Responsible Machine Learning Pilot Test Projects: A Medical Coding Case Study

Samantha Champagnie, Janis L. Gogan

Responsible Machine Learning Pilot Test Projects: A Medical Coding Case Study

Samantha Champagnie[1] and Janis L. Gogan[2] (0000 0002 4770 2515)

[1] Muma College of Business, University of South Florida, Tampa FL USA
[2] Bentley University, Waltham MA USA

Abstract. Prior studies reported on many machine learning (ML) projects that under-performed. What steps can leaders take during ML pilot projects to identify and mitigate project risks and systems risks, before implementing new ML systems at scale? We report on an exploratory case study of a U.S.-based healthcare provider organization's ML pilot project, undertaken when a software vendor proposed an automated solution that would combine natural language processing (NLP) and ML, to improve medical claims coding quality. We reveal tactics the client took during the pilot project, to spot and limit risks that could ultimately harm the firm, its healthcare providers, and its patients. We conclude with suggestions for further research on responsible ML.

Keywords: AI, Machine Learning, Ethics, Governance, NLP

1 Introduction

"AI hype has far exceeded the state of AI science, especially when it pertains to validation and readiness for patient care" [32, p. 51]. In Winter 2021 IBM announced its planned sale of its Watson Health AI business [15] – an acknowledgement of a gap between the potential and actual realized value of AI in healthcare. Because some of the gap is attributed to human design mistakes that affect machine learning (ML) algorithms (e.g., when developers specify inaccurate or incomplete data sets for algorithms to analyze), ethicists propose that project sponsors and developers should be held accountable for ML mistakes that could harm patients and other stakeholders [21, p. 132; 34]. Consistent with this view, we define responsible machine learning (RML) as *the use of ethically-sound governance policies and controls to prevent ML errors and adverse events, to detect errors that nevertheless occur, and to minimize stakeholder harm, by correcting mistakes and appropriately adjusting relevant systems, processes, controls and policies.* Similar to [36], our definition acknowledges the duality of human fallibility and accountability, and recognizes that project leaders seek harm-free collaborative value creation [7, 35].

An IT pilot test -- a disciplined … time-bound, limited-scope, limited-participation project" [11] -- can flag some project risks or system risks before they cause harm.

In this paper we report on findings from an exploratory case study of an ML pilot test in healthcare (from before its launch to its end). First, we review relevant prior ML research in healthcare. After describing our research method, we present our case study findings, and discuss their implications for responsible ML pilot projects.

The Conclusions section discusses study contributions and limitations and offers suggestions for further RML research.

2 Prior ML Research: Insights and Perspectives

The "cognitive generation of decision support" [33] offers potential value in many industries. However, further research is needed to develop a better understanding of socially- and technically-constructed "synthetic knowing" challenges [23]. Several reviews summarize ML research in [29, 14, 32, 12], in healthcare and other contexts.

ML reportedly improves processes such as organizational sense making [1], judges' legal decisions [39], IoT data analytics for improved support of installed equipment [5], and cyber-security [9]. Analyzing huge structured and unstructured datasets [1, 25, 31], ML has been deployed to classify, compare, and detect patterns, and to optimize, predict and/or offer recommendations [29]. Healthcare datasets based on electronic medical records, claims, images, and/or social media content support efforts to improve operations and services [5, 32]. Some ML projects focus on disease diagnosis and treatment [4]. Other projects demonstrate ML potential for identifying triggers and risk factors, such as for asthma care [40] and detecting individuals at risk for suicide [8]. ML applications aim to support remote patient tele-monitoring [41] and to predict in-hospital mortality [36]. Thus, ML projects target many aspects of informed healthcare [14 16], including screening, triage, and treatment [32, 24].

ML design teams confront tradeoffs among algorithm explainability, simplicity, speed, and accuracy [36]. Many prior studies reported unintended ML consequences [12], including consequences linked to a common risk factor: the so-called ML "Black Box" (difficulty explaining and evaluating opaque algorithms) [32, 27]. To address this problem, developers are urged to ensure multidimensional data quality (e.g., validity, accuracy, completeness), design limited-scope algorithms in modules [2] and to take other steps to improve algorithmic explainability [22, 28].

