The COMET Sleep Research Platform

Deborah A. Nichols  
Stanford Sleep Medicine Center, Redwood City, CA, dnichols@cableone.net

Steven DeSalvo  
Stanford Sleep Medicine Center, Redwood City, CA

Ric Miller  
Stanford Sleep Medicine Center, Redwood City, CA

Darrell Jonsson  
Stanford Sleep Medicine Center, Redwood City, CA

See next pages for additional authors

Follow this and additional works at: http://repository.academyhealth.org/egems

Part of the Health Information Technology Commons, and the Health Services Research Commons

Recommended Citation
Nichols, Deborah A.; DeSalvo, Steven; Miller, Ric; Jonsson, Darrell; Griffin, Kara S.; Hyde, Pamela R.; Walsh, James K.; and Kushida, Clete A. (2014) "The COMET Sleep Research Platform," eGEMs (Generating Evidence & Methods to improve patient outcomes): Vol. 2: Iss. 1, Article 16.  
DOI: http://dx.doi.org/10.13063/2327-9214.1059  
Available at: http://repository.academyhealth.org/egems/vol2/iss1/16

This Model/Framework is brought to you for free and open access by the the EDM Forum Products and Events at EDM Forum Community. It has been peer-reviewed and accepted for publication in eGEMs (Generating Evidence & Methods to improve patient outcomes).

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.
The COMET Sleep Research Platform

Abstract

**Introduction:** The Comparative Outcomes Management with Electronic Data Technology (COMET) platform is extensible and designed for facilitating multicenter electronic clinical research.

**Background:** Our research goals were the following: (1) to conduct a comparative effectiveness trial (CET) for two obstructive sleep apnea treatments—positive airway pressure versus oral appliance therapy; and (2) to establish a new electronic network infrastructure that would support this study and other clinical research studies.

**Discussion:** The COMET platform was created to satisfy the needs of CET with a focus on creating a platform that provides comprehensive toolsets, multisite collaboration, and end-to-end data management. The platform also provides medical researchers the ability to visualize and interpret data using business intelligence (BI) tools.

**Conclusion:** COMET is a research platform that is scalable and extensible, and which, in a future version, can accommodate big data sets and enable efficient and effective research across multiple studies and medical specialties. The COMET platform components were designed for an eventual move to a cloud computing infrastructure that enhances sustainability, overall cost effectiveness, and return on investment.

**Acknowledgements**

Funding provided by 1R01HS019738-01 from the Agency for Healthcare Research and Quality (AHRQ) – PI Clete A. Kushida.

**Keywords**
comparative effectiveness research, COMET, research platform, obstructive sleep apnea, business intelligence, cloud computing

**Disciplines**
Health Information Technology | Health Services Research

**Creative Commons License**
Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License.

**Authors**
Deborah A Nichols, Stanford Sleep Medicine Center, Redwood City, CA; Steven DeSalvo, Stanford Sleep Medicine Center, Redwood City, CA; Ric Miller, Stanford Sleep Medicine Center, Redwood City, CA; Darrell Jonsson, Stanford Sleep Medicine Center, Redwood City, CA; Kara S Griffin, Sleep Medicine and Research Center, Chesterfield, MO; Pamela R Hyde, Stanford Sleep Medicine Center, Redwood City, CA; James K Walsh, Sleep Medicine and Research Center, Chesterfield, MO; Clete A Kushida, Stanford Sleep Medicine Center, Redwood City, CA.

This model/framework is available at EDM Forum Community: http://repository.academyhealth.org/egems/vol2/iss1/16
The COMET Sleep Research Platform

Deborah A. Nichols, MS; Steven DeSalvo, MM; Richard A. Miller, MA; Darrell Jónsson; Kara S. Griffin, MA; Pamela R. Hyde, MA; James K. Walsh, PhD; Clete A. Kushida, MD, PhD

Abstract

Introduction: The Comparative Outcomes Management with Electronic Data Technology (COMET) platform is extensible and designed for facilitating multicenter electronic clinical research.

Background: Our research goals were the following: (1) to conduct a comparative effectiveness trial (CET) for two obstructive sleep apnea treatments—positive airway pressure versus oral appliance therapy; and (2) to establish a new electronic network infrastructure that would support this study and other clinical research studies.

Discussion: The COMET platform was created to satisfy the needs of CET with a focus on creating a platform that provides comprehensive toolsets, multisite collaboration, and end-to-end data management. The platform also provides medical researchers the ability to visualize and interpret data using business intelligence (BI) tools.

Conclusion: COMET is a research platform that is scalable and extensible, and which, in a future version, can accommodate big data sets and enable efficient and effective research across multiple studies and medical specialties. The COMET platform components were designed for an eventual move to a cloud computing infrastructure that enhances sustainability, overall cost effectiveness, and return on investment.

Introduction

Health care is currently experiencing a major shift in the way data are collected and utilized. With advances in technology, the opportunity for clinicians, allied health professionals, researchers, and patients to collect health-related information has grown exponentially. The key is to confirm that these data are converted to meaningful outcomes. To properly identify which outcomes are the most significant, greater importance is being placed on obtaining input from numerous individuals with diverse areas of expertise and perspectives. Similarly, we felt that input from a broad base of stakeholders would be the best way to objectively determine which features or requirements were most important for our primary function: designing an informatics platform to support clinical research.

This manuscript comprehensively describes the Comparative Outcomes Management with Electronic Data Technology (COMET) platform, including its three layers and two subsystems. Multiple examples are provided with data from a comparative effectiveness trial (CET); however, it should be stressed that the platform is designed to be repurposed within other research disciplines. An overview of concepts are provided that can be adapted by project managers for other intentions, with sufficient details for technical implementation by informatics experts.

Background Analysis for Platform Development

In a requirements analysis performed with input from over 40 stakeholders, including principal investigators, research staff, biostatisticians, clinicians, and informaticians, similar components were found to be desired across most clinical trials. A total of 18 essential components (listed in Table 1) were identified to address the technical, clinical, and research needs for an informatics platform and became the conceptual foundation for the COMET research platform.

