bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Next Generation Precision Medical Record

A Framework for Integrating Genomes and Wearable Sensors with Medical Records

Daryl Waggott, Euan Ashley

To Review & Approve: Anja Bog, Anil Ankisettipalli, Surbhi Sheth-Shah, Enakshi Singh, Megan Grove, Prag Batra, Mark H Wright, Matthew Wheeler, Jason Merker, Werner Eberhardt, Carlos D Bustamante

Affiliations
1. Stanford Department Medicine
2. Stanford Department of Biomedical Data Science
3. Stanford Center for Cardiovascular Innovation
4. Stanford Division of Cardiovascular Medicine
5. Stanford Center for Computational, Evolutionary, and Human Genomics
6. SAP Innovation Center, Silicon Valley
7. SAP Healthcare Development Unit

Funding
SAP Stanford Innovation Grant

Acknowledgments
SAP Innovation Team
SAP Healthcare Team

Abstract

Current medical records are rigid with regards to emerging big biomedical data. Several examples that already exist in clinical practice include whole genome sequencing and wearable sensors for real time monitoring. Genome sequencing not only enables conventional diagnostic interrogation, but forms the fundamental baseline for precision health throughout a patient's lifetime. Mobile sensors enable tailored monitoring regimes for both reducing risk through precision health interventions and acute condition surveillance. To address the absence of these data in the EMR we worked with SAP Healthcare to re-envision a modern medical record with these components. The pilot project used 37 patient families with complex medical records, whole genome sequencing and some level wearable monitoring. Core functionality include patient timelines with integrated text analytics, personalized genomic curation and wearable

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

alerts. The current phase is being rolled out to over 1500 patients in clinics across the hospital system. While fundamentally research, we believe this proof of principle platform is the first of its kind and represents the future of data driven clinical medicine.

Background

Electronic Medical Records are the core of modern health care systems. Over the last 20 years they have slowly emerged from paper based records into an essential tool for tracking contact information, patient encounters, quality of care metrics and billing. While they exist as successful tools for administrative purposes they tend to be technically unsophisticated when it comes to clinical care. Increasingly, high volumes of clinical data, such as imaging and genomics, are being generated and immediately siloed and often discarded due to lack of central infrastructure. While potentially transformative to clinical care, the demands of stability and curse of legacy, leave EMR systems limping at a glacial pace towards cost efficiency and medical innovation.

The complexities of American health care system were discussed in detail in `*The Anatomy of Health Care in the United States*` (Moses et al. 2013). The system is staggering, as of 2011, 15.7% of the US workforce was employed in health care, with expenditures of $2.7 trillion (17.9% of GDP). Even with increases in resources, prices are rising (hospital charges +4.2%/yr, drugs and devices +4.0%/yr) and basic quality metrics such as life expectancy are behind other developed nations. The authors discuss several major forces of change in health care that speak directly to the challenge and opportunity that this framework works to address. First, consolidation moving from horizontal aggregation of similar entities to vertically integrated heterogeneous and generally insular organizations. Second, information technology investments have focused on administrative efficiencies and improving quality of care. While admirable objectives these goals have a great deal of potential for improvement. For example US billing costs are 13% vs 6.6% in Canada. While, for coordinated clinical care to be realized, information needs to be collected, digital, standardized, shared and ultimately innovative clinical applications created. This is not trivial in a large, slow moving, regulatory burdened system.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The existing medical record is opaque to most modern analytical tools (Wang et al. 2015; Tamang et al. 2015; Jung et al. 2015; Curé et al. 2015). Clinicians struggle to search for their own patients using keywords, and in many cases cannot complete this simple task. The core of the clinical encounters are recorded as a series of clinical notes to form a timeline. Notes are free text and subjective to each clinician of varying quality and absolutely lacking any type of semantic structure. Moreover, external reports, images or assays are generally stored as bitmap data blobs without metadata and no direct method of querying. Even if clinical labs or diagnostic tests have been run they are often entered manually and therefore riddled with errors or simply missing. Shockingly, extracting the relevant disorder or ICD9 code is often not even possible without manual intervention (Mark and David 2015). Luckily, there has been developments to embed greater support for medical ontologies to better capture diagnosis and symptoms (Girdea et al. 2013; Robinson et al. 2008). There are also a number of disruptively nimble EMR platforms such as the cloud based Practice Fusion and Athena Health (http://www.practicefusion.com/; http://www.athenahealth.com/).

In this document we propose an alternative model focused on deep data centric patient integration. Specifically, we propose a framework for linking patient medical records with genomic data and wearable monitoring. A key dependency for this integration is the ability to transform unstructured clinical encounters into informatically amenable data structures (Liao et al. 2015). This transformation will enable contextual interpretation of genomic results and wearable monitoring for continuous care. The concept of the a linked record has been worked on with varying success in most national healthcare systems, but has been difficult within the balkanized nature of US system (Weber, Mandl, and Kohane 2014). What has not been not been tackled sufficiently is the integration of large, complex data within a patient's actual record. In this paper we demonstrate the integration of two main data types, genomics and mobile health within an intelligent medical record.

Clinical genetic testing is quickly moving from single marker tests to large multi gene panels digitally captured from exome and whole genome (Yang et al. 2014; Lee et al. 2014; Zhu et al. 2015). Most major EMR's have no capacity to store genomic data, and if so, only store high level metadata or a few variants that are deemed reportable. This is changing and some custom systems have been incredible examples of what is possible (Dumitrescu et al. 2015; Alterovitz et al. 2015). Even when genomics is collected it is often excluded from from the medical record, or at least limited to a report of a single key finding. Several reasons (beyond ethical) for exclusion include informatic practicality (100s of millions of data points per patient) and lack of clinical utility or genetic actionability. The former is a technical issue that we address in our framework, while the latter is more subtle.