CIOs who already oversee IS project portfolioss with varied risks may need expert help to evaluate unique ML risks (e.g., the Black Box and other technical, ethical and regulatory risks [5]). Committed partners [33] need to be both willing and able to collaborate effectively [26]. Ethicists and lawyers can help spot and mitigate some risks [21, 36]. In addition to partners with ethical and legal expertise, some new technical and managerial IS capabilities and roles are needed on ML teams [5, 22].

Prior ML studies suggest that ML projects need some new controls [33, 21]. IS project managers already seek to balance tight versus loose controls [35]. Tight

formal controls include strict deadlines and performance metrics, while looser informal controls include mechanisms for building strong relationships between developers and their customers [7]. Thus, some unique ML risks (such as the Black Box) reportedly need to be subject to relatively tight controls [2]. Looser informal controls needed to encourage a fact-based culture [33], promote realistic ML expectations [20] and ensure clear communication [13, 2]. Agile techniques [5, 39] -- including pilot testing of minimally-viable algorithms [28] – can help reveal risks or stakeholder concerns before problems arise.

Prior ML studies provide a helpful foundation, yet raise important questions. Operating a new system without mitigating a known risk would violate a tenet of responsible ML -- limit stakeholder harm. What specific new controls can mitigate unique ML risks? For example, what controls can help mitigate the Black Box problem? Under what circumstances is it necessary to redesign a planned ML system? Our case study aimed to address the following question: *In a collaboration between a healthcare organization and a software vendor, to pilot-test a minimally-viable ML system, how are ML project risks identified and mitigated?*

3 Research Method

3.1 Overview

This pilot ML project involved healthcare administration (claims coding and billing) and patient care (since training data relies on providers' medical documentation). "ProCo," (disguised) located in the U.S. East, handles claims processing for 500 physicians, physician assistants and nurse practitioners (hereafter, "providers"), in 45 medical specialties, for 3.5 million encounters per year. Its 200 staff work in ProCo's central office; 700 non-clinical support staff are located in provider clinics. About half of its patients are Medicare beneficiaries (age 65 and up). ProCo's ML pilot was initiated in early winter 2018, when executives became aware of a vendor's potential solution (at that time, they did not call it a pilot test; some executives hoped they were purchasing a low-cost coding solution).

A case study is a suitable research method for exploring complex new phenomena holistically and with a focus on "how" and "why" questions [37]. Our participant-observation case study began in March 2018. One author, a ProCo employee, had ongoing access to managers and documents, and sat in on project-related meetings. In interviews conducted April 2018 to December 2018 (10 ProCo employees, 3 ProCo providers, 3 SofCo employees), interviewees described contractual issues, stakeholder expectations, and technical and operational challenges in this "coding automation" pilot project. In a final June 2019 interview and follow-on emails, ProCo's Vice President of Revenue Cycle Management ("VP") described developments that took place in spring 2019. Our study archive contains hand-written field notes, vendor status reports, ProCo documents (e.g., weekly coding quality reports, relevant emails), and project meeting notes that the VP prepared for the executive team. This paper emphasizes the VP's perspective. A newcomer to the organization, she challenged many taken-for-granted assumptions – which was

valuable for revealing risks other ProCo managers either did not see or were not willing to disclose.

3.2 Background: A Medical Claims Coding Tutorial

Healthcare providers produce encounter documentation (free-form notes plus highly structured codes) to describe patient evaluation, condition, and treatment. Current Procedure (CPT) codes describe treatment procedures, equipment, and medications prescribed. A CPT subset -- Evaluation & Management (E&M) codes -- describe the patient's status (new/existing), care setting (inpatient/ outpatient, medical unit, etc.), provider's review of their medical history, details relevant to patient's presenting condition, and provider's examination of the patient. International Classification of Disease (ICD) codes describe a patient's medical condition. CPT and ICD codes should align, since a provider performs E&M tasks based on a patient's presenting condition and diagnosis, their in-hospital medical record, and their personal health record (containing details of past patient encounters with a primary care provider and specialists). Specific E&M codes link to specific reimbursement amounts.

CPT codes are input into medical billing software via manual data entry or automated data transfers from other systems. Claims (submitted to private insurance companies or government agencies like Medicare) are denied, adjusted or delayed, and hefty financial and licensure penalties may be imposed, if they contain incorrect codes. Providers complain about overly complex coding rules [17, 19, 38]. Studies report that while computerized provider order-entry (CPOE) systems reduce many errors, new errors arise due to usability issues [6, 3, 10].

