# A Strategy for Deploying Secure Cloud-Based Natural Language Processing Systems for Applied Research Involving Clinical Text

David Carrell
Group Health Research Institute
carrell.d@ghc.org

## Abstract

Natural language processing (NLP) of clinical text offers great potential to expand secondary use of high-value electronic health record (EHR) data, but a barrier to adopting NLP is the high total cost of operation, driven mainly by the costs and limited availability of technical personnel in applied health research settings. To overcome this barrier we propose a cloud-based service systems model by which entire NLP systems deployed in the cloud are cloned and provided to the adopting institution for their exclusive and unlimited use. Useful algorithms that perform various information extraction and classification tasks are built in to the NLP system. A rationale and model for cloud-deployed NLP is presented and the inherent data security and patient privacy issues it raises addressed. Both technical and socio-institutional security issues are discussed in the context of the unique challenges associated with processing highly regulated clinical text in an unconventional computing environment. Results of a June 2010 survey of Institutional Review Board (IRB) managers in applied research settings are presented. Survey questions address IRB managers' readiness to approve research projects involving cloud-based NLP technologies. Useful next steps and information needs are presented in the conclusion.

## 1. Introduction

Natural language processing (NLP) of clinical text offers great potential to expand high-value secondary use of electronic health record (EHR) data.[1] This is evidenced by recent requests for proposals, including the Office of the National Coordinator for Health Information Technology's Strategic Health IT Advanced Research Projects (SHARP) initiative, which is advancing NLP methodologies to extract comparable measures from heterogeneous EHRs nationwide.[2] The National Cancer Institute (NCI) is similarly promoting the use of NLP to introduce efficiencies into its Surveillance, Epidemiology and End Results (SEER) cancer registry program.[3] The Electronic Medical Records and Genomics (eMERGE) consortium uses NLP to extract data from EHRs for genome wide association studies (GWAS), and plans to expand its consortium to include other research institutions in the near future.[4]

One of the perennial challenges in applying NLP to clinical text is the disjunction between the location of the clinical text and the location of state-of-the-art deployments of NLP systems to "mine" or process the text. Proposals to transmit large quantities of clinical text outside the local institution to another institution with NLP processing capacity are widely considered infeasible. Attempts to bring NLP capacity to the text are more feasible but also difficult; local deployment of complex systems can be costly and requires technical expertise generally uncommon in the health care and health research settings where most clinical text resides.

Release of the open-source UIMA/cTAKES NLP system in 2009 by IBM and Mayo Clinic[5] reduced the cost of NLP deployment but it remains a non-trivial undertaking. Beyond deployment costs, a dearth of high-level NLP expertise in applied settings is a significant barrier to developing the NLP algorithms that can perform useful tasks. For these reasons clinical NLP continues to flourish mainly in academic institutions with substantial biomedical informatics departments.

Our objective is to make useful NLP capacity available in applied research settings where traditional deployment and NLP algorithm development is either technically or economically infeasible. To accomplish this objective we propose a somewhat unusual case of a cloud-based service systems model. In a typical service systems model a service is purchased by and delivered to the customer. From the customer's perspective the *service* is what counts; the apparatus that performs or generates the service is behind the scenes and generally of no concern.

Our model is different. In it *the apparatus itself* is what is delivered to the customer, and it is delivered by cloning an entire virtualized system. After the apparatus is cloned, the customer owns and uses it exclusively to process their own clinical text using the information extraction algorithms built in to

1530-1605/11 $26.00 © 2011 IEEE

the system. Why deliver the apparatus instead of the service? The apparatus is delivered to give the adopter/customer maximal ability *to secure the patient health information they process*. This model of NLP deployment offers a mechanism for transcending place, making it possible for functioning NLP systems and clinical text to reside in the same computing space, under the control and ownership of a single institution.

Cloud computing is defined as "a model for enabling on-demand network access to a shared pool of configurable IT resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. It allows users to access technology-based services from the network cloud without knowledge of, expertise with, or control over the technology infrastructure that supports them."[6] The essential characteristics of cloud computing include on-demand self-service, ubiquitous network access, location independent resource pooling, rapid elasticity, and measured service. Users of cloud services may "rent" applications (referred to as Software as a Service or SaaS), ready-to use virtual operating systems (referred to as Platform as a Service or PaaS), or the underlying virtualized hardware (referred to as Cloud Infrastructure as a Service or IaaS).[6] Because cloud systems rely heavily on virtualization to implement rapid up- and down-scaling of resources, cloning is a widely available technique for replicating entire cloud-deployed systems—for one's own use or for sharing them with other customers of the same cloud provider (e.g., Amazon or Microsoft). Such cloning, we postulate, can dramatically reduce the cost of technology distribution in the realm of clinical NLP.

Previously published surveys of a wide variety of users and potential users of cloud-deployed systems indicate that security concerns remain the primary impediment to adoption of cloud-computing.[7] Evidence we have gathered from a survey of institutional review board (IRB) managers in health research settings indicates similar concerns, but also suggests opportunities for adoption exist.