First, genomic sequencing typically originates as a specific diagnostic test and therefore belongs within the scope of the original clinical purpose. While technically this is logical it highlights that genome sequencing should be part of patient's core medical record and shared among clinicians and supplemented improved over time. Second, clinical actionability is dependent on the diagnostic context and fails to recognize the huge potential for application in

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

multiple health domains over the lifetime of a patient. In addition, clinically relevant genomic information is systematically evolving through initiatives such as ClinGen. The arguments against storing genomic data are somewhat circular. That is, the future clinical value of genomic data is dependent on collection and collection is argued against by its value proposition. Third, the current clinical record by definition is an atomic unit specific to a patient as defined by HIPAA. This can be at odds with the current clinical genetics diagnostic process which depends on inclusion of family members or patients with similar conditions for diagnostic power (Dewey et al. 2011). Moreover, ACMG guidelines on reporting incidental findings recommend notification of family members to potential carrier status, which is not possible using this model. Storing all of a patient's clinical data within EMR would therefore entail risk of future incidental findings and consequent ethical and resource obligations. We systematically tackle the practical issues of storing genomes within the EMR in our framework.

Mobile health is in it's infancy relative to genomics within clinical practice. The vast majority of devices are consumer oriented with sophisticated collection of health information on activity, sleep patterns and heart rate. With current clinical practice these metrics are only opportunistically captured at discrete time points. This fails to capture latent risk situations such as changes in magnitude or variability over time or under specific conditions. As an example, atrial fibrillation which is a major risk factor for stroke, has an estimated prevalence ~2% and growing rapidly (3.7-4.2% >49yrs) (Colilla et al. 2013; Chugh et al. 2014; Zoni-Berisso et al. 2014). The presence of the condition is unknown to most patients due to gradual onset or infrequent episodes. For those that do have a moderate or early stage diagnosis, very little is done to continuously monitor and consequently prevent severe events. Similarly, cardiomyopathy which affects 1/500 is a major risk factor for AF and sudden death, but is rarely monitored outside of clinic. Wearables are capable of making these measurements and for providing affordable non-invasive monitoring of patients by clinicians. While many clinics are working on innovative applications for these data they struggle with not being able to incorporate the results into the clinical record in a meaningful integrated way. For example, patients with elevated risk could be monitored more closely and targeted for aggressive engagement strategies. A modern clinical record is crucial for this type of intelligent profiling i.e. integrating genetic risk predisposition, observed subclinical risk factors such as stationary behavior, presence of comorbidities or taking a medication with predicted adverse events based on a specific genotype.

In this document we describe an analytical framework for integrating classical EMRs with genomics and wearables to create the next generation clinical record. All of this can be done without changing the underlying EMR investment. We believe the future of healthcare is going to be heavily dependent on granular patient measurement with continuous monitoring and an emphasis on machine learning based algorithms to inform precision health. The solution is to re-imagine what an integrated system would look like with these core fundamentals built in. To the best of our knowledge this is the first application of integrating genomics and wearable sensor data with medical records for actual patients.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results and Features

Patient Triage

A key challenge to a system that integrates many users, patients and data is task or key information prioritization. Upon logging into the framework a clinician or informatician will be presented with patients ranked by two main criteria. First, patients will be seen ordered relative to outstanding tasks for their record. This could include various aspects of variant/gene curation to Sanger validation or additional test ordering. Second, patients will be ordered relative to key information or alerts pertinent to their condition or general health. The representation of the alerts and tasks will be specific to the roles and assigned patients for the user. An example screenshot can be seen in Figure 1.