4 Case Findings

4.1 SofCo Proposes a Medical Claims Coding Solution

ProCo's certified medical coders (paid $23/hour) in the central business office (CBO) input complex codes describing in-hospital care, while ProCo providers were responsible for office-visit notes and coding. Most providers produced these in real time, using speech-to-text software. Like many provider organizations, ProCo struggled to achieve consistently high E&M coding accuracy. In eight of 25 medical specialties, coding compliance overall (per internal and external audits) was less than 80% (20% or more claims contained at least one incorrect code). In some specialties, a few individual providers produced many inaccurate codes. These chronically non-compliant providers increased the risk that they and/or ProCo would incur penalties.

"SofCo" (disguised) proposed to create software that would audit all ProCo providers' office visit E&M codes. Its planned "engine" would rely on natural language processing (NLP) and ML. NLP software would scan real ProCo office encounter records and "interpret" their meaning (via pattern-recognition). First, SofCo would need to train the software. "Expert" certified offshore coders based in India would save their own E&M code decisions into the NLP/ML training dataset.

With each pass through this training data, the ML algorithm would scan for patterns to enact as coding rules. Although the ProCo CIO and the central business office (CBO) head expressed some skepticism at that time, the CEO and CFO, envisioning a potentially very large financial benefit, signed a "limited-scope" contract with SofCo (scope limited to office-visit CPT E&M code auditing).

Between signing the contract in early winter 2018 and the start of our case study in late March 2018, the CBO head resigned. A new VP for Revenue Cycle Management ("VP") spent her first month on the job learning about ProCo and the planned "coding automation project" (not then referred to as a "pilot"). In an April demonstration, SofCo's Sales Director boasted their software would be the first NLP/ML system to perform E&M code auditing. The Sales Director claimed their existing NLP platform was already capable of suggesting E&M codes; it just needed "tuning;" within six months (she claimed) SofCo's software would successfully auto-code at least 60% of ProCo's encounter charts. She claimed that by May 2019 (one year from a planned May 2018 kick-off) SofCo's auto-coded charts would be comparable to provider's codes. At that point, audited codes (those containing no discrepancies, per the experts) would be routed directly to ProCo's billing system (without human intervention). Table 1 summarizes project milestones.

4.2 The ProCo VP Attempts to Establish Control

The new ProCo VP some came to recognize that this project was risky. No project plan or other "standard" vendor documentation existed (specifying deliverables, roles and responsibilities, data and process flows, etc.). Colleagues informed her about problematic communications issues. For example: although the project was contractually limited to E&M coding for office visits, SofCo's project manager inexplicably spent time learning about "hospital" encounters and associated coding. The Sales Director stated an E&M coding platform already existed, yet the VP soon learned that SofCo's NLP/ML "engine" was not yet operational. She learned that IT consultancy Gartner described NLP+ML solutions for medical coding as in a nascent stage of development. While the VP was "reassured" to learn that SofCo had NLP experience (ProCo providers used SofCo's speech-to-text software to produce encounter notes), she was "concerned" to discover that SofCo did not have a record of strong machine learning. The project goal was clear: produce a system capable of automatically producing E&M codes based on providers' encounter documentation, to a 95% level of accuracy. However, the means to achieve it were not clear. The VP also learned that ProCo did not issue a formal Request for Proposal (which, e.g., should spell out how ProCo would select and securely send charts to SofCo). The contract did not include necessary details (e.g., step-by-step explanations of how the proposed solution or coding accuracy verification would work, or whether SofCo was required to return or destroy ProCo's data at the project conclusion).

Hoping to clarify the project scope and roles, the VP arranged a second interactive online demonstration of SofCo's NLP+ML prototype in May. This session

confirmed that SofCo's coding platform was not fully developed; further work was necessary to comply with E&M coding guidelines, and it could not become operational until human coders fed the training database. During this session, the VP clarified that the contract limited the scope to office visit encounter coding. Now, she narrowed the scope further, by limiting the project in two ways: 1) utilize only office encounter charts from ProCo's family practice specialty, and 2) include only charts of those providers whose coding quality was less than 80% (per government audits and ProCo internal audits).