The COMET sleep research platform development was funded with a grant provided by the Agency for Healthcare Research and Quality (AHRQ) with a primary goal of creating a new platform for the effective use of electronic data in comparative effectiveness research (CER).1

The secondary goal of the AHRQ project was to complete a CER trial to test the efficacy of the newly developed platform. Thus, CET, a randomized multicenter research study was implemented to compare the effectiveness of two treatments for overweight, hypertensive obstructive sleep apnea patients: positive airway pressure (PAP) versus oral appliance (OA) therapy. Four sites were involved...
in CET including Stanford University, Harvard University, the University of Pennsylvania, and the University of Wisconsin–Madison. Medical specialties represented included psychiatry, neurology, pulmonology, cardiology, and sleep.

During the early development phase for the platform, a new team of stakeholders—including principal investigators, research staff, clinicians, nurses, project managers, and developers—individually identified requirements that were necessary to perform CET. This was a crucial test to confirm that the conceptual development roadmap using the 18 components enabled all necessary requirements. Upon review, it was determined that implementation of the 18 COMET components would support all the desired CET requirements, and that agile software methodology development (see “Agile software development” in the COMET Platform Layers section; also see “Agile Methodology” in Appendix D: Glossary of Terms) should begin immediately following this framework.

Currently, the COMET sleep research platform is a standardized platform for collecting, organizing, storing, analyzing, presenting, and sharing research data from multicenter clinical research trials. This platform was designed with a focus on stability, standardization, and scalability so that both smaller and larger projects can be accommodated. Close attention to cost effectiveness, ease of use for researchers, reusability by other research projects, and governance were also prioritized. This platform is further enhanced by applying business intelligence (BI)\(^2-5\) to the COMET project, which are approaches first tested in the business environment that was flexible enough to accommodate diverse multicenter clinical research studies. Referencing a common expandable toolset creates a foundation for rapidly building and deploying varying research studies, thus reducing overall costs and increasing return on investment (ROI) by providing access to quality electronic data in a timely manner for analytic purposes. With regard to the future of hosted systems, the intent for the COMET platform is to facilitate a move of the platform or component parts of the platform to cloud computing services in future versions of the platform.

### Developing the COMET Platform

Several technology platforms were evaluated to identify the best fit for the COMET platform based on input from numerous stakeholders on what requirements are most necessary to perform research studies. Ultimately, the team chose a broad family of servers and other technologies that are interoperable to allow selection of only the components required for a particular research project. Not all COMET components need be selected for every trial.

The Microsoft family of servers and development products were selected for the COMET infrastructure in order to minimize software development costs, to allow for the integration of products, and to have the future ability to move to cloud computing. Cloud computing is the practice of using multiple remote servers hosted on the Internet to store and process data instead of local servers or a personal computer. To facilitate the latter design goal, the COMET platform comprises three technology layers that map to the cloud computing service model\(^6\) (1) Infrastructure Layer, (2) Server Software Layer, and (3) Application Software Layer (Figure 1). The characteristics of the COMET design are outlined in the following section in addition to the technology solutions that were ultimately selected.

### COMET Platform Layers

**Figure 1. COMET Platform Layers**

This system has the potential to improve health care by providing a more efficient method for integrating research projects across multiple sites and multiple research disciplines. Rather than repurposing existing research platforms to fit new research needs, or designing whole new research platforms each time a study is funded, we designed COMET to provide an infrastructure that was flexible enough to accommodate diverse multicenter clinical research studies. Referencing a common expandable toolset creates a foundation for rapidly building and deploying varying research studies, thus reducing overall costs and increasing return on investment (ROI) by providing access to quality electronic data in a timely manner for analytic purposes. With regard to the future of hosted systems, the intent for the COMET platform is to facilitate a move of the platform or component parts of the platform to cloud computing services in future versions of the platform.
The Infrastructure Layer

The infrastructure layer provides the foundation (hardware, network, and storage) for the platform, similarly to how the foundation of a house is built using a concrete slab and footings. Desired features of the platform include reliability, security, scalability, extensibility, and sustainability—features that are comparable to features of many informatics infrastructures. These features produce value by delivering a dependable, industrial-strength infrastructure that optimizes computing performance and has the capacity to accommodate future growth. Currently, the COMET platform is located at a secure data center at Stanford University, which enhances our network security. The system includes three Dell rack-mount host computers with a dozen virtual machines (VMs). For a more detailed view of the COMET platform, see Appendix B.

Server Software Layer

The server software layer includes all the computer programs that are installed on the infrastructure layer. The elements used in the infrastructure layer have impacts on the overall computer performance metrics, including software application response time. The COMET infrastructure is running Microsoft Windows Server 2008 R2 and 2012 as server operating systems (OS) on which all of the server software applications run as listed in Table 2.

By selecting a solution that is based on the technologies from one vendor, Microsoft, this provided a level of integration, interoperability, and time to delivery that we believe is worth the cost in relation to less expensive technology offerings and open source solutions that often do not provide the level of support required. In addition to interoperability of components, some of the key components such as MS SQL Server and MS SharePoint come in various editions that roughly map to different levels of feature sets, performance, and cost (Table 3). By identifying optional interoperable components in various editions, researchers have fine grained control to select the technology that fits the project in terms of feature set, scalability, cost, and duration of use. Selection of data presentation features is one example of the granular control researchers have in making technology choices to match their needs.

An example of a low-cost, self-service BI solution may include a single server housing a database server (SQL Server) and a web server (IIS). Research team members use client-based tools (Excel) to visualize data with multidimensional analysis and pivot tables.

A higher cost, team-based BI solution may add two servers including a directory service for user authentication (Active Directory) and a collaboration and presentation subsystem (SharePoint). In addition to hosting Excel PowerPivot tables, this configuration provides for more data visualization features (dashboards, scorecards, database reports) in a secure, permissioned environment. Furthermore, team members only need to enter their personalized username and password; then they can use any browser on any OS to access and interact with the data from any internet connected location. This eliminates the need for the installation of additional software on the team member’s local computer.