While the cloned NLP deployment model we advance here is entirely under the control of the adopter/customer, it is still based in the cloud. This introduces data security issues, some of which are unique to the health care domain. Section 2 of this paper presents technical details of our "minimal risk" deployment scheme. Section 3 explores technical and socio-institutional aspects security when patient clinical text is involved, and concludes with a presentation of the results of our IRB managers' survey on the use of cloud computing for research involving patient clinical text. In section 4 we offer summary observations and suggest useful next steps.

## 2. A model for cloud-deployed NLP

The cloud-deployed NLP model we propose leverages 1) virtualization functions of public cloud services to vastly reduce the local burden of traditional local NLP deployment and 2) a deployment scheme that minimizes or eliminates many of the data security risks associated with cloud computing. In this section we describe the architecture of the system and its technical feasibility; later sections address implicit security issues.

### 2.1. Virtual local deployment of NLP

Deploying an open source NLP system such as UIMA/cTAKES requires installing and configuring multiple components (programming languages and development environments, database systems, dictionaries and ontologies, indexes, utilities, etc.). Customizing NLP processing pipelines for specific information extraction purposes requires integration of many additional custom-built components and/or tailored lexical resources. Programming costs associated with the deployment, customization, and maintenance of NLP systems accounts for a substantial portion of the total cost of operation and remains a key barrier to distribution of these technologies. The need for local NLP/informatics expertise to guide algorithm development is another major impediment to adoption, particularly in applied research settings where such expertise is scarce.

Cloud-based virtualization enables low-cost, low-effort, and highly reliable replication, or cloning, of entire software systems, including systems that entail custom-deployed operating systems (OS), firewalls, and applications built upon cloud-provided infrastructure (IaaS). The overall cost of such cloning is so low that it provides an efficient mechanism for propagating even relatively minor tweaks—not to mention extensive revisions— from a master system to its clones, simply by re-cloning the updated master. In this way virtualization is means of distributing, updating, and maintaining complex NLP systems.

Allowing institutions to obtain and use privately-owned deployments of NLP systems through cloning has the potential to dramatically reduce the cost of and other barriers to adoption. This is already the