Table 1. Timeline of Key Events in the Coding Automation Pilot Project

Date	**Project Event**
2018	
winter	Limited-scope contract signed.
Mar	Case study begins. New VP-RCM ("VP") meets with outgoing VP-RCM, CFO, CIO, and IT Director to learn about planned NLP+ML collaboration with SofCo
April	Early April: SofCo conducts product demonstration VP meets with ProCo compliance officer, coding manager, a coding auditor, and some providers to learn if/how current processes would need to change during the pilot project and after system rollout. CBO updates CEO and CFO on coding compliance project concerns.
May	SofCo Project Manager agrees to weekly Project Status meetings SofCo conducts a second product demonstration. VP asks SofCo's Project Manager to provide a project plan.
June	VP seeks further clarification of project details on ProCo side
Aug	Detailed project discussions with CIO and IT Director. Outcome: How ProCo ambulatory encounter charts would be sent to SofCo.
Sept	First Auto Feed of ambulatory encounter charts sent to SofCo. Daily feeds (M-F) thereafter until May 2019.
Oct	Review of feedback loop to Providers (how Indian coders would notify ProCo providers of suggested code changes based on their reviews, and expectations re timely ProCo provider responses)
Nov	ProCo 10% audits of SofCo coders' accuracy begin (continues until May 2019)
Dec	News that SofCo would soon be acquired by a very large company leads ProCo VP to have a "scope clarification" conversation with SofCo Sales Director.
2019	
March	SofCo starts providing weekly written status reports
April	SoftCo announces invoicing ($.50/claim) will start May 1. SoftCo provides "cryptic" weekly code quality status reports (per VP)
May	Project escalation to CIO, CEO and CFO; contract terms renegotiated
June 26	The Coding Automation pilot project is dissolved. ProCo VP provides a final update describing spring 2019 developments. Case study ends.

The May 2018 demo resolved some important concerns, but other concerns arose that summer. The contract specified ProCo would send its office visit encounter records to SofCo on a daily basis, starting in May 2018. However, it took most of the summer to work out exactly how to transfer data securely to SofCo. The first

data transfer took place in September, and thereafter, ProCo sent Family Practice office encounter records to SofCo each weekday.

The ProCo VP aired her concerns in weekly meetings with SofCo personnel. In fall 2018 she asked how SofCo measured the offshore coders' coding accuracy. SofCo replied that ProCo was welcome to audit their work. To that end, the VP added a U.S.-based certified quality auditor to the project team, tasked with spotting and correcting offshore coders' errors. These audits began in November. These, and weekly meetings revealed that SofCo coders' accuracy was not as strong as SofCo's sales pitch predicted. This greatly concerned the ProCo VP; she felt claims auto-coding should not move forward until SofCo "experts" achieved 95% accuracy. She reasoned that "garbage-in/garbage out" applies to ML: if offshore coding accuracy was weak, the data set would train the ML algorithm to "learn" incorrect coding rules. She expressed surprise that SofCo did not evaluate its coders' accuracy.

In mid-December 2018, SofCo announced it had agreed to be acquired by a Fortune 100 coding technology company; the deal was to be finalized in Q1 2019. SofCo's Sales Director assured the surprised ProCo VP "Nothing will change;" the name on their project materials, email signature and letterhead would reflect the acquiring company, yet the acquisition would not impact the project. The Sales Director expressed enthusiasm about their future parent company's considerable technical resources, which would further their development and design efforts. In turn, the ProCo VP reported to ProCo's executive team that the acquisition would bring additional resources to the coding automation project and should have no adverse impact on the project timeline.

In winter and spring 2019, ProCo's VP saw little improvement in offshore coders' quality, and she learned little about the opaque ML algorithm. In March, SofCo finally began providing weekly status updates. These mostly reported on corrective actions taken to improve offshore coders' quality. The VP stated that SofCo "minimally addressed the ML engine development; they merely indicated it was 'on track'." In her view, SofCo's report format was uniquely "cryptic … [and] at such a high level that I had to request multiple follow-up meetings just to understand it."

In April, SofCo stated it would invoice ProCo for coding services, starting May 1 (one year after the "effective" project start date of May 1, 2018, per the contract). ProCo responded by proposing a new agreement; ProCo would keep sending SofCo the data feeds they needed to train their algorithm, but SofCo should issue no invoices until its human coders successfully achieved a 3-month cumulative accuracy score of 95%. In her weekly updates to the CIO, CFO and CEO, the VP now reported the project status as "at risk." SofCo had yet to demonstrate an ability to deliver an automated solution that could produce compliant E&M coding.