Table 2. Server Software Applications: Versions, Features, and Costs

| Software Component | Edition | Feature Set | Cost |
|--------------------|---------|-------------|------|
| MS SQL Server      | Express | Limited     | Free |
|                    | Standard| Basic       | Low  |
|                    | BI      | Business intelligence | Medium |
|                    | Enterprise | Full featured | High |
| MS SharePoint      | Foundation | Limited     | Free |
|                    | Online  | Full featured | Low  |
|                    | Server  | Full featured | High |
Application Software Layer

The application software layer is dependent upon the server software layer to execute tasks with the infrastructure layer hardware. The application software layer enables developers to design computer programs that perform tasks specified by the end user.

Agile software development. The agile software development methodology is used for the development of applications in this layer. The agile method is an iterative approach by a collaborative team of developers, knowledge engineers, and business analysts to introduce successive improvements through collaboration that ultimately produces a sustainable model that decreases costs and increases quality improvements while maintaining usability. This methodology allows the project team to focus the majority of the development effort on the solution rather than on theoretical proposals or overly detailed software documentation. The development team members are listed in Table 4.

Table 4. COMET Sleep Research Development Team

| Team Role              | Members                        |
|------------------------|--------------------------------|
| Stakeholders           | Principal Investigator, Project Manager |
| Information Technology (IT) | Director of IT, Developers          |
| Content                | Content Specialist, Knowledge Engineer |
| Ontology               | Ontologist                      |
| Business Management    | Business Analyst, Research Manager |

The agile methodology is especially useful in developing data visualizations such as pivot tables, dashboards, scorecards, and database reports where data must be presented in a simple, self-evident fashion while at the same time being information dense. It is extremely difficult to create a seemingly simple, insightful graphic representation of complex data without an iterative approach such as agile that involves input from a diverse team.

Automating data processing such as receipt, messaging, and processing of device data also requires an agile development process. Frequently device interchange files are not well documented and are sometimes ambiguous. Errors in hand annotations must also be detected, documented in a clear manner for correction, and ultimately received, interpreted, cataloged, and imported into the study database. Numerous iterations in an agile process involving clinicians, content specialist, database administrators, and programmers must be completed to create comprehensive rules that are properly coded for accurate and automated operation. The agile creation of a comprehensive library of rules, messages, and data structures of standard medical device interchange files are also a valuable contribution of the COMET platform.

Selection of development languages. Development languages were selected for rapid development, optimization, and ease of maintenance. Technologies such as Active Server Pages (ASP), Hyper Text Markup Language (HTML), and Java Script (Jscript) are used for browser-based web forms. Microsoft Access 2010, Microsoft SQL Server Management Studio 2008 R2, and Structured Query Language (SQL) are used for data manipulation within the SQL Server and MySQL Databases. Related statements to a given task or process are combined into stored procedures for ease of maintenance and integration with the Windows Task Manager and COMET report scheduling engines. Microsoft Visual Studio is utilized for the creation of code for data file imports utilizing the Microsoft C# (that is, C “Sharp”) language and Visual Basic Scripting (VBScript) languages. The change control approach for software development is used so code is accurately versioned and documented. The Microsoft Business Intelligence Development Studio 2008 R2 and 2012 is used to create the multidimensional cubes from the SQL Server star schema designs for fast compilation of results into relevant outcomes utilizing predefined dimensions for data slicing.

Table 4. COMET Sleep Research Development Team

Application Software Layer Subsystems

At the highest level, CET data management requires two subsystems for the Application Software Layer: (1) a data pipeline subsystem for content collection and processing of data into data set results, and (2) a collaboration and presentation subsystem to provide for a consolidated location for investigators to access, view, and query data ultimately leading to reporting and analysis (Figure 2).

Data Pipeline Subsystem of the Application Software Layer

The content collection processes for CET had significant requirements that were important to consider early in the development process. These core requirements to accommodate multiple data inputs (federated data, web forms, and device data) and the solutions fulfilled for CET are discussed in detail in Appendix C.

Any data system that regularly accumulates and transforms validated data to desired output formats demands a data pipeline subsystem that implements extract, transform, and load (ETL) features with some level of automation. Automated services are central to the COMET platform and enable recurring tasks to be regularly and dependably repeated with detailed tracking logs to minimize the need for human intervention during this processing (Figure 3).

The automated COMET ETL processes include reliable scheduling of data acquisition, quality control, and cleansing, as well as ongoing data verification during the clinical trial and research period. The automation features support archived reports documenting data exceptions, exclusions, missing data, and errors while providing mechanisms to manage unhandled or unrecognized data, minimizing the need for manual intervention.
There are three platform components that support these automated and repeatable processes: (1) Windows Task Manager and WS_FTP; (2) Extract, Transform, and Load; and (3) The COMET Report Engine.

Windows Task Manager and WS_FTP
Windows Task Manager is used to schedule the execution of code to upload the various files collected from both clinical centers and scoring centers into the COMET Data Coordinating Center (DCC) platform. Differential scheduling needs including execution at fixed times, varying dates, or varying intervals can be configured. On a scheduled basis, the folders are checked for the upload of file sets at the four research sites and WS_FTP is then used to host the transfer of files to the DCC at Stanford for further processing. Additionally, all the valid file sets are archived upon transfer, noting both the original upload location in addition to the archival location to accurately log the file upload history. The log also notes all validation processes including passes and failures.

The COMET platform is extensible, and is able to accommodate diverse content, including file sets with different components containing an array of files with various formats. It supports raw data file types including the following: Standard CSV/American...
Standard Code for Information Interchange (ASCII), ASCII/Nonstandard data files, and Non-ASCII proprietary data files. The varying file types uploaded can include alphanumeric data files, instrument data files, and more complex file types such as image and video files.

**Extract, Transform, and Load (ETL)**

Extraction requires that the files collected from Windows Task Manager scheduled jobs and web forms are consolidated into the COMET SQL Server database at Stanford. Scheduling of automated tasks is flexible (e.g., daily, weekly, or monthly) to support: (1) ongoing administration of studies; (2) regularly scheduled data refresh cycles to approximate real-time status updates (e.g., recruitment progress); and (3) systematic scheduling, accumulation, transformation, and presentation of data in a timely format. Transformation includes a documented series of reproducible steps that are taken to insure the data are validated for correct values, duplicates are removed, missing values are flagged and corrected, and scoring algorithms and derived variables are calculated.

