be noted that the use cases described in this report were selected based on the immediate needs among researchers in our department. However, similar open-source tools could be similarly envisioned to integrate with anatomic pathology data and the EHR, to automatically phenotype tissue specimens as they are processed in the laboratory. While the majority of data elements in the clinical laboratory are discrete, identifying tissue specimens in the anatomic pathology laboratory may require technologies such as natural language processing (NLP) to process semistructured and unstructured data, such as those commonly found in pathology reports.[25] While not used for this implementation, custom NiFi-processors would allow the users to develop more complex filters and integrate NLP or machine learning-based technology for free text or nested data structures commonly found in anatomic pathology. Similarly, the platform can also be used to identify the patients who may be eligible to consent and enroll in studies, rather than simply for biospecimen collection.

In the era of digital and personalized medicine, novel approaches to increase the efficiency of biospecimen identification will be crucial to accelerate discovery. Modern data architectures as described here can be used to address the fundamental challenges in the procurement of biospecimens in support of biomedical research. Future work will seek to integrate additional data types, including the analysis of unstructured data, to enable more complex case and biospecimen identification.

Disclosure

Wade Schulz was an investigator for a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; is a technical consultant to HugoHealth, a personal health information platform, and co-founder of Refactor Health, an AI-augmented data management platform for healthcare; is a consultant for Interpace Diagnostics Group, a molecular diagnostics company.

Financial support and sponsorship

Nil.

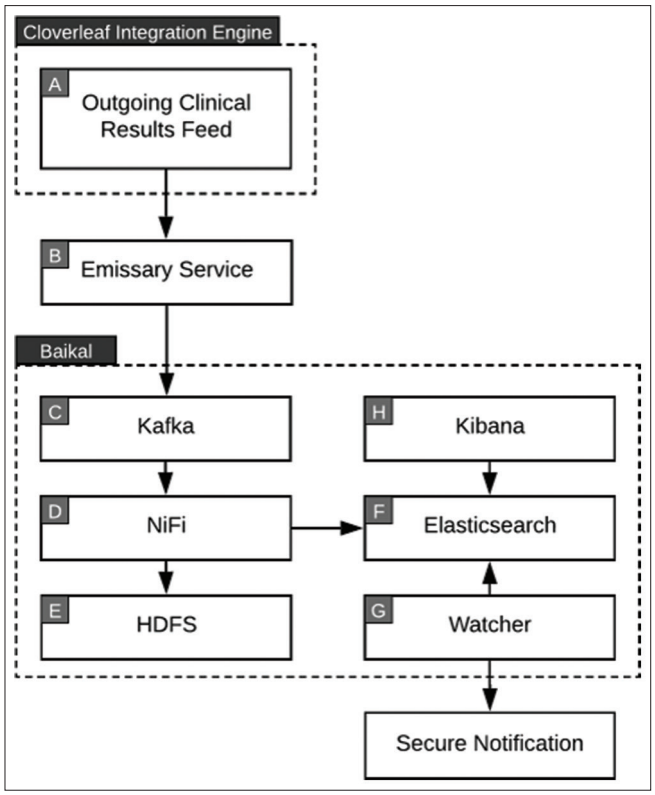
Conflicts of interest

There are no conflicts of interest.