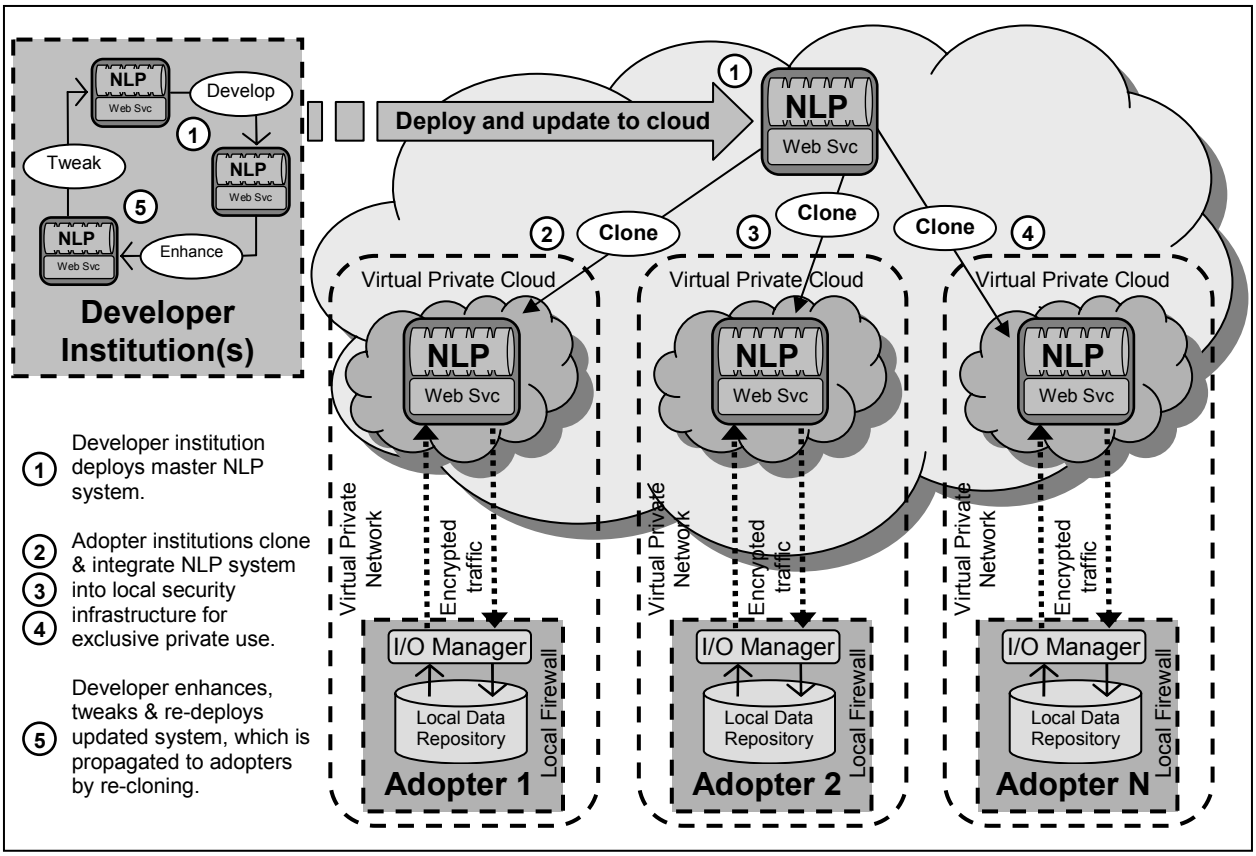


Figure 1. Schematic of propose model for cloud-deployed clinical NLP

case in the context of collaborative consortia such as the Open Health Natural Language Processing (OHNLP) Consortium,[8] the HMO Research Network,[9] the eMERGE consortium,[10] and the SHARP Program,[11] where expertise and programming resources are being deliberately marshaled for collective reuse.

The technological aspects of this model are quite straightforward and are depicted in Figure 1. NLP experts at one or a small number of collaborating institutions develop and deploy to the cloud a master copy of a complete NLP system, periodically updating the master as improvements or tweaks are made in the course of their regular development cycle of expansions and improvements. Prospective adopters of the system obtain permission from the owners of the master to create individual, private clones in their own privately controlled spaces in the same cloud. The adopting institution then integrates their clone of the system within their own security infrastructure, including VPN integration creating, as one cloud provider calls it, a "Virtual Private Cloud"[12] for their exclusive use. Further reducing the need for high-level NLP programming expertise, adopting institutions re-clone the master to avail themselves of updates and enhancements.

To interact with the cloud-deployed system the adopter/user deploys an open-source application (also developed by the NLP experts), that implements extract, transform, and load (ETL) services. This application encrypts and transmits local text, one document at a time, to a web service running in the cloned system, and receives back encrypted, NLP-annotated text and extracted data in structured format, which it then decrypts and stores in a local data repository. At this point use of the cloud-based clinical NLP system is complete. Researchers at the local institution then perform the analytical components of their investigations as they would with other forms of structured data.

The primary responsibilities of local IT in this model are the manipulation and management of clinical text prior to NLP processing, and the storage and manipulation of NLP-generated structured data after processing—both functions that IT staff in applied health research settings perform routinely.

The technology needed to facilitate this deployment model is clearly available now. However, because cloud computing is a novel technology it introduces a number of unfamiliar security issues, particularly when it involves protected health information (PHI) as defined in the

Health Insurance Portability and Accountability Act of 1996 (HIPAA).[13] In Section 3 we address these challenges in terms of risk mitigation procedures and policies and the opinions of Institutional Review Board (IRB) managers. The *designs* of cloud-based systems also affect risk exposure. In the remainder of this section we discuss system design elements that can be used to minimize risk exposure.

### 2.2. Minimal-risk design for cloud-based NLP

Because security concerns are paramount, risk reduction should be built in to the cloud-based system whenever possible, even at the expense of performance and functionality. We advance four key design elements intended to reduce cloud-related risk exposure: 1) non-persistence of data, 2) encryption, 3) VPN integration, and 4) consuming cloud services at the infrastructure level (IaaS).

Persisting data in the cloud, even if encrypted, introduces unnecessary risk exposures and is therefore avoided entirely in the proposed model. Instead, documents are received via a web service, processed one at a time *entirely in memory*, and returned to the local repository without storing any of the data or NLP-created annotations to disk—even temporarily. The slight performance hit this entails is well worth the advantage of not having to consider security issues related to cloud-based storage.

Encryption of sensitive data in transit is standard practice and routinely implemented without difficulty. We recommend using encryption even when the clinical text in question is entirely risk-free (i.e., contains no patient PHI). Again, the performance cost of encryption is worth the added protection in the event that identified data were mistakenly processed and this mistake coincided with a security breach of some kind.

An evolving and very attractive feature of cloud computing services is the ability to incorporate cloud-deployed resources into a virtual private network (VPN). Microsoft and Amazon, among others, offer such functionality. VPN-integration is a "win-win" in that it adds additional assurance that the system will be accessed exclusively by trusted/internal users, and eliminates other risk exposures such as externally launched intrusion attacks. The cost associated with VPN-integration is modest, primarily entailing additional local VPN and firewall configuration.

The security rational for building the NLP system on IaaS is that it gives greater control to system designers and users over security on many levels. Under the IaaS model the user is responsible for all systems deployed above the virtual machinery, including the OS and firewall. While security is necessarily a shared responsibility in any model of cloud deployment, with IaaS the responsibility of the cloud provider (e.g., Amazon) is limited to the physical data center and the virtualization layer.[14] This means there are fewer aspects of the overall deployment where users must "trust but verify" the cloud provider, whether first-hand or by third-party auditing service. Reducing the need for verification makes management of the deployed system more like traditional locally owned and operated infrastructure, though of course significant difference remain.