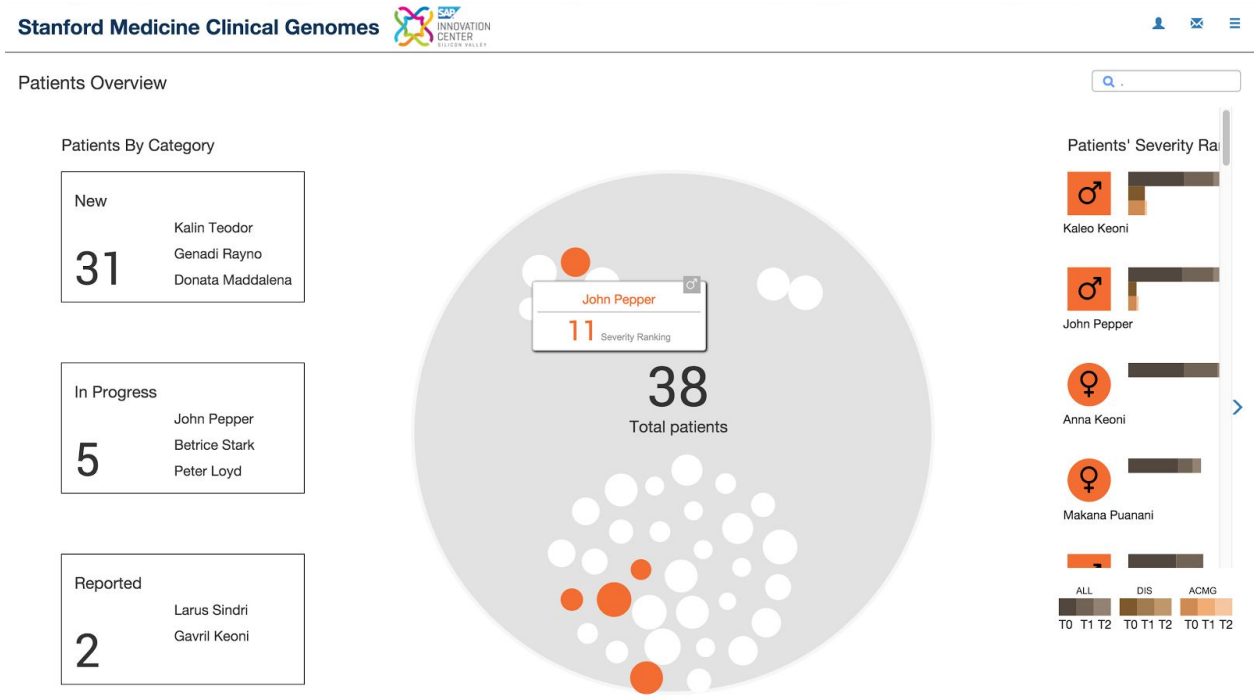

Figure 1 - Patient Triage

Patient Dashboard

The patient triage process then directs a user to a patient dashboard. This is a simplified yet rich overview of a patient's integrated precision medical record. Key elements include an interactive patient clinical timeline, conventional lab results, wearable sensor alerts and genomic summaries. The genomic summaries can are stratified into key clinical findings (either incidental, actionable or condition specific), drug response variants, ABO/HLA genotypes and

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

global ancestry. Each of the elements can be selected for more in-depth review and will be discussed below. An example screenshot can be seen in Figure 2.

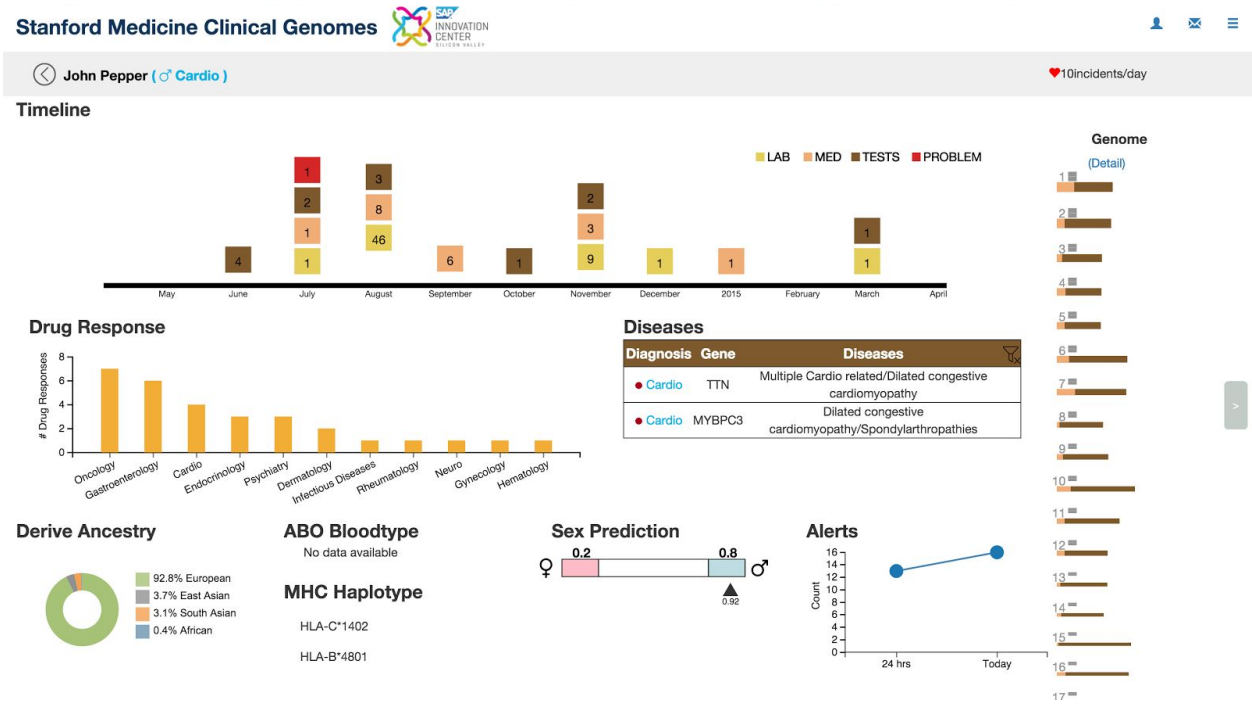

Figure 2 - Patient Dashboard

Scalable Genome Storage

First and foremost the infrastructure was designed to scale to handle all genomes for all patients within a large national healthcare systems. Our ten node system in its current configuration scaled with negligible performance difference to include all 1000 genomes data. In its current configuration we estimate that this can be extended to 100,000 genomes. Any further extension would require straightforward addition of more HANA nodes.

Complex annotation joins and aggregate queries are on the order of millisecond. Initial genomic load times from complex VCF files currently take 5 ½ minutes per whole genome (~10M variants). This step happens only once. It's anticipated that we can dramatically improve these load times by not parsing and temporarily storing all the INFO and FORMAT tags present in the VCF. Ideally, the variant calling pipeline and output VCF is specifically optimized for downstream storage. The genome data model can be reviewed in more detail in supplementary file 1.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Automated Genetic Curation

Genetic curation for risk features and pathogenic variants is central to our vision of the precision medical record. In this framework the medical record is mined for key terms which inform candidate gene lists for variant prioritization. The objective is to empower curation experts and clinicians to focus on high value tasks and not data wrangling.