References

1. van Ommen GJ, Törnwall O, Bréchot C, Dagher G, Galli J, Hveem K, *et al*. BBMRI-ERIC as a resource for pharmaceutical and life science industries: The development of biobank-based Expert Centres. Eur J Hum Genet 2015;23:893-900.
2. Langhof H, Kahrass H, Illig T, Jahns R, Strech D. Current practices for access, compensation, and prioritization in biobanks. Results from an interview study. Eur J Hum Genet 2018;26:1572-81.
3. Shabihkhani M, Lucey GM, Wei B, Mareninov S, Lou JJ, Vinters HV, *et al*. The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. Clin Biochem 2014;47:258-66.
4. Ellervik C, Vaught J. Preanalytical variables affecting the integrity of human biospecimens in biobanking. Clin Chem 2015;61:914-34.
5. Massett HA, Atkinson NL, Weber D, Myles R, Ryan C, Grady M, *et al*. Assessing the need for a standardized cancer HUman Biobank (caHUB): Findings from a national survey with cancer researchers. J Natl Cancer Inst Monogr 2011;2011:8-15.
6. Moore HM, Kelly AB, Jewell SD, McShane LM, Clark DP, Greenspan R, *et al*. Biospecimen reporting for improved study quality (BRISQ). Cancer Cytopathol 2011;119:92-101.
7. Simeon-Dubach D, Burt AD, Hall PA. Quality really matters: The need to improve specimen quality in biomedical research. J Pathol 2012;228:431-3.
8. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, *et al*. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res 2009;19:1675-81.
9. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, *et al*. Health Care and Precision Medicine Research: Analysis of a Scalable Data Science Platform. J Med Internet Res 2019;21:e13043.
10. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, *et al*. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol 2012;8:615.
11. Guasch-Ferré M, Hruby A, Toledo E, Clish CB, Martínez-González MA, Salas-Salvadó J, *et al*. Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. Diabetes Care 2016;39:833-46.
12. Schulz WL, Durant TJ, Siddon AJ, Torres R. Use of application containers and workflows for genomic data analysis. J Pathol Inform

2016;7:53.
13. Eiseman E, Haga SB. Handbook of Human Tissue Sources. A National Resource of Human Tissue Samples. Handbook of Human Tissue Sources a National Resource of Human Tissue Samples. Washington, D.C; 1999.
14. Gee S, Georghiou L, Oliver R, Yuille M. Financing UK Biobanks: Rationale for a National Biobanking Research Infrastructure. Final Report of Stratum Work Package; 2013. p. 7.
15. Simeon-Dubach D, Watson P. Biobanking 3.0: Evidence based and customer focused biobanking. Clin Biochem 2014;47:300-8.
16. Simeon-Dubach D, Henderson MK. Sustainability in biobanking. Biopreserv Biobank 2014;12:287-91.
17. Vaught J. Biobanking comes of age: The transition to biospecimen science. Annu Rev Pharmacol Toxicol 2016;56:211-28.
18. Betsou F, Barnes R, Burke T, Coppola D, Desouza Y, Eliason J, *et al*. Human biospecimen research: Experimental protocol and quality control tools. Cancer Epidemiol Biomarkers Prev 2009;18:1017-25.
19. Vaught J, Rogers J, Myers K, Lim MD, Lockhart N, Moore H, *et al*. An NCI perspective on creating sustainable biospecimen resources. J Natl Cancer Inst Monogr 2011;2011:1-7.
20. El Messaoudi S, Rolet F, Mouliere F, Thierry AR. Circulating cell free DNA: Preanalytical considerations. Clin Chim Acta 2013;424:222-30.
21. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, *et al*. All of Us Research Program Investigators. The "All of us" research program. N Engl J Med. 2019;381:668-76.
22. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, *et al*. The UK biobank resource with deep phenotyping and genomic data. Nature 2018;562:203-9.
23. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, *et al*. China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011;40:1652-66.
24. Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman E, Million veteran program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70:214-23.
25. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, *et al*. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23.



Supplemental Figure 1: System architecture for Baikal. (A) Health-Level 7 messages generated by the laboratory information system and received by clinical integration engine. (B) Health-level 7 messages validated by custom emissary service and mapped to Java Script Object Notation. (C) Java Script Object Notation submitted to Kafka for downstream processing. (D) Custom Python scripts executed in NiFi to parse and denormalize messages. (E) Denormalized data are stored in Hadoop-distributed File System. (F) Java Script Object Notation and quality improvement metrics are stored in Elasticsearch. (G) Watcher runs scheduled queries on Elasticsearch data and generates notifications. (H) Kibana is used for visualizations and dashboards