Building an NLP system on IaaS is more costly to developers compared to developing on PaaS because it requires management and configuration of more systems (e.g., OS and firewall). An IaaS-based system also entails additional costs to the adopter/user in the form of additional OS and firewall monitoring, whether. Nevertheless, we feel these costs should be absorbed in the interest of producing a system that maximizes opportunity for local control over security.

To begin exploring the mechanics of a cloud-based system we deployed components of a simple, proof-of-concept system within our institutional firewall and the Amazon cloud. It did not involve patient data of any kind. We used this system to explore connectivity, firewall, encryption, and other security issues. To date we have encountered no significant problems with this system, though further exploration with a complete deployment of the NLP system and a more extensive security apparatus is needed.

## 3. Achieving security in the cloud

While the proposed model appears feasible from a technical standpoint, meaningful adoption is unlikely unless local security concerns can be addressed. Constructively addressing these concerns, we believe, requires a holistic approach that recognizes both technical and socio-institutional aspects of institutional security infrastructure. The former includes encryption techniques, firewalls, intrusion detection schemes and auditing mechanisms; the latter is the domain of policy, governance, risk tolerance, historical experience and trust.

In this section we address both aspects of security with respect to adoption of cloud-based clinical NLP to support health research. Though the two aspects often overlap we focus first on technical matters and then on the socio-institutional issues. At the end of this section we present the results of our IRB manager survey.

### 3.1. Technical aspects of security

In traditional user-owned on-site computing settings where sensitive information is processed, IT departments generally implement standard or tailored models of security that that are well developed and understood. As technology and/or threats evolve these security practices are adapted. Cloud computing represents a disruptive technology that renders some important aspects of traditional security practice obsolete, creating the need for local IT departments to acquire new knowledge and experience.

An array of for- and non-profit resources and consulting services have emerged to address the resulting gaps in knowledge and experience. The non-profit Cloud Security Alliance is one such resource.[15] Cloud service providers also offer security resources and technologies tailored to their respective proprietary offerings.[16] Numerous for-profit consulting firms offer design, testing, and monitoring services. Industry observes note that cloud service "brokers," offering intermediation, monitoring, governance, provisioning, and integration consulting are emerging as an "important component in the overall cloud ecosystem."[14] In the near term, at least, we expect local institutions will need to rely on some level of external expert security consulting.

Though no standard cloud security solutions have yet emerged security experts agree on the basic steps the end-user must follow to identify and implement workable solutions:

1) Identifying the asset being considered for the cloud.
2) Determining the sensitivity of the asset to a security breach.
3) Mapping the asset to a cloud deployment model (e.g., private vs. public cloud, internal vs. external access).
4) Assessing readiness and comfort levels among key stakeholders and their risk mitigation requirements.
5) Evaluating cloud service providers (e.g., Amazon, Microsoft) and service level (SaaS, PaaS, or IaaS).
6) Considering special requirements for handling regulated data (e.g., HIPAA in the case of patient information).
7) Sketching potential data flows, highlighting specifics as to how data move into and out of the cloud, enumerating risk exposure points and audit requirements.
8) Implementing the necessary security practices, from governance to auditing.

This is a process each institution must undertake for itself. Nevertheless, the cloud-based clinical NLP use case suggests themes and issues we believe will surface in most settings. The extremely high value health care institutions place on patient privacy and the extensive federal and state regulatory requirements surrounding the use and stewardship of protected health information (PHI) raise salient issues relevant in all health care and health research settings. We now consider the above eight-step process from the perspective of a typical adopter institution.

In the case of cloud-based clinical NLP **the assets** (step 1) are electronic copies of clinical text generated in the normal course of delivering care to large numbers of patients over long periods of time within a single health care institution. These include physician chart notes, radiology reports, and pathology reports. As electronic *copies* of original electronic documents these assets have little or no intrinsic value. However, because clinical text often contains PHI (e.g., names, contact information, medical record numbers) there is an enormous imperative to avoid unauthorized disclosure, within as well as outside the institution. The asset in this case is thus best described as *clinical text which is securely managed throughout the course of its travel into, through, and back from the cloud-based system*.

It is difficult to overstate this asset's **sensitivity to security breach** (step 2). Disclosure, for any reason, is a violation of patient privacy, resulting in potential harm to individual patients and certain harm to the institution in the form of damaged reputation, costly and burdensome remediation obligations, and financial penalties under state and federal law, defined by HIPAA[13] and the Health Information Technology for Economic and Clinical Health Act (HITECH, passed in 2009 as part of ARRA).[17] Depending on the extent and prior history of a breach, fines may reach hundreds of thousands of dollars, costs which may be dwarfed by those resulting from damage to an institution's reputation.

Considering these security requirements a **suitable cloud model** (step 3) for clinical NLP, we believe, is likely to include the following:

a) public cloud hosting (to support adoption by geographically dispersed institutions),
b) NLP system development on IaaS (to give control over security from the OS level up),
c) "distribution" of NLP capacity through cloning of the entire virtualized NLP system, at the initiative of the adopting institution (to create a wholly owned, 100% locally controlled cloud-based instance for exclusive private use), and
d) integration of the cloud instance within local network security infrastructure, as some cloud

providers are now offering[12] (to add an additional, familiar layer of protection against external threats).