The heart of the automated curation is based on a software we developed called Sequence To Medical Phenotypes (STMP) (Dewey et al. 2015). One of the core modules tiers variants within a geneset according to potential pathogenicity. The primary tiers are (1) clinvar pathogenic 2+ stars (2) LOF or canonical splice (3) High across species amino acid conservation site (4) Predicted protein deleterious across multiple algorithms. Additional STMP modules focus on prioritization based segregation within a pedigree and PGx Haplotyping using and HMM. Preloaded gene lists exist for numerous conditions and classes of actionability. Using STMP within the Hana system allows for substantially improved query performance and simplified management of the large number of external database dependencies. Within the patient dashboard of Patient X (Figure 2), note that the identified disease, candidate gene and variant from the automated prioritization agreed with previous manual curation efforts (Cardiomyopathy: *MYBPC3*).

As the underlying medical record changes and genetic annotation evolves, so too does the curation. Most importantly the genomic information is directly in the medical record whereby relevant information is “pushed” to the clinician and not manually “pulled” by a curator. The concept of continuous annotation updates and “push” curation makes re-analysis an automated ongoing task. Figure 3 shows an example of tiered variant prioritization along with a query interface for adjusting thresholds i.e. allele frequency, inheritance model, gene tolerance.

Upon automated prioritization candidate genes and their respective variants can be reviewed independently with extensive opportunity for comments/citations to improve ongoing knowledge base development. Variants can be queried for frequency and clinical associations across all other patients in the system. In fact, these aggregate queries can be included in clinical genetics reporting. Templates exist not only for final clinical genetic reports but intermediate presentations for team discussion. All reports contribute back to the knowledge base for improved future curation.

Beyond complexities around billing, the primary limitation to scaling genomic services in a clinical setting is time spent on curation of genes and variants. The personnel costs alone are substantial. Our approach of integrating genomics into the medical record with substantial back end analytics simplifies human resource burden, streamlines process and ultimately reduces cost.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

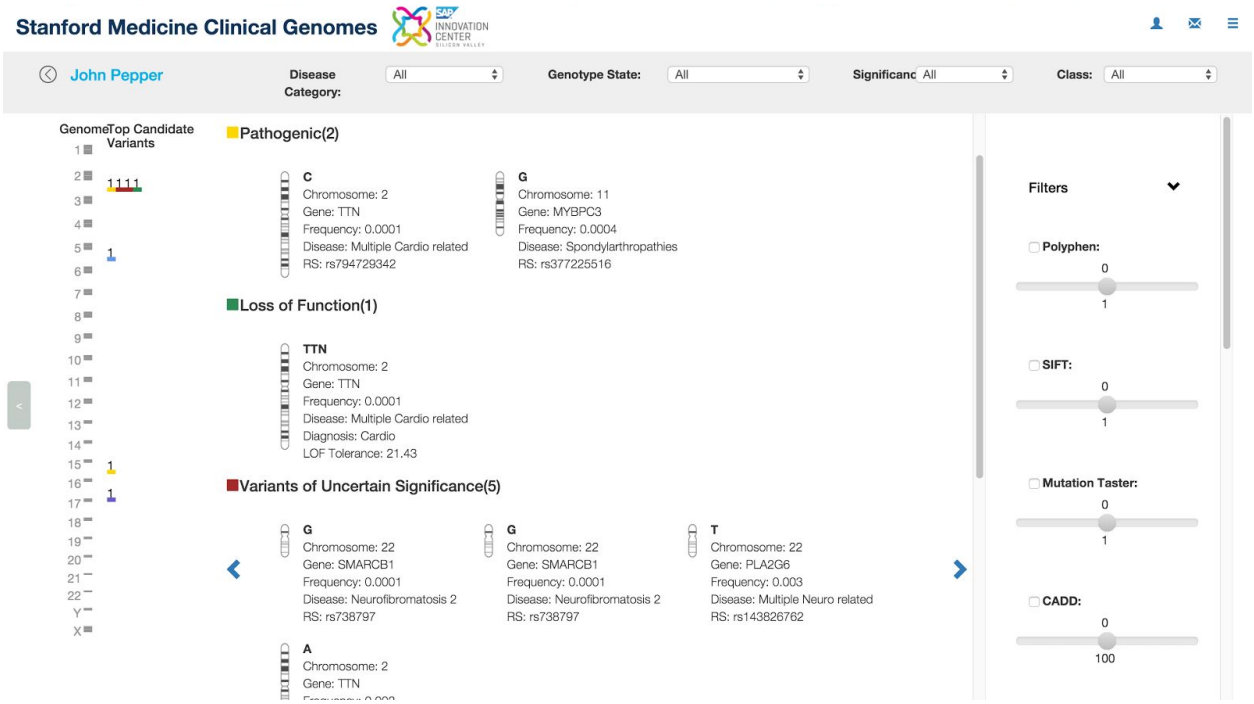

Figure 3 - Variant and Gene Curation

Complex Genomic Annotation

The genome provides a large amount of accessible information relevant to clinical care. Single diagnostic tests for pathogenic variants ignore the use of this rich information for ongoing precision health. As such we added a number of annotations that depend combining information across multiple variants. Several examples include ABO/HLA haplotypes, PGx haplotypes, sex, relatedness and ancestry. Figure 4 shows a summary of how variants can be interpreted in the context of family segregation and local ancestry (Maples et al. 2013). Additionally, genetic risk scores for common disease are in the process of being added.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4 - Complex Genomic Annotation