In June 2019 SofCo informed ProCo that their new parent company would transfer SofCo's ML project to the parent's ongoing NLP+ML development effort, in order

to consolidate resources. SofCo assured ProCo they would reengage once the parent's auto-coding software was "ready for market." On June 26, ProCo's executive team decided to end the coding automation project.

5 Discussion

As discussed above, prior studies advise ML project leaders to choose willing and capable partners and set realistic expectations. In retrospect, the ProCo VP believes SofCo made unwarranted promises (predicting their algorithm would be ready to recommend codes within six months of project initiation, and would correctly auto-code 60% of ProCo charts within one year). The ProCo VP, CIO, ProCo IT staff, and ProCO executives lacked ML experience. The new VP sensed a "disconnect" between the optimism of the CEO and CFO (who focused on potential financial benefits) and the CIO, who seemed cynical about, and disengaged from, this project.

The new VP played a valuable role, both by challenging taken-for-granted assumptions and by drawing on her prior expertise as a project manager in a coding compliance context. After the project ended, she reflected: "In previous software implementation projects, a lack of expertise on our end was not necessarily a problem; we relied on vendors' assurances that their products were ready for use." The VP did know how to evaluate SofCo's medical coding expertise, and she came to recognize why this was important (for training the algorithm). Her ability to evaluate SofCo's ML claims improved during the pilot project (thanks to Gartner reports and other authoritative sources that helped educate her about NLP and ML).

Because of the Black Box (algorithm explainability) problem, many prior studies advised ML project leaders to utilize a modular design. The pilot collaboration ended before SofCo was ready to release their ML software for ongoing operations. Up through that point, SofCo's status updates were seen (by the ProCo VP) as vague. After the project ended, she expressed the opinion that both SofCo's weak project management expertise and the black box challenge affected the project from the outset; she suspected that weak project management was the root cause. Given that three product demonstrations were necessary (because of questions the VP and others had about how the NLP+ML engine would learn), we believe a Black Box issue was evident. We do not know if SofCo attempted to design for modularity or explainability, but ProCo's VP stated that in meetings, SofCo personnel were unable to convey how their software worked, and their written status reports were "cryptic."

The case findings about SofCo's medical coding accuracy difficulties point to a vitally important ML issue. If humans produce data that will be used to train an ML algorithm, a) the data (in this case, medical E&M codes) must be correct, and b) the human process of producing that data (in this case, choosing codes based on providers' medical documentation) should be explainable. From this we infer that the "black box" of the human brain can be an antecedent to the ML algorithm "black box". The ProCo VP saw evidence that U.S.-based human ProCo coders were more proficient at E&M coding than the certified coders SofCo hired in India. Had this project been designed to rely on ProCo's coders to train the ML algorithm, she said,

ProCo would have negotiated a very different contract with SofCo (since U.S.-based coders earn much higher wages than India-based coders).

Start small and use appropriate data: The contract indicated the project would focus on E&M coding for office visits. The ProCo VP limited the scope further (just the Family Practice specialty and only those providers with weak prior coding quality). This latter choice added complexity to the project and contradicts prior advice to tackle easier problems first and gradually introduce complex patterns into ML training data sets [18]. SofCo's choice to request daily data feeds also added unnecessary complexity (as did their use of offshore coders to train the algorithm). SofCo could have asked ProCo for historical claims data (considered best practice for those ML projects involving processes with verifiably "correct" solutions).

Prior studies emphasize the importance of identifying clear success criteria and metrics, and designing controls that can detect mistakes. Both partners agreed that a successful system would pick correct E&M codes based on providers' documentation. Use of offshore coders as arbiters of correctness was problematic, but the VP overcame that problem by hiring a U.S.-based certified medical claims coder to audit their work (how the VP came to realize that the offshore coders were less skilled than SofCo claimed). The VP attempted to impose relatively tight formal control by requesting written project status updates based on project milestones and coding quality metrics. For months, SofCo did not send the requested reports. Perhaps this was because SofCo was wrestling with algorithm explainability issues? Perhaps they chafed at ProCo's attempted tight control (did not feel like a "partner")?