Assessing **readiness and comfort levels** and identifying local **risk mitigation preferences** (step 4) is an iterative process among key stakeholders, including HIPAA compliance officers, IRB members, IT leadership, legal counsel, researchers, and health system leadership. Such a process will reveal and must navigate local cultural and perhaps even political interests. Because of the high degree of potential institutional risk involved, we anticipate the process will be as constructive and educational as it is frustrating and time-consuming.

**Evaluating potential cloud service providers** (step 5) should be a relatively generic process, informed by the cloud model identified in step 3. A useful set of "nasty questions" to guide this process is provided by the Jericho Forum.[18]

Federal HIPAA and HITECH regulations specify that patient clinical text is **regulated data requiring special handling** (step 6) wherever it may be stored, accessed, or transmitted. Text containing PHI can only be accessed by approved personnel for purposes of delivering care, performing administrative functions such as quality control, or conducting IRB-approved research—and only to the extent *minimally necessary* for the purpose at hand. Outside this narrow scope access to patient health information is considered a disclosure and a breach of privacy. HIPAA defines two forms of "safe harbor" available to avoid disclosure risks: *encryption* and *de-identification*. These special handling provisions suggest the broad contours of a security infrastructure we believe will be necessary to develop a secure cloud-based NLP system:

a) data management policies and access controls permitting access only to clinical text intended for NLP processing, and limited to IT personnel and researchers developing and/or using the system,
b) 100% de-identification of clinical text, with manual verification, as a bridge measure providing complete security assurance during the early stages of security infrastructure development and vetting, and
c) encryption of clinical text in transit to and from the cloud or at rest in local databases; cloud-based storage is avoided entirely through 100% in-memory processing.

Safe harbor via de-identification entails removal of 17 HIPAA-specified identifiers, including names of patients, family members, addresses and other information that could be used to identify or re-identify (i.e., establish identity by inference based on available data) the patient.[13, 17] Manually verified de-identification is an onerous requirement that will limit the volume of text processed in the cloud. However, we believe this is a reasonable price to pay for a realistic but risk-free environment in which to perform proof-of-concept testing and to obtain early buy-in from relevant stakeholders. Even a complete security failure would not constitute a breach under HIPAA because none of the text involved would contain PHI.

When institutions are ready they may graduate from processing no-risk de-identified text to low-risk types of text. The "final diagnosis" section of a surgical pathology report is an example. This is where pathologists summarize their findings of the tissue study. The summary rarely contains PHI, and when it does it tends to be physician or institutional names as opposed to patient identifiers, which some IRBs and HIPAA compliance offices may consider to be a lower level of risk exposure. In any case, a security breach involving low-risk text would result in disclosures that were greatly reduced in quantity if not severity.

The final two steps in the security assessment process, **enumerating risk exposure points and audit requirements**, and **implementing security practices**, are analogous to those in traditional (non-cloud) IT settings and do not need elaboration here. We will only note that these activities represent ongoing responsibilities as technologies and threats evolve—in all IT settings, including the cloud.

The themes and issues explored in this section underscore the challenges in implementing cloud-based clinical NLP. They also suggest feasible solutions may be found, provided the adopting institution marshals the requisite technical expertise, pursues a carefully planned implementation strategy, has reasonable expectations about timelines, and includes the appropriate stakeholders. The next section focuses on those stakeholders and related socio-institutional prerequisites of adoption.

## 3.2. Socio-institutional aspects of security

Security is attained when and only when relevant local stakeholders declare it to be so. As we have seen, a number of different stakeholders play a role in identifying security requirements and deciding when they have been implemented. Because cloud computing introduces novel technologies with unfamiliar security challenges stakeholder engagement and education must, we feel, receive careful attention throughout the planning and adoption process. Similarly, risk exposure must be carefully managed through strategies that assure the highest possible exposures never exceed stake-

holders' lowest common tolerance level. Two basic principles underlie the non-technical side of establishing security: 1) persistent stakeholder engagement and 2) incrementalism.

Because all stakeholders (see section 3.1) play some material part or have a compelling interest in the cloud-based system they must be engaged early, often, and continuously in the planning and implementation phases. Engagement includes commitment to transparency. Transparency will speed identification of latent concerns, build trust, and help avoid "U turns" and other setbacks.

Education is another element of engagement. Providing stakeholders the information and knowledge they need to fairly and effectively fulfill their roles in the process is the responsibility of those who are leading the adoption effort. In an area as technical and unfamiliar as cloud computing addressing educational needs may require securing outside professional expert consultation. There is truth in the adage "If you think education is expensive, try ignorance."

Engagement also implies that all aspects of the project advance in an appropriate sequence and at a reasonable rate. Technical implementation must not outpace stakeholder readiness or tolerance for risk. Though advancing according to a "lowest common denominator" rule may seem overly cautious, the socio-institutional perspective argues against less inclusive and less deliberate approaches.