Live Wearable Monitoring

Figure 5 shows a two day window of Patient X monitored using an Apple watch (3118 observations). Patient X suffered from cardiomyopathy and therefore any heart rate abnormalities would be of clinical interest. We chose to focus on defining a specific metric for HR relative to predicted energy expenditure. If the patient deviates from the expected relationship an alert is triggered and this is relayed back to the medical record and monitoring physician. You'll notice that toward the right of the figure there exists an increased density of elevated HR relative to predicted energy expenditure. Several caveats include: (1) Our data was collected with up to a 48 hour synchronization lag that needs to be optimized. (2) Based on a previous clinical validation we found Apple HR reasonably accurate but energy expenditure (especially older algorithms based on tri-axial accelerometer sensors) was poor *(clinical device validation - in preparation)*. We are currently working a core set of similar personalized sensor metrics that could be applied to different health/disease states.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

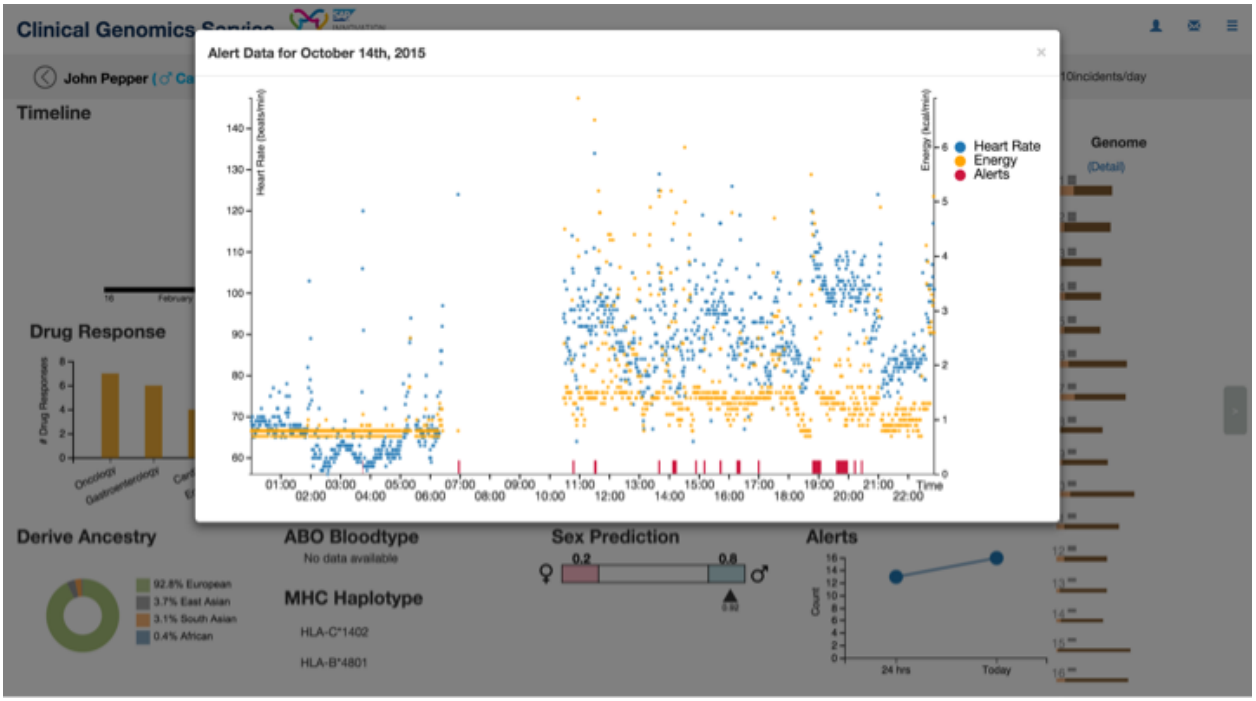

Figure 5 - HR and Energy Expenditure Monitoring

EMR Adapters for Augmentation not Replacement

Stanford Medicine has two independent installations of EPIC. Access to the records are typically restricted to limited forms with no bulk import/export. Our initial pilot of this framework focused on a static EPIC export using the STRIDE platform. This approach is only suitable for research and not production. For clinical utility real time bulk access needs to be enabled. Based on a collaboration with ASCO a solution was developed that creates an EMR abstraction layer for controlled access (Saphire_2015; ASCO_Connection; ASCO_Meeting_2016). These adapters cover the majority of major EMR platforms and are strictly compliant with HIPAA/PHI regulations. As such, data flow is primarily pulled from the EMR and not pushed back. One highly valuable byproduct of the use of abstraction is that unstructured EMR data i.e. clinical notes, is transformed into structured information such as key terms and timelines. The use of EMR adapters serves to augment without changing the underlying informatic health structure such as billing and operations.

Text Mining of Disease Terms

The majority of a clinical encounter is recorded as a “clinical note”. This is primarily free text with a large amount heterogeneity in quality/quantity between clinicians. In significant number of cases we reviewed, there was no clear or easily machine accessible differential, diagnosis or structured clinical traits/symptoms. This information therefore required extraction from the clinical notes. Based on an existing SAP technology called Medical Research Insights we built a

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

natural language processing pipeline to several key clinical features ("MRI Demo," n.d.; "MRI Accelerate Personalized Trx" 2014). The primary objective was to extract candidate diagnosis. For the most part this depended on identifying ICD9 codes and disease/diagnostic terms i.e. MESH or HPO values such as cardiomyopathy. The terms were then assigned to a timeline where they were mapped to candidate gene sets using tools such as DisGenNET, OMIM and ClinVar. When possible well curated disease gene panels such as GeneDx or Invitae would be selected. If a candidate gene panel was not available an *in silico* panel would be created based on the described ontology lookup. The digital candidate gene panel is incredibly important for specificity during curation.

The disease terms further act to tune the use wearable sensor data collection. Different patient classes would have different algorithms and relevant clinical alerts. For example a cardiology patient will have a number of HR and exertion measurements where as a neurology patient may have more of a focus on motility, fine-motor and coordination.