Another critical step during the transformation is to harmonize disparate data. For COMET, a physical data model (PDM) is utilized where detailed study-specific data dictionaries associate metadata to content. Metadata include information related to the structure, formatting, creation, naming conventions, purpose, and the location of all elements in the data dictionary. These data can be mapped to data from other systems given the separate data model is well documented. The processed data is then loaded into table structures designed to better accommodate the type of data uploaded and ultimately facilitate visualizations and data set extractions.

**COMET Report Engine**

The COMET report engine is programmed to run at scheduled intervals (hourly, daily, weekly) throughout the study, transforming data with scoring and processing algorithms, while simultaneously archiving progress over time. Health Insurance Portability and Accountability Act (HIPAA) and protected health information (PHI) compliance procedures occur at this stage using random renaming functions to de-identify unique patient identifiers. SQL Server Agent facilitates scheduled weekly runs of processing algorithms and report generation. The report engine design allows point-in-time views of study data (beginning with recruitment) to monitor relevant statistics for study management.

**Collaboration and Presentation Subsystem of the Application Software Layer**

The COMET Collaboration and Presentation Subsystem is accessible through the SharePoint infrastructure, which provides permissioned access to information and a common area for communication and collaboration. Collaboration features, dashboards and scorecards, and reports and analysis components of the subsystem are described and demonstrated below.

**Collaboration Features**

The collaboration features include document libraries, calendars, contacts, task lists, workflows, and permissioned work spaces. Document libraries allow for easy storage and retrieval of files related to the ongoing development of the COMET platform as well as those documents related to the clinical research and CET. Calendars manage conference calls and the associated materials discussed during the calls. Contacts facilitate communication between team members. Task lists track progress to date and deadlines for uncompleted tasks by research site and task owners. Workflows allow for the automation of tasks designed to save staff time.
time. These collaboration features permit check-in and check-out of documents with version control enabling multiple individuals to provide input to a given document. All features are permissioned for accessibility based on a user’s role in the study.

Dashboards and scorecards. A SharePoint website—for presentation of relevant, real-time data during the study recruitment period—allows the clinicians, clinical coordinators, research staff, and centralized scoring center staff to access information related to study management. All of the sites are permissioned for accessibility based on a user’s need for access and study role. For example, a three-column dashboard created for research staff involved in the processing of 24-hour ambulatory blood pressure monitoring (ABPM) data provides the status for this particular outcome. The first column provides more static summary information including the study workflow and key team members (with links to contact information). The second column provides access to live key performance indicators (KPI), e.g., ABPM files transferred by visit; live data reports, e.g., the number of files transferred and scored by clinical center; and a document library that provides links to study protocols, manuals, templates, references, and other pertinent files. The third column includes links to collaboration features such as calendars and task lists. This graphical user interface (GUI) was designed to be intuitive (more general information to the left, more detailed information to the right) and to present a balance of visual information (e.g., graphs and illustrations) and textual information.

Reports and analysis. While BI and multidimensional analysis (MDA) are used primarily for competitive advantage for businesses in industries such as retailing or banking, BI can also allow an effective review and more efficient analysis of vast amounts of clinical and research data in health care. The use of BI and MDA within COMET enables easier, more insightful querying of big data sets and integration of data across disparate research systems. With MDA, individual researchers can rapidly and effectively answer questions of interest using real-time data. By providing user-driven experiences, MDA decreases staff hours necessary to run custom queries or write and rewrite customized reports. The elements of BI in the COMET platform include (1) The COMET Data Warehouse; and (2) Microsoft SharePoint, Access, and Excel.

The COMET Data Warehouse. The COMET data warehouse comprises a series of processed and transformed outcome variables collected during a study with a prespecified set of specific aims. These outcomes are organized into sets of star schema design data sets containing specified study outcomes in measures and dimensions for slicing. These data sets are then utilized to create multidimensional cubes that can be used to better visualize data using BI solutions. BI can present the measures of interest for a given study or group of studies and can quickly and efficiently slice data using relevant dimensions. For CET, one of the multidimensional cubes focused on the primary and secondary cardiovascular outcomes including variables derived from 24-hour ABPM and vascular ultrasound devices. Multidimensional cubes can also be de-identified.

Microsoft SharePoint, Access, and Excel. While the study is ongoing, progress reports can be generated and displayed using SharePoint, Access, and Excel. Excel Power Pivot and Excel Power View can be used as the BI front-end tools to provide a rich user-driven experience of the data accumulated on a robust research platform that enables easy access to these tools. Figure 4 shows one of the dashboards accessible by permissioned users to select, view, and interact with a series of Excel Power Pivot Reports.

Figure 5 is a screenshot of one of the Power Pivot visualizations used during CET to monitor the number of web forms entered by study staff. It depicts three filters that allow the user to view the number of web forms entered after selecting various combinations of study visit, patient randomization status, and site. Data in tables and graphs update immediately as different filter options are selected. Final study reports can also be constructed using these BI tools.

Discussion

Frequently, informatics tools for clinical research are created at the time of the award of a grant and commencement of clinical trials. Tools created may fulfill one or more purposes for the trial, but may be limited for future applications. Many times, when the grant is over, development of one or more tools ceases and may or may not remain available due to the costs to keep the infrastructure and staff funded. Performing a comprehensive requirements analysis prior to development with a diverse group of stakeholders will increase the generalizability and sustainability of informatics tools developed for a specific research protocol.

Implementation

In a requirements analysis phase performed prior to the commencement of COMET, we strove to identify key features that would most likely lead to a sustainable research infrastructure. Based on this input, development of the COMET platform focused on designing components that provide a robust, standardized, extensible, and scalable clinical research platform. The robust features of the platform include its use of industry standard OSs and servers with thorough third-party documentation and ample avenues for technical support. Utilizing products from a single vendor reduces overall development costs due to the established integration of various server products; databases; and applications with features for governance, and data-access and use permissions. Standardization of table structures, data imports, and use of a data dictionary with associated metadata provides the means to share data across different institutions. The normalization of key reference fields from medical devices provides a generalizable mapping structure to easily accommodate multiple devices. Extensible applications developed using agile processes are built to be data driven, allowing the flexibility to support diverse content as study needs arise.