Incrementalism means starting small and advancing in carefully chosen, graduated steps. It increases the likelihood of acquiring needed knowledge and building a history of positive experience. It also reduces the chance of large-scale failure. The low-risk approach we propose using 100% de-identified text implements an incremental strategy that allows the security apparatus to be develop, tuned, and vetted before it is relied upon to protect sensitive patient data. This approach also gives local IT personnel an opportunity to develop experience and a deeper working knowledge of the respective systems, further enhancing stakeholder confidence.

The principles of persistent engagement and incremental implementation are based largely on common sense. The findings of our IRB survey, which we turn to next, suggest these principles can provide useful guidance in actual research settings.

### 3.3. IRB manager survey findings

As a preliminary exploration of the knowledge, attitudes, and beliefs of key stakeholders in applied health research settings we conducted a survey about potential uses of cloud computing. Briefly, the methodology employed was as follows. We developed an eight-item survey instrument and had it reviewed by our local IRB to confirm that it was exempt from human subjects review because it was limited to issues of infrastructure development (IRB reference number NR-10-013). In June 2010 15 IRB managers, one from each of the 15 HMO Research Network centers, were invited by email and one follow-up phone call to complete a brief, eight-item anonymous online survey implemented using SurveyMonkey.[19] The survey began with a 450-word preamble (provided in the Appendix) describing cloud computing and potential IRB-related issues raised when patient clinical text is involved. This was followed by eight closed-ended questions including one Yes/No question and seven ordinal-level rating scales. The questions assessed familiarity with cloud computing and elicited opinions regarding the expected difficulty of obtaining IRB approval for studies employing cloud-based technologies. After seven days the survey was closed and responses were tallied. Eight IRB managers completed the survey for a response rate of 53%.

**Table 1. Internal Review Board (IRB) Manager Responses to a Cloud Security Questionnaire Conducted in June 2010 with Percentage Distribution of Responses by Question[1]**

| Survey question | Response options | Pct |
|---|---|---|
| A. Before taking this survey how familiar were you with the concept of cloud computing? | 1. Not at all familiar | 63 |
| | 2. Slightly familiar | 25 |
| | 3. Somewhat familiar | 13 |
| | 4. Very familiar | 0 |
| | 5. I'm not sure | 0 |
| | Total: | 100% |
| B. Have researchers at your center expressed interest in cloud computing? | 1. No | 63 |
| | 2. Yes | 25 |
| | 3. I'm not sure | 13 |
| | Total: | 100% |
| C. If a researcher at your center proposed using cloud computing to process patient data how important would data security and patient privacy be in considering whether to grant IRB approval? | 1. Not at all important | 0 |
| | 2. Slightly important | 0 |
| | 3. Somewhat important | 0 |
| | 4. Very important | 88 |
| | 5. Hard to imagine approval under any circumstances | 13 |
| | 6. I'm not sure | 0 |
| | Total: | 100% |
| D. Defining institutional risk as "threat to an institution's reputation or ability to carry out its core mission," if a researcher at your center proposed using cloud computing to process patient data how important would institutional risk be in considering whether to grant IRB approval? | 1. Not at all important | 0 |
| | 2. Slightly important | 0 |
| | 3. Somewhat important | 0 |
| | 4. Very important | 88 |
| | 5. Hard to imagine approval under any circumstances | 13 |
| | 6. I'm not sure | 0 |
| | Total: | 100% |
| E. Compared to average projects with typical IRB issues, how much more difficult would it be to grant IRB approval to a project that is average/ typical except that it involves cloud computing to process clinical text containing HIPAA-protected health information (PHI)? [2] | 1. No more difficult | 0 |
| | 2. Slightly more difficult | 0 |
| | 3. Somewhat more difficult | 38 |
| | 4. Much more difficult | 50 |
| | 5. Hard to imagine approval under any circumstances | 13 |
| | 6. I'm not sure | 0 |
| | Total: | 100% |
| F. Compared to average projects with typical IRB issues, how much more difficult would it be to grant IRB approval to a project that is average/ typical except that it involves cloud computing to process clinical text containing no PHI, is 100% de-identified, and cannot be re-identified? [3] | 1. No more difficult | 25 |
| | 2. Slightly more difficult | 13 |
| | 3. Somewhat more difficult | 38 |
| | 4. Much more difficult | 13 |
| | 5. Hard to imagine approval under any circumstances | 0 |
| | 6. I'm not sure | 13 |
| | Total: | 100% |

Notes:
1. Anonymous survey administered via SurveyMonkey.com in June 2010; 8 of 15 IRB managers from research centers in the HMO Research Network responded yielding a response rate of 53%.
2. The seventh question (G, not shown), was similar to question E but asked respondents to assume respected researchers at two HMO-RN centers whose IRB managers were trusted had obtained approval for similar projects. Question G responses are discussed in section 3.3.
3. The eighth and final question (H, not shown), was similar to question F except as described in Note 2 above. Responses to question H are also discussed in section 3.3.