Patient Timeline and the Longitudinal Genome

A concept that is not fully developed in our EPIC medical record installation is the patient timeline. Through a combination of clinical note parsing and fuzzy date matching with external tables (procedures, prescriptions, labs), we were able to derive a patient's clinical timeline. The longitudinal perspective of patient encounters and treatment is important for better organization, capturing change in health over time, integration with sensor monitoring and maintenance of clinical care. From the genomics standpoint, automated genetic curation is relevant to a specific window of time. As expected the relevance of different genomic features will change over time within a patient. For example cardiovascular risk or treatment vs oncology variants would be used independently in time and by practitioner. Improved focus on a longitudinal view of patient's health helps emphasize the relevance of prospective health as opposed to reactionary treatment. Moreover, it demonstrates that the utility of genomics is multifaceted and should exist as a fundamental baseline within all medical records. Finally, continuous sensor monitoring can adapted over time and simply correlated with clinical management. Fundamentally, patient health is not cross sectional and we show proof of principle the utility of its incorporation into the medical record.

Version-ome

The fundamental algorithms and supporting annotation used for calling and interpreting features is evolving at terrific pace. In addition, over a patient's lifetime it's possible to have multiple tissues sequenced for different purposes such as germline risk or tumor profiling. Given that a person's genomic profile is not static, we chose to explore the concept of a "version-ome". Essentially, can we apply conventional database management techniques to allow for regulatory compliant, evolving representation of one's genome. For example access datestamps, storing incremental changes on updates and strictly versioning the hundreds of secondary data sources used for curation. This versioning allows for reproducible curation queries, so that specific

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variants to be revisited years in the future. For example any variant found in a clinical report will have have an auditable query associated with it.

Contextual Clinician Alerts

Another key design thought for our next generation medical record system is some degree of automated intelligence. With semi-automated prioritization/curation the system can alert clinicians to only the most fundamental changes in information relevant to a patient's current condition or latent risk. For example cardiovascular metrics become elevated and patients genomic risk is high. Subsequently, PGx variants both in terms of efficacy/safety could inform treatment choice. Also, as underlying ClinVar pathogenicity assertions evolve, an alert could be initiated if a diagnostically relevant threshold is met for a VUS.
Similarly, alerts can generated based on wearable sensor data streams. As patients move outside of personalized tolerance zones (based on EMR tuning), an alert to a clinician would be initiated. More systematic trends could be summarized during regular clinical encounters and would not trigger alerts.

Methods and Technology

Development Process

The process of joint development between SAPs innovation team and Stanford Medicine was firmly rooted in design based thinking. Briefly, for six week extensive time was spent by SAP working with a broad spectrum of Stanford stakeholders to deeply understand the problem space. Key challenges were explored and presented back to the Stanford team. A direction and scope were agreed upon and then the project proceeded to agile development with a multi-disciplinary team. The final prototype was presented back to the Stanford stakeholders and which initiated revisements and an additional 3 months of operationalization.

Cohort

During development four initial cohorts of subjects were tested. Initial prototyping was done on 1000 genomes data (2,504 WGS), Personal Genome Project and 250 exomes from the ELITE project (elite.stanford.edu). The above samples were not patients, lacked accessible electronic medical records and wearable data. Subsequently, we focused development on a cohort of 37 whole genomes from a Stanford Medicine clinical genome pilot project and 8 from the Stanford Undiagnosed Disease Network clinical site. All of these patients had Stanford medical records which were accessed (de-identified) under a combination of 4 IRB protocols. Patients were a combination of cardiology, neurology, inherited cancer and multi-system presentations. A subset of the patients were selected for continuous mHealth sensor monitoring using three consumer devices (apple watch, basis peak and polar with chest strap).

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

EMR

The existing medical record system at Stanford Medicine is EPIC and a technology called STRIDE was used to map table structure and export data (Lowe et al. 2009). Among the 37 pilot patients there were 124 detailed diagnosis, >1400 medications, >22,000 lab results and >1000 procedures.

Backend Architecture

The data model, analytical engine and user interface were all implemented using the SAP HANA platform. The setup consisted of a distributed 10 node setup (9 active and one hot standby). Each node was outfitted with 1TB of memory and 80 Intel(R) Xeon(R) CPU E7 - 4870 @ 2.40GHz. More specific details on performance metrics for in memory systems can be found in the following references.

The Impact of Columnar In-Memory Databases on Enterprise Systems (Plattner 2014)
SAP HANA Database - Data Management for Modern Business Applications (Färber et al. 2012)
A Course in In-Memory Data Management (Plattner 2013)

Genetic Data Model

All data was extracted from variant call format files (VCF). The structure is incredibly extensible in order to allow tool specific annotation and QC metrics ("[PDF]The Variant Call Format (VCF) Version 4.2 ... - Samtools," n.d.). This creates a large amount of heterogeneity in data parsing and overall representation. Moreover, multiple variants or genotypes can exist at a single position, especially within large cohorts. Additionally, any one subject has 2 copies of DNA or alleles at 3 billion positions. All of these positions potentially could have annotation and QC data. Finally, for the system to be scalable it needs to be able to handle 100s of thousands even millions of genomes. Together these challenges emphasize the needs for robust yet streamlined data model.