The team chose a broad family of servers and other technologies, both open source and commercial software, which are interoperable to allow selection of only the components required for a particular research project. Not all COMET components need be
selected for every trial. As COMET addresses a superset of clinical research needs, use of a commercial software provider is warranted because of the breadth and depth of features required (e.g., extensibility) that could not be economically provided by a small academic team. This approach helps ensure a platform that is sustainable into the future and provides the latest cutting edge tools on an integrated platform. Additionally, comprehensive training, support, and development resources are accessible worldwide.

We chose the Microsoft family of servers and development products to minimize software development costs, to allow for the integration of multiple products, and to reduce the time to delivery; and also to have the ability to move to a cloud computing service model in the future. By identifying optional interoperable components in various editions (ranging from freeware to Enterprise), researchers have a fine-grained control for selecting the technology that fits the project in terms of feature set, scalability, cost, and duration of use. The researcher can also choose to apply selected components to portions of the project or utilize the full end-to-end data management solution offered by the COMET platform. The reusable nature of the COMET Platform applications and application software layer (see Appendix B: COMET Infrastructure Detail View) will save time and money, increase data quality, and be easier to use for research than one-off partial solutions. Note that informatics solutions cannot address all potential barriers to data sharing endeavors, as there are institutional, regulatory, and governance issues that must be addressed in concert with informatics issues.

Governance
Several administrative requirements were also identified during the requirements analysis. For a multicenter clinical trial, governance issues such as approval by an Institutional Review Board (IRB) and monitoring by a Data and Safety Monitoring Board (DSMB) were important to address. We found it beneficial for the parent site to take the lead on the necessary IRB submissions and to obtain the initial IRB approval because this permits the creation of templates to aid the remaining sites with their individual IRB submissions. A primary function of the DSMB is to monitor patient safety and study metrics. Constructing queries to track and report on important metrics such as participant enrollments and randomizations in addition to participants who withdrew or dropped out (patient driven; pre- and post-randomization), were excluded or disqualified (staff driven; pre- and post-randomization) from the study provides the necessary parameters for reporting.

Strengths of the COMET Platform
The COMET platform is unique in its approach to analysis of research data with the implementation of MDA and BI for visualization and rapid access to data and results. The platform is scalable in its ability to accommodate data sets, by both visit and accumulation. Rather than build a new research platform each time a study is funded, COMET was designed to provide a general purpose infrastructure that is flexible enough to support diverse multicenter clinical research study management. Inclusion of a common expandable toolset creates a foundation for rapidly building and deploying varying research studies, thus reducing overall costs and increasing ROI. This system, while developed within a large academic center, could ultimately provide the greatest value to smaller research centers lacking access to the infrastructure and information technology (IT) staff required to implement a collaborative solution.
Limitations of the COMET Platform

As in all technology offerings, there is a team of informatics experts and business analysts behind the design of the COMET platform. While the COMET platform is generalizable to accommodate diverse study aims and extensible to span several medical disciplines, implementation for new research studies will still require the involvement of informaticians and business analysts. While the platform provides a unique tool for medical research, several more iterative deployment cycles are required to account for most, if not all, permutations of technical, clinical, and research requirements. With further attention given to documentation and automation, a future version of the platform could accommodate development by individuals outside of our existing team. In addition to the internal use by study investigators, the COMET platform has implemented a public website (http://comet-sp.stanford.edu) with information on the COMET platform and potential uses by other organizations and research projects. This external public website can be used in the future for the dissemination of published research results and direct correspondence with other interested investigators.

Conclusion

The COMET platform has the capacity to unite clinical and research needs with IT to create an end-to-end data management solution. Future versions of the platform may accept additional types of user input and enhanced mechanisms for instrument data collection. It is envisioned that dynamic generation of clinical input mechanisms and their corresponding database tables will allow for the rapid inclusion of new and varied types of content.

The platform was designed to be flexible enough to accommodate requirements for many different studies. The platform's architecture can be scaled up to accommodate big data sets with greater complexity, automation, and dependence on Enterprise-level solutions or scaled down to provide specific features at much lower financial and manpower costs using free or low-cost components for basic research. The integration of technology early in the study start-up period greatly increases study efficiency and reduces the time and effort spent on manual processing. Furthermore, BI and MDA provide for an extensible model for analysis that can harmonize data from data warehouses across medical disciplines and multiple CER studies.

With the inclusion of MDA and BI, the platform could eventually be utilized for predictive analysis that can facilitate early detection and the implementation of prescribed protocols for disease prevention. Predictive analytics, a technology feature that finds patterns and trends within large volumes of data that may predict outcomes, is included as part of the Microsoft Business Intelligence products contained within the SQL Server Analysis engine. With the proper input from clinicians, knowledge engineers, and biostatisticians, predictive analysis has great potential for use in health care systems to predict disease outcomes, causes, and yet-unknown medical correlations that may exist in the data.

The ultimate model for sustainability may be to plan for a future move to cloud computing services, and the COMET platform or component parts of the platform can be moved to cloud computing services in future versions. The COMET platform uses a segmented, componentized architecture where the technology layers logically map to the cloud computing service model (NIST Definition) including Infrastructure as a Service (IAAS), Platform as a Service (PAAS), and Software as a Service (SAAS) (Figure 6). The benefit of moving the COMET architecture to the cloud architecture is the reduction of operational costs, particularly in the area of infrastructure and platform.

A robust community of users, developers, and supporters must evolve together with a collaborative workspace for ongoing contributions, concurrently with the technical and infrastructure iterative development of the COMET platform. A COMET “Hub” has been identified as such a place with support and funding currently being pursued.

Figure 6. Future Cloud Computing Service Model

![Cloud Computing Service Model](image-url)
This endeavor has the support of interested academic researchers from multiple institutions and industry leaders such as Microsoft. The website for the COMET Hub will launch in the winter of 2014 and will eventually provide hands-on access to the COMET platform and applications in a subscription-based cloud environment that allows collaborators to use key features, evaluate suitability for their own projects, contribute to the growth of the platform, and ultimately use the platform for their own medical research projects.