A response rate of 53% raises the possibility of non-response bias. To be conservative we thus avoid making inferences from this survey to HMO-RN research centers *as a whole*. Further, because the HMO-RN represents only a subset of all health-oriented research settings, inferences beyond the HMO-RN must be avoided. Nevertheless, IRB managers from eight different HMO-RN institutions completed the survey. This is a sufficiently large number of respondents to allow consideration of whether there is *any evidence of readiness* to experiment with cloud-based technologies within these institutions, which is the substantive question we address here.

Responses to survey questions are reported in Table 1. Overall, the results indicate substantial readiness among some IRB managers to consider cloud-based technologies. While 88% of managers were "not at all" or only "slightly" familiar with cloud computing concepts (Table 1, Question A), and only 25% indicated their researchers had expressed interest in the technology (Table 1, Question B), several IRB managers expressed the opinion that using these technologies is viable *under certain circumstances*. Clearly, securing patient data and mitigating risks to the health care institution are "very important" issues across the board (Table 1, Questions C and D). Even so, 38% of the managers felt obtaining IRB approval for cloud-based processing of de-identified patient clinical text was only "slightly more difficult" or "no more difficult" than approval for a typical project, and another 38% considered it to be only "somewhat more difficult" (Table 1, Question F). While most (63%) believed IRB approval for projects involving *identified* text would be "much more difficult" to obtain relative to "typical" projects—or even unimaginable—a surprising 38% saw this as only "somewhat more difficult" (Table 1, Question E). Interestingly, managers' assessments of IRB challenges were not influenced by whether other, trusted research centers had already granted IRB approval to such projects (presented as a hypothetical; see footnotes 2 and 3 in Table 1). This suggests IRB managers consider these issues to be predominantly local matters.

We interpret these results to be suggestive of meaningful readiness to pursue cloud-based NLP technologies in a cautious and graduated manner in at least some of the research centers surveyed. Additional education about cloud computing is clearly needed as evidenced by Question A results. Data de-identification is likely to be an important part of viable implementations, at least early on in an institution's experimentation with cloud models.

# 4. Conclusion

The cloud-based service systems model of low-cost NLP deployment we propose offers the potential to bring clinical NLP capacity to applied research settings. The core technologies appear to be feasible and economical, though additional proof-of-concept testing is needed. Such testing will extend existing knowledge of the adoption challenges inherent in the proposed technologies, and may suggest additional approaches for mitigating risks.

De-identification of patient text is likely to facilitate stakeholder readiness in this area, according to the IRB survey. We are unaware of any open-source de-identification software tools suitable for use in these applied settings; were such tools available they could alleviate implementation challenges.

Important questions remain about the acceptability of the institutional risks associated with cloud technologies, and these questions should be explored with relevant local stakeholders. It is clear that progress in addressing security concerns will require local experience and local stakeholder engagement. The IRB survey findings suggest that decisions about the acceptability of cloud computing is essentially a local matter.

The community of potential cloud-based NLP adopters would benefit from descriptions of the experiences of institutions that undertake all or part of the security assessment process described in section 3. Also, information about which if any steps in this process may benefit from the involvement of external expert consultants would be helpful. First-hand accounts of actual cloud-based implementation experiences would be extremely valuable. In particular, it would be useful to know whether, when, and under what circumstances institutions may be inclined to approve processing of imperfectly de-identified text or even un-redacted text that naturally contains very low levels of PHI. Our expectation is that such approval may develop over time as local knowledge and experience grow and as cloud security technologies and best practices mature, but this remains an unknown.

Prudence encourages those contemplating adopting cutting-edge technologies to ask: is it too early to adopt? Are these technologies more futuristic than feasible? Recent announcements by mainstream corporations suggest cloud processing of patient data is rapidly becoming part of our existing technology landscape. In August 2010 IBM announced its offering of a cloud-based electronic health record developed in collaboration with Aetna insurance.[20] Several other companies have similar

offerings. Outside the health area examples of cloud-based systems managing sensitive data are accumulating, particularly in the financial sector.[21] Given the growing interest within the Federal health technology arena and Federal health research institutes in exploiting cloud-based technologies, prudence would appear to be on the side of continuing to explore opportunities the cloud computing model offers.

## 5. Acknowledgements

This work was supported by grants from the Office of the National Coordinator for Health Information Technology (90TR0002), the National Human Genome Research Institute (1U01HG004610), and the National Cancer Institute (1RC1CA146917).

## 6. References

[1] SM Meystre, GK Savova, KC Kipper-Schuler, et al., Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Year Book of Med Informatics 2008;47(Supp 1):128-44.

[2] Office of the National Coordinator for Health Information Technology, Strategic Health IT Advanced Research Projects (SHARP) Web site, http://healthit.hhs.gov/portal/server.pt?open=512&objID=1436&parentname=CommunityPage&parentid=8&mode=2&in_hi_userid=11113&cached=true (last accessed Sept. 15, 2010).