As discussed above, prior studies advise ML teams to partner with legal and ethical experts, and to especially rigorously evaluate clinical ML systems [32, 36]. Both ProCo and SofCo apparently framed this pilot project as having an administrative focus. Yet, medical coding is not merely an administrative job. Treatment efficacy studies, clinical trials, and public-health studies rely on accurately-coded medical records; poor data quality in this context can ultimately jeopardize care quality. The ProCo VP questioned why SofCo "was willing to use coders whose accuracy was only in the 60% range," to feed the ML training dataset. "I was surprised they did not hire auditors to verify the offshore coders' accuracy; everyone just assumed their codes were accurate," she said. In the context of "regulatory scrutiny … isn't a failure to verify accuracy unethical?" The VP was also concerned about possible federal penalties: "What would we say to the government? That the … algorithm coded it, so we assumed it was right?" Her comments emphasize that responsible data governance is a necessary element of responsible machine learning.

Prior ML studies emphasize the importance of clear communication among collaborators. The ProCo VP stated that weak communication was a problem, from start (e.g., scope confusion led SofCo to waste time mapping hospital processes) to finish ("nothing will change" statement by SofCo's Sales rep, just one month before the project's dissolution). After the pilot project ended, the VP stated she now believes ML projects "require more than traditional governance." One prior study suggested ML projects should be located in business units, not in IT [25]. The ProCo

VP believes an internal partnership is needed: "While it is logical to embed a project like this in a business unit, the IS team needs … to play an important role in the overall project management and governance. The business unit understands what the ML engine needs to do, but the IS team should understand how to manage the IS project risks."

6 Contributions, Limitations and Conclusion

A prior study reports that provider resistance doomed an NLP+ML medical coding project in a German hospital [28]. Our case study, of a similar NLP+ML medical coding pilot project, revealed other impediments. Provider resistance did not impede this pilot; instead, the Black Box problem seems to have exacerbated communication, planning, and shared governance. Prior ML studies advise leaders to establish an appropriate ML project governance structure, including agreed-upon formal and informal preventive, detective and corrective project controls [22]. Our case study followed an ML pilot from launch to dissolution, to track specific risks the ProCo VP identified and attempted to mitigate. We note that each organization entered this collaboration with some unresolved internal governance challenges, and that the collaboration suffered from several shared-governance issues. The ProCo VP recognized a need for stronger governance, and took several appropriate steps to impose control (requesting weekly meetings and written status updates, adding a U.S.-based medical claims coding auditor to the team, etc.). While SofCo did not disclose specific technical issues in their ML algorithm, the training data quality was implicated (human coders struggled to produce accurate codes and could not explain some coding decisions). ProCo's VP, with 20 years' relevant prior experience, recognized that human mistakes would contaminate the training data set that fed this ML algorithm. She attempted to exercise both formal and informal control, by requesting written status reports (formal control) and insisting on weekly meetings (informal control). Physical distance and lack of direct access to the offshore coders impeded some of her attempts at control.

A study limitation is that we cannot verify why SofCo's new parent company put the collaboration on hold. Had SofCo proposed to train the algorithm with prior approved claims from ProCo's high-quality providers (removing incorrect claims, such as those denied by insurers or flagged in internal and government audits), we believe this pilot might have succeeded. A fruitful next case study would focus on an organization that uses verified prior claims data for their training data set. That study would seek to answer a similar research question: What project risks and system risks arise? How (if at all) does a project sponsor or project manager mitigate known risks, prior to authorizing an ML system for operational use?

Schuetz & Venkatesh [30] propose that some prior IS practices and assumptions do not fit ML projects. Other studies link the ML Black Box problem (one unique ML challenge) to adoption issues [28]. Our case study reveals suggestive evidence that a human Black Box/explainability problem affected an ML pilot project. No one on

the ProCo side understood how the ML algorithm would choose codes, and SofCo personnel could not explain "in understandable terms" (ProCo VP's phrase) how their software or human coders did or would spot patterns or how specific patterns did or would guide its coding decisions. The VP did not want to "blindly trust" the machine, the vendor liaison, or SofCo's claims coders. Both weak human explainability and weak machine explainability limited this manager's control options. How to impose preventive process controls in the face of opaque algorithmic or human decisions? She focused on what she knew about SofCo coders' performance. Behind the curtain, SofCo apparently struggled to "tune" its algorithm, but the ProCo VP was unable to deploy detective controls pointing directly to specific SofCo ML algorithm problems. The VP did recognize that an algorithm cannot be considered reliable if its training data is not verifiably reliable. Further design studies could attempt to develop automated detective controls that reveal why specific ML problems occur. Until then, smart systems need capable human partners. New case studies are needed, to continue to explore how humans and machines collaborate effectively or ineffectively in ML projects.