In conclusion, COMET is a research platform that is a generalizable and extensible infrastructure that, in a future version, will be able to accommodate big data sets and enable efficient and effective clinical research across multiple studies and medical specialties. The COMET platform components were designed for an eventual move to a cloud computing environment that significantly enhances overall cost effectiveness, and may be a potential key to sustainability for this and other platforms.

Acknowledgements
Funding provided by 1R01HS019738-01 from the Agency for Healthcare Research and Quality (AHRQ) – PI Clete A. Kushida.

References
1. Hamilton Lopez M, Holve E, Sarkar IN, Segal C. Building the informatics infrastructure for comparative effectiveness research (CER): a review of the literature. Med Care. 2012 Jul;50 Suppl:S38-48.
2. Luhn H P. A business intelligence system. IBM Journal of Research and Development.1958;2(4):314–19.
3. Codd EF. A relational model of data for large shared data banks. Communications of the ACM. 1970 Jun;13(6):377-87.
4. Kimball R, Ross M, Thornthwaite W, Mundy J, Becker B. The data warehouse lifecycle toolkit (2nd ed.) Wiley; 2008.
5. Mettler T, Vimarlund V. Understanding business intelligence in the context of healthcare. Health Informatics J. 2009 Sep;15(3):254-64.
6. The NIST Definition of Cloud Computing. National Institute of Standards and Technology. Retrieved 24 July 2011.
7. Schweitzer EJ. Reconciliation of the cloud computing model with US federal electronic health record regulations. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):161-5.
8. Beck K, et al. Manifesto for agile software development. Agile Alliance; 2001.
## Appendix A. Acronyms and Terms

| Acronyms | Terms |
|----------|-------|
| ABPM     | 24-hour Ambulatory Blood Pressure Monitoring |
| AD       | Active Directory |
| AHRQ     | Agency for Healthcare Research and Quality |
| ASCII    | American Standard Code for Information Interchange |
| ASP      | Active Server Pages |
| ASQ      | Alliance Sleep Questionnaire |
| BI       | Business Intelligence |
| CET      | Comparative Effectiveness Trial |
| CMS      | Content Management System |
| COMET    | Comparative Outcomes Management with Electronic Data Technology |
| CSV      | Comma Separated Value |
| DCC      | Data Coordinating Center |
| DMS      | Document Management System |
| DSMB     | Data and Safety Monitoring Board |
| EDF      | European Data Format |
| ESS      | Epworth Sleepiness Scale |
| ETL      | Extract, Transform, and Load |
| FTP      | File Transfer Protocol |
| GUI      | Graphical User Interface |
| HIPAA    | Health Insurance Portability and Accountability Act |
| HTML     | Hyper Text Markup Language |
| IAAS     | Infrastructure as a Service |
| IIS      | Internet Information Services |
| IRB      | Institutional Review Board |
| Jscript  | Microsoft dialect of Java Script (ECMAScript) |
| KPI      | Key Performance Indicator |
| MDA      | Multidimensional Analysis |
| NIST     | National Institute of Standards and Technology |
| OA       | Oral Appliance |
| OS       | Operating System |
| PAAS     | Platform as a Service |
| PAP      | Positive Airway Pressure |
| PDF      | Portable Document Format |
| PDM      | Physical Data Model |
| PHI      | Protected Health Information |
| POMS     | Profile of Mood States |
| ROI      | Return on Investment |
| SAAS     | Software as a Service |
| SAQLI    | Calgary Sleep Apnea Quality of Life Index |
| SQL      | Structured Query Language |
| VBScript | Visual Basic Scripting |
| VM       | Virtual Machine |
Appendix B. COMET Infrastructure Detail View

Application – End Users
- Study Management
  - System Administrators
- Project Management
  - Principal Investigators
- Patients
- Governing Bodies
- Funding Agencies

Applications
- Application Admin
  - Setup application
  - Register a study
- (CET) Comparative Effectiveness Trial
- Sleep Apnea Registry
- COMET Website
- Patient Portal

3) Application Software Layer
- Governance
  - Permissioned Use
  - Permissioned Access
  - HIPAA Procedures
  - Security Procedures
  - IRB Document Library
- Content
  - Documents & Data Libraries & Templates
  - Content Request
  - Data Collection
  - Data Pipeline
- Visualization
  - Analysis
  - Reports
  - Pivot Tables
  - Key Performance Indicators
  - Dashboards
- Language
  - Terminology
  - Taxonomy
  - Ontology
  - Content Type Definitions
  - Data Dictionary

2) Server Software Layer
- IT Management
  - Team Foundation Manager
  - Data Protection Manager
  - Operations Manager
- Security
  - Active Directory
  - Users & Groups
  - User Management
- Content Mgmt
  - Sharepoint Server
  - Web Server
  - Content Type Hub
  - Migration Tools
- Database Mgmt
  - SQL Server
  - MySQL Server
  - Analysis Server
  - Report Server
- Communications
  - Content Transfer
  - Email Server
  - FTP Server

1) Infrastructure Layer
- On Premise Hosted
  - 12 tier
  - Virtualized servers
  - Stand alone servers
- Hybrid Hosted
- Cloud Hosted
  - Office 365
  - Azure
  - Team Foundation Service

Platform – Development Team
- Stakeholders
- Information Technology
- Content
- Ontology
- Business Management
Appendix C. Comparative Effectiveness Trial (CET) Content Collection Requirements and Solutions

Per specification by the original grant application, the CET was the primary "Use Case" to guide the development and test the functionality of the infrastructure design for supporting a prospective multicenter research study. The content collection processes for CET had significant requirements that were important to consider early in the development process. Enabling diverse mechanisms (Federated Data, Web Forms, and Device Data) for collecting data was identified as a key requirement (Figure 7).

1. Federated Data.
The ability to collect federated data was listed as a content collection requirement for CET. Federated data originate from geographically dispersed and heterogeneous database systems, but these data are harmonized using a federated query. A federated query retrieves disparate data sets to return a result set that joins data in a new and meaningful way for comparisons and analyses. A federated approach was selected because it provides maximum institutional flexibility, while maintaining a high degree of sharable information.