To address these challenges we focused on only storing the genotypes in bit format with only the most fundamental QC metrics. The assumption being that a subset of QC metrics could be generated on the fly as requested. Conventional annotations (i.e. clinvar, protein domain, eQTL site...) were stored as tertiary tables for real time querying to allow asynchronous updates. The real time nature of queries enabled arbitrary multi-subject aggregation statistics and dynamic interrogation of hundreds of annotation sources.
For functional effect prediction (such as performed by Annovar, VEP and SNPEff) we selected a hybrid precompute versus on demand compute model. First we precomputed all high value SNP positions such as coding, dbSNP, clinvar and cosmic. This includes observed and unobserved genotypes in regions of functional importance. Similarly, we calculated all previously observed INDELs. For unobserved INDELs that could not be precalculated we annotate with transcript,

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

size and frame. As new SNP and INDEL positions are loaded into the database they are exported, calculated and then loaded for future caching.
A relatively simple but important distinction with VCF format is that the table grows vertically with the addition of new genomes (VCFs grow horizontally adding columns).

The specifics of the data model can be found in Supplementary File 1 and in the following patent (Patent# 20150234870 - DYNAMIC MAPPING OF EXTENSIBLE DATASETS TO RELATIONAL DATABASE SCHEMAS).

Genomics Data

Pipeline. The genotypes were called using GRCH37-GATK 2.7 best practices.
Pathogenicity. Variants were annotated with complete clinvar dumps, including assertions and reviewing organization. Prioritization of pathogenic variants was limited to two plus star likely pathogenic, pathogenic variants. Contradictory assertions and variants of unknown significance were only considered for curation purposes.
Actionable. Variants were determined to be actionable if they were annotated as likely deleterious and in ACMG guidelines or curated lists such as 114 genes proposed by (Dorschner et al. 2013).
PGx. FDA labels, pharmadme, protein atlas, cyp alleles
Risk scores. Varimed and other validated scores such as Knowles et al.
b Sex prediction was based on homozygozity of the X chromosome.
HLA. Haplotypes were estimated several tools based panels of phased polymorphic SNPs
Local Ancestry. Haplotypes were generated using high quality variants present on the Illumina Omni2.5 and phased using ShapeIT vX with 1000 genomes v3 reference panel. Local ancestry was estimated using a second generation optimization of the RFMIX algorithm (Maples et al. 2013). Briefly, the input phased haplotypes were corrected for statistically inconsistent ancestral and long range LD patterns. Next, a random forest was used to classify the most likely population for each allele within a sliding window. The reference populations were chosen for purity from recent admixture. Global ancestry was assessed by summing all segment assignments and subsequently compared to an extension of ADMIXTURE (Alexander, Novembre, and Lange 2009). The extension is based on projection of a reference population onto the target population of interest. The approach enables robust application to target datasets with confounding structure such as relatedness.

Mobile Data

Three devices were used for data collection. Specifically, the Apple watch, Basis Peak and Polar watch with chest strap. Devices were put collection mode during any modest to strenuous activity. Data was downloaded using device specific APIs. This approach is in the process of being automated with a clinical porting of the app MyHeartCounts (“MyHeartCounts,” n.d.). MHC has enrolled 50,000 users, with the collection of conventional cardiovascular risk factors in combination with mobile based sensor data through HealthKit. Porting a mirror of the back end to Stanford servers will allow for maintenance of regulatory compliance when dealing with

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

patient data. We foresee health care providers managing suites of medical apps that can be prescribed for seamless integration with medical record systems. MHC-clinical is working example of this model.

References

Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64.

Alterovitz, Gil, Jeremy Warner, Peijin Zhang, Yishen Chen, Mollie Ullman-Cullere, David Kreda, and Isaac S. Kohane. 2015. "SMART on FHIR Genomics: Facilitating Standardized Clinico-Genomic Apps." *Journal of the American Medical Informatics Association: JAMIA* 22 (6): 1173–78.

Chugh, Sumeet S., Rasmus Havmoeller, Kumar Narayanan, David Singh, Michiel Rienstra, Emelia J. Benjamin, Richard F. Gillum, et al. 2014. "Worldwide Epidemiology of Atrial Fibrillation: A Global Burden of Disease 2010 Study." *Circulation* 129 (8): 837–47.

Colilla, Susan, Ann Crow, William Petkun, Daniel E. Singer, Teresa Simon, and Xianchen Liu. 2013. "Estimates of Current and Future Incidence and Prevalence of Atrial Fibrillation in the U.S. Adult Population." *The American Journal of Cardiology* 112 (8): 1142–47.

Curé, Olivier C., Henri Maurer, Nigam H. Shah, and Paea Le Pendu. 2015. "A Formal Concept Analysis and Semantic Query Expansion Cooperation to Refine Health Outcomes of Interest." *BMC Medical Informatics and Decision Making* 15 Suppl 1 (May). biomedcentral.com: S8.

Dewey, Frederick E., Rong Chen, Sergio P. Cordero, Kelly E. Ormond, Colleen Caleshu, Konrad J. Karczewski, Michelle Whirl-Carrillo, et al. 2011. "Phased Whole-Genome Genetic Risk in a Family Quartet Using a Major Allele Reference Sequence." *PLoS Genetics* 7 (9): e1002280.

Dewey, Frederick E., Megan E. Grove, James R. Priest, Daryl Waggott, Prag Batra, Clint L. Miller, Matthew Wheeler, et al. 2015. "Sequence to Medical Phenotypes: A Framework for Interpretation of Human Whole Genome DNA Sequence Data." *PLoS Genetics* 11 (10): e1005496.

Dorschner, Michael O., Laura M. Amendola, Emily H. Turner, Peggy D. Robertson, Brian H. Shirts, Carlos J. Gallego, Robin L. Bennett, et al. 2013. "Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes." *American Journal of Human Genetics* 93 (4): 631–40.