[3] Federal Business Opportunities Solicitation Number N02PC05001-78, Tools for Electronic Data, Web site: https://www.fbo.gov/index?s=opportunity&mode=form&id=e4f9850461fc5f6eb5a41b19c4a3c3ea&tab=core&_cview=1 (last accessed Sept. 15, 2010).

[4] The Electronic Medical Records and Genomics (eMERGE) Network, Phase II (RFA no. RFA-HG-10-009), http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-10-009.html (last accessed Sept. 15, 2010).

[5] Guergana K Savova, James J Masanz, Philip V Ogren, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation, and applications. J Am Med Inform Assoc 2010 17:507-513.

[6] Richard Kissel, Editor, Draft Glossary of Key Information Security Terms, Revision 1 (NIST IR 7298,

Draft), US Department of Commerce, May 26, 2010.

[7] International Data Corporation (IDC) Enterprise Panel, August 2008.

[8] Open Health Natural Language Processing (OHNLP) Consortium Web site: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP (last accessed Sept. 15, 2010).

[9] The HMO Research Network Web site: http://www.hmoresearchnetwork.org/ (last accessed Sept. 15, 2010).

[10] The electronic Medical Records and Genomics (eMERGE) Consortium Web site: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page (last accessed Sept. 15, 2010).

[11] SHARP Project Area 4: Secondary Use of EHR Data: http://informatics.mayo.edu/sharp/index.php/Main_Page (last accessed Sept. 15, 2010).

[12] Amazon Virtual Private Cloud (VPC) Web site: http://aws.amazon.com/vpc/ (last accessed Sept. 15, 2010).

[13] Health Insurance Portability and Accountability Act of 1996 (HIPAA) http://www.hhs.gov/ocr/privacy/hipaa/understanding/special/research/index.html (last accessed Sept. 15, 2010).

[14] Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing V2.1, December 2009.

[15] Cloud Security Alliance Web site: http://www.cloudsecurityalliance.org/ (last accessed Sept. 15, 2010).

[16] Amazon Web Services: Overview of Security Processes, Web site: http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1697 (last accessed Sept. 15, 2010).

[17] Brad Rostolsky and Reed Smith, HHS Regulations Impose Federal Security Breach Notification Requirements, 10/6/2009. The National Law Review, http://www.natlawreview.com (last accessed Sept 15, 2010).

[18] The Open Group, Jericho Forum Self Assessment Scheme, March 2010.

[19] Survey Monkey online surveys. Web site: www.SurveyMonkey.com (last accessed Sept. 15, 2010).

[20] D. H. Kass, IBM offers cloud-based patient information service to health car providers, 8/16/2010, http://www.itchannelplanet.com/technology_news/article.php/3898751/IBM-Offers-Cloud-based-Patient-Information-Service-to-Health-Care-Providers.htm (last accessed Sept. 15, 2010).

[21] Elisabeth Horwitt, Cloud Security: Oxymoron? Computerworld, June 7, 2010.

## 7. Appendix

The following preamble was used to introduce the concept of cloud computing in the IRB manager survey conducted using SurveyMonkey. Following the preamble were the questions shown in Table 1.

**Description of Cloud Computing**

**What is cloud computing?** Traditionally, businesses have met computing needs by owning and operating computer systems *locally*—as all HMO-RN IT departments currently do. **Cloud computing** is a rapidly emerging alternative that provides computing capacity through the internet, analogous to a public utility providing electricity or natural gas. Also like a utility "the cloud" delivers computing capacity on-demand, cheaper, with higher quality and availability. The term "cloud" comes from the familiar clipart image of a cloud used to depict the internet in computer network diagrams. The major cloud service providers include Amazon, Google, IBM and Microsoft. A powerful feature of cloud-deployed systems is the ability "clone" them. With the click of a mouse a complex, painstakingly engineered data processing system can be replicated—any number of times—to meet growing needs or to give to another cloud user for their private use.

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction

**Why consider cloud computing?** Some software systems are difficult to deploy, such as those that "mine" information from clinical text using Natural Language Processing (NLP). Many HMORN investigators believe *local access to NLP systems* would create opportunities for compelling, fundable research, but the high cost of deploying NLP precludes this. Cloud computing offers a solution: NLP experts deploy a fully functional NLP system in the cloud (this has already been done), and then invite others to create their private clones of the system for their exclusive private use.

**What are the IRB issues?** The "catch" in this cloud-based solution is that it requires *temporarily* transmitting and *temporarily* processing patient data outside your institution's network firewall (before receiving back within your local firewall the NLP-mined data and text). Because the cloud computing service provider (e.g., Amazon or Microsoft) owns and manages the hardware infrastructure that makes the cloud services possible (typically in large data/server centers), security is unavoidably a shared responsibility (e.g., between Amazon as the cloud "provider" and your institution as the cloud "consumer"). This raises novel security issues and **novel IRB questions** about protecting the security and privacy of patient data. How can one be really certain the cloud provider is doing its part to secure the hardware? (There are answers to such questions but we do not need to go into them here.) The point is that security in the cloud contrasts with the traditional on-site, user-owned model of computing where all data and responsibility for security is 100% under the control of the local institution.