CET demonstrated the collection of federated clinical data using four Oracle MySQL databases housed at each of the clinical centers. The Alliance Sleep Questionnaire (ASQ) is a branching-logic electronic questionnaire that was installed locally at each of the four clinical centers to demonstrate geographic dispersion and was then used to collect baseline clinical data directly from patients. The ASQ was selected for use in this study because it was developed by a consensus group of stakeholders from multiple institutions and various areas of expertise. It inquires not only about symptoms of obstructive sleep apnea, but also about other sleep disorders including insomnia, restless legs syndrome, and narcolepsy. On a nightly basis, data were automatically transferred to the DCC at Stanford while remaining separately hosted and permissioned. These data sets, synchronized to local data sets, were queried separately with the results integrated into the COMET database and, subsequently, the COMET report data warehouse, which were both housed on SQL Server. These procedures enabled federated queries to be performed across clinical centers.

2. Web Forms.
The collection of content using electronic web forms was also listed as a requirement for CET. Web forms are displayed through internet browsers using display languages such as Hyper Text Markup Language (HTML). The use of electronic web forms can increase data quality and staff efficiency, and by using web-based forms for data entry, end users do not need to download large applications to their computers.

Figure 7. CET Content Collection: COMET Data Pipeline Subsystem

Stanford
Harvard
UPenn
Wisconsin

Federated Data
Web Forms
Device Data

Content Collection
Data Processing
Data Set Results
During CET, patient observations, questionnaires, instrumentation data, and clinical data were collected from patient visits across the eight-month enrollment period and transferred to the centralized repository at the DCC using 57 web forms provided by the COMET platform. Not only do these forms include front-end data validations, they were also designed to rapidly enable the addition of new forms to the platform, thus providing an extensible platform for capturing longitudinal data. Examples of validated questionnaires for which web forms were used for data entry include the Epworth Sleepiness Scale (ESS), Profile of Mood States (POMS), and Calgary Sleep Apnea Quality of Life Index (SAQ-LI). In this version of the platform, web entry was performed by research assistants after collecting data from patients using paper forms. In a future iteration, direct data input by clinicians or research participants will be possible.

In order to demonstrate technologies that could be used by research staff to quickly enable direct data input, a prototype proof-of-concept web form for collecting clinical research data was created using Microsoft SharePoint, InfoPath Forms, and Nintex Automated Workflows. The goal was to create a mechanism that collected a limited set of specific clinical data, using validations to confirm data quality, which could ultimately provide useful summary data. Data entered by staff included blood pressure, weight, height, and other clinical measures. Microsoft SharePoint and InfoPath forms were used to enable research staff to design user-friendly web forms that could validate data fields prior to submitting the form. SharePoint lists, which are easily modified by research staff, stored valid values for various fields. Valid values are “looked up” from lists of answers for specific questions and are commonly called “look up tables.” Nintex Automated Workflows were used to initiate, track, and sign off on procedures with a goal of significantly increasing staff efficiency.

3. Device Data.
The final requirement for CET content collection was to interface with device data. Device data are collected from electronic instruments that measure one or more physiological functions and converts the information into meaningful outcomes for health care decision-making. Access to these data enables clinicians and allied health professionals to personalize and optimize diagnostic and treatment decisions for individual patients. For research, investigators can use raw data to create precise, well-defined outcomes.

Throughout CET, direct data feeds exported from various devices were consolidated and uploaded to the DCC from each clinical center. Email messages were also integrated to communicate file-set transfer status to the clinical centers. File sets were diverse in size, content, and file formats (e.g., CSV, PDF, Image files, proprietary formats). Many of the cardiovascular data sets, including 24-hour blood pressure monitoring and vascular ultrasound, were initially uploaded as a complex series of folders, image files, and data files. Upon transfer, these data sets required further back-end processing including centralized scoring and the consolidation of summary variables. Similarly, polysomnography (sleep study) data were exported with .REC (recordings in European Data Format [EDF]) extensions to enable centralized scoring using standardized software. Data sets that required further back-end processing by the DCC included the two devices that collected treatment adherence across time for PAP (Encore Anywhere: Philips Respironics, Inc.) and OA therapy (TheraMon: Handelsagentur Gschlagt, Hargelsberg, Austria). These longitudinal data sets were archived in their original formats and consolidated into the DCC database to create comparable variables that could be used for reporting and analysis.
Appendix D. Glossary of Terms