Dumitrescu, Logan, Nicole A. Restrepo, Robert Goodloe, Jonathan Boston, Eric Farber-Eger, Sarah A. Pendergrass, William S. Bush, and Dana C. Crawford. 2015. "Towards a Phenome-Wide Catalog of Human Clinical Traits Impacted by Genetic Ancestry." *BioData Mining* 8 (November): 35.

Färber, Franz, Sang Kyun Cha, Jürgen Primsch, Christof Bornhövd, Stefan Sigg, and Wolfgang Lehner. 2012. "SAP HANA Database: Data Management for Modern Business Applications." *SIGMOD Rec.* 40 (4). New York, NY, USA: ACM: 45–51.

Girdea, Marta, Sergiu Dumitriu, Marc Fiume, Sarah Bowdin, Kym M. Boycott, Sébastien Chénier, David Chitayat, et al. 2013. "PhenoTips: Patient Phenotyping Software for Clinical and Research Use." *Human Mutation* 34 (8): 1057–65.

Jung, Kenneth, Paea LePendu, Srinivasan Iyer, Anna Bauer-Mehren, Bethany Percha, and Nigam H. Shah. 2015. "Functional Evaluation of out-of-the-Box Text-Mining Tools for Data-Mining Tasks." *Journal of the American Medical Informatics Association: JAMIA* 22 (1). jamia.oxfordjournals.org: 121–31.

Lee, Hane, Joshua L. Deignan, Naghmeh Dorrani, Samuel P. Strom, Sibel Kantarci, Fabiola Quintero-Rivera, Kingshuk Das, et al. 2014. "Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders." *JAMA: The Journal of the American Medical Association* 312 (18): 1880–87.

Liao, Katherine P., Tianxi Cai, Guergana K. Savova, Shawn N. Murphy, Elizabeth W. Karlson, Ashwin N. Ananthakrishnan, Vivian S. Gainer, et al. 2015. "Development of Phenotype Algorithms Using Electronic Medical Records and Incorporating Natural Language Processing." *BMJ* 350 (April). British Medical Journal Publishing Group: h1885.

bioRxiv preprint doi: https://doi.org/10.1101/039651; this version posted February 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Lowe, Henry J., Todd A. Ferris, Penni M. Hernandez, and Susan C. Weber. 2009. "STRIDE--An Integrated Standards-Based Translational Research Informatics Platform." *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2009 (November): 391–95.

Maples, Brian K., Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. 2013. "RFMix : A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference." *American Journal of Human Genetics* 93 (2). The American Society of Human Genetics: 278–88.

Mark, Craven, and Page C. David. 2015. "Big Data in Healthcare: Opportunities and Challenges." *Big Data* 3 (4). online.liebertpub.com: 209–10.

Moses, Hamilton, David H. M. Matheson, E. Ray Dorsey, Benjamin P. George, David Sadoff, and Satoshi Yoshimura. 2013. "The Anatomy of Health Care in the United States." *JAMA: The Journal of the American Medical Association* 310 (18): 1947.

"MRI Accelerate Personalized Trx." 2014. *SAP News*. February 24. http://news.sap.com/medical-insights-project-from-sap-to-accelerate-personalized-treatment-for-cancer-patients/.

"MRI Demo." n.d. *SAP Innovation Center Network*.

"MyHeartCounts." n.d. https://med.stanford.edu/myheartcounts.html.

"[PDF]The Variant Call Format (VCF) Version 4.2 ... - Samtools." n.d. https://samtools.github.io/hts-specs/VCFv4.2.pdf.

Plattner, Hasso. 2013. *A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases*. Springer Publishing Company, Incorporated.

———. 2014. "The Impact of Columnar In-Memory Databases on Enterprise Systems: Implications of Eliminating Transaction-Maintained Aggregates." *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* 7 (13). VLDB Endowment: 1722–29.

Robinson, Peter N., Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. "The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease." *American Journal of Human Genetics* 83 (5): 610–15.

Tamang, Suzanne, Manali I. Patel, Douglas W. Blayney, Julie Kuznetsov, Samuel G. Finlayson, Yohan Vetteth, and Nigam Shah. 2015. "Detecting Unplanned Care from Clinician Notes in Electronic Health Records." *Journal of Oncology Practice / American Society of Clinical Oncology* 11 (3). jop.ascopubs.org: e313–19.

Wang, Guan, Kenneth Jung, Rainer Winnenburg, and Nigam H. Shah. 2015. "A Method for Systematic Discovery of Adverse Drug Events from Clinical Notes." *Journal of the American Medical Informatics Association: JAMIA* 22 (6). jamia.oxfordjournals.org: 1196–1204.

Weber, Griffin M., Kenneth D. Mandl, and Isaac S. Kohane. 2014. "Finding the Missing Link for Big Biomedical Data." *JAMA: The Journal of the American Medical Association* 311 (24): 2479–80.

Yang, Yaping, Donna M. Muzny, Fan Xia, Zhiyv Niu, Richard Person, Yan Ding, Patricia Ward, et al. 2014. "Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing." *JAMA: The Journal of the American Medical Association* 312 (18): 1870–79.

Zhu, Xiaolin, Slavé Petrovski, Pingxing Xie, Elizabeth K. Ruzzo, Yi-Fan Lu, K. Melodi McSweeney, Bruria Ben-Zeev, et al. 2015. "Whole-Exome Sequencing in Undiagnosed Genetic Diseases: Interpreting 119 Trios." *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, January. doi:10.1038/gim.2014.191.

Zoni-Berisso, Massimo, Fabrizio Lercari, Tiziana Carazza, and Stefano Domenicucci. 2014. "Epidemiology of Atrial Fibrillation: European Perspective." *Clinical Epidemiology* 6 (June): 213–20.