| Terms | Definition |
|-------|------------|
| Active Directory (AD) | A directory service provided by Microsoft that assigns and enforces policies on the network and authorizes all users and computers in the network domain. |
| Active Server Pages (ASP) | Are an internet presentation technology provided by Microsoft that enables a server-side scripting engine for generating web pages that can be viewed by a browser. |
| Agile Methodology | An iterative approach to software development where development teams focus more on requirements and developed solutions rather than on vast amounts of documentation and proposals. |
| Applications | Are individual collections of code that provide a specific business or end user need in a computing environment. Applications include everything from desktop word processors to advanced applications for task automation. |
| Business Intelligence (BI) | A set of techniques, methodologies, architectures, and technologies that transform raw data into meaningful and useful information for business purposes. BI can handle enormous amounts of unstructured data to help identify, develop, and create new opportunities for discovery and interpretation of large volumes of data in a user-friendly manner. Utilizing BI, one can discover new trends, opportunities, and correlations in the data that would otherwise go unnoticed. |
| Change Control | An approach to software development whereby changes to the code are tracked by a software development collaboration tool, and then changes are checked into the project by individual developers for inclusion in the product. |
| Cloud Computing | Is the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer to reduce overall costs utilizing a subscription model. Cloud computing divides into three models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). |
| Dashboards | Are a presentation style of relevant data used to present key performance indicators (KPIs) to show progress in areas of interest to a given business application and targeted audience. |
| Data Collection Schema | Is a database design that is implemented for import of data for further cleansing, aggregation, duplicate removal, and verification. This schema is typically used during the data transformation process to insure validity of the data collected. |
| Data Coordinating Center (DCC) | Is a centralized repository where study data are collected from various sources, often dispersed geographically, for ETL processing by the data pipeline and for housing the reports and analytical formats for final data presentation. |
| Data Pipeline | Is a subsystem where a series of interim steps occur including ETL, to bring data from its import source to its final destination format for reporting and analysis. |
| Data Warehouse | Is a database design that differs from the application database in that it is designed and optimized to respond to analysis questions that are critical for business applications. Data warehouses are also intended to house very large amounts of data and present analysis results very rapidly to end users. |
| De-Identification | Is a process used to comply with HIPAA regulations to protect the identity of patients and patient data. To achieve this, all patient IDs are re-generated, usually to a random number that then becomes the patient’s new identification within the software system. |
| Development Team | Refers to all of the people involved in the development of a business application solution. This includes the business analysts, developers, projects managers, testers, and any support personal related to the overall project delivery and success. |
| Device Data | Are collected from electronic instruments that measure one or more physiological functions and converts the information into meaningful outcomes for health care decision-making. |
| Excel Power Pivot | Is an add-on to the Microsoft Excel application that allows for presentation of multidimensional data (dimensions and facts) from multidimensional cubes. |
| Extract, Transform and Load (ETL) | Refers to the process of transforming data into a schema that is more appropriate for reporting and analysis. The final objective is to be loaded into a database for reporting or creation of multidimensional cubes that enable BI solutions. |
| Federated Data | May originate from geographically dispersed and heterogeneous database systems, but these data are harmonized using a federated query. |
| Federated Query | Is the ability to query data across geographically dispersed data sets to return a result set that joins the data in new and meaningful ways for comparisons and analysis. |
| File Transfer Protocol (FTP) | Is a network protocol on a Transmission Control Protocol (TCP) and Internet Protocol (IP)—commonly referred to as “TCP/IP”—network infrastructure used to transfer files between computers. |
| Graphical User Interface (GUI) | Allows end users to view electronic information using more visual or graphic presentations of data. It is often designed with a goal of enhancing the user experience. |
| Hyper Text Markup Language (HTML) | Is a standardized system for tagging text files to achieve font, color, graphic, and hyperlink effects on World Wide Web pages to be presented in standard web browsers. |
| Infrastructure as a Service (IaaS) | Is one of the three models provided by the cloud computing model. The IaaS provides resources such as virtual machines, disk image libraries, file-based storage, firewalls, load balancers, Internet Protocol (IP) addresses, virtual local area networks (VLANs), and software bundles. |
| Java Script (Jscript) | Is an object-oriented, computer programming language commonly used to create interactive effects within web browsers. |
| Key Performance Indicator (KPI) | Is a visual representation that is used to evaluate the status of a well-defined question. For research, study status or efficiency of a certain component may be presented. |
| Longitudinal Data | Are data that are viewed over a period by connecting the data to the time of entry or import so that changes to the data can be observed over time. In a research study, data are collected across multiple visits with prespecified time points. |
| Metadata | Contain specific information related to the structure, formatting, creation, naming conventions, purpose, and location of data elements. |
| Multidimensional Analysis (MDA) | Refers to the process of summarizing data across multiple levels of data called “dimensions,” and then presenting the results in a multidimensional grid format. This process is also referred to as OLAP, Data Pivot, Decision Cube, and Crosstab. Microsoft utilizes Excel with Power Pivot to provide MDA support. |
| Multidimensional Cubes | Are views into large amounts of data utilizing SQL Server Analysis Server. The database design is a star schema of dimensions and fact tables to present data for analysis by the Excel PowerPivot front-end tool for multidimensional analysis. |
| Operating System (OS) | Is the software system that is installed on hardware servers and desktop computers to provide an environment for running applications and browsers and to enable connecting to the internet and network infrastructures. |
Ontology is a structured hierarchy of information within a given domain designed to establish a common understanding of terms that enables the sharing of information.

Platform as a Service (PAAS) is one of the three models provided by the cloud computing model. The PAAS is provided as a computing platform, typically including OS, programming-language execution environment, database, and web server.

Predictive Analysis is part of the Microsoft SQL Server Analysis Server technology that is used to analyze known facts in order to make predictions about future events. The analysis is run on large volumes of data and looks for trends within the data. Predictive analytics features are customizable within the Microsoft Business Intelligence suite of products.

Return on Investment (ROI) is a term often used in business to quantify the benefits (or profits) of a given project given the cost of the initial outlay.

Requirements Analysis is the process in software design where the stakeholders are questioned in order to gain understanding of all of their requirements for the system design. The success of the final software product is highly dependent on the complete understanding of the requirements gathered during this phase.

SQL Server Agent is a scheduling tool included with Microsoft SQL Server that allows for the scheduling of tasks and scripts against the SQL Server databases. Jobs are scheduled and then logged as successes or failures.

Software as a Service (SAAS) is one of the three models provided by the cloud computing model. The SAAS provides access to application software and databases that run the applications.

Stakeholders are a group of individuals with a wide array of expertise and diverse backgrounds who provide focused input with a goal of optimizing outcomes. For COMET, stakeholders included principal investigators, research staff, biostatisticians, informaticians, developers, clinicians, researchers, nurses, and project managers.

Star Schema is a database design methodology utilized in BI systems that de-normalizes the data to facilitate the presentation of data for research and analysis. It is called “star schema” design because it consists of a central fact table surrounded by dimension tables (the points of the star).

Structured Query Language (SQL) is a software development language utilized to access, change, present, and manipulate data within a database.

Virtual Machines (VMs) are software-based emulations of a single computer. In a server or hosted environment, one computer can provide the ability to have multiple VMs running concurrently for separation by business logic and computing tasks performed.

Visual Basic Scripting (VBScript) is an Active Scripting language developed by Microsoft that is modeled on Visual Basic. It is designed as a “light-weight” language with a fast interpreter for use in a wide variety of Microsoft environments.

Web Forms are forms displayed through browser interfaces connected to the internet using display languages and technologies such as Hyper Text Markup Language (HTML). Common examples of browsers used to display web forms include Microsoft Internet Explorer, Google Chrome, Mozilla Firefox, and Apple’s Safari.

Windows Task Manager is used to display the programs, processes, and services that are currently running on a Microsoft Windows computer. Task Manager monitors a computer’s performance, and it can be used to close a program that is not responding.

WS_FTP is an FTP client produced by Ipswitch, Inc. for the Windows OS.