Unrealistic expectations constrained managers' and clinicians' readiness to participate in this case study, similar to findings of prior ML studies [27]. A CIO can temper unrealistic expectations by establishing project governance that fully addresses project planning, controls and oversight. This is especially important for those healthcare ML projects at the intersection of administrative and clinical practice. Such projects bring financial and regulatory risks, along with threats to patient privacy, quality of patient care, and public health. Stakeholders include patients, regulators, healthcare systems, payers, and clinicians [14].

A study limitation is that this paper focused on one key informant – a well-qualified newcomer VP who took responsibility for oversight of this pilot. A fuller exposition would closely examine the perspectives of other stakeholders (starting with the other participants whom we interviewed). New studies are also needed that look closely at specific ML risks that threaten harm in terms of diversity, equity and inclusion (with important social and ethical implications; see [26].

There is much to learn about challenges revealed in responsible (or irresponsible) ML pilot projects, and implications for subsequent large-scale ML implementation projects. We encourage other researchers to join this effort, with new design science, action research, critical incident studies and case studies that can shed further holistic light on early-stage collaboration in client-vendor ML pilot projects.

References

1. Abassi A., Zhou Y., Deng S., Zhang P. 2014. Text analytics to support sense making in social media: A language-action perspective. MIS Quarterly (42:2): 427-464.
2. Asatiani A., Malo P, Nagbøl P.R., et al.. 2020. Challenges of explaining the behavior of Black-Box AI systems, MIS Quarterly Executive (19:4.

3. Ash J., Sittig D., Poone E., et al. 2007. The extent and importance of unintended consequences related to computerized provider order entry. Journal of the American Medical Informatics Association (14:4), 415-423.
4. Bardhan I., Chen H., Karahanna E. 2020. Connecting systems, data and people: A multidisciplinary research roadmap for chronic disease management. Introduction to the Special Issue on IT and Chronic Disease, MIS Quarterly (44:1), 185-201.
5. Bilgeri D, Gebauer H., Fleisch E., Wortmann F. 2019. Driving process innovation with IoT field data. MIS Quarterly-Executive (18:3), Article 5, August.
6. Campbell E., Sittig D., Ash J., Guappone K., Dykstra R. 2006. Types of unintended consequences related to computerized provider order entry. Journal of the American Medical Informatics Association (13:5), 547-556.
7. Cardinal L.B., Kreutzer M., Miller C.C. 2017. An aspirationl view of organizational control research: Re-invigorating empirical work to better meet the challenges of 21st century organizations. Academy of management Annals (11:2), 559-592.
8. Chau M., Li T.M.H., Wong P.W.C., et al. 2020. Finding people in emotional distress in online social media: A design combining machine learning and rule-based classification. MIS Quarterly (44:2), 933-955.
9. Ebrahim M., Nunamaker J.F., Chen H. 2020. Semi-supervised cyber threat identification in dark net markets: A transductive and deep learning approach. JMIS (37:3), 694-722.
10. El Shayib M. & Pawola L. 2020. Computerized provider order-entry related medication errors among hospitalized patients. Health Informatics Journal (26:4), 2834-2859.
11. Gogan J.L. & Rao A. 2011. When vendors participate in IT pilot test projects: Pitfalls and challenges. Engineering Management Journal (23:3), 2-29, Sept.
12. Gogan J.L. 2021. Responsible machine learning projects. Proceedings of the 27th Americas Conference on Information Systems (AMCIS).
13. Goul M. 2018. Poised Between 'a Wild West of Predictive Analytics' and 'an Analytics of Things Westworld Frontier (17:4), MIS Quarterly –Executive Article 9, Dec.
14. He J., Baxter S.I., Xu J., Zhou X., Zhang K. 2019. The practical implications of artificial intelligence technologies in medicine. Nature Medicine (25:1), 30-36.
15. Hernandez D. & Fitch A. 2021. IBM's retreat highlights hurdles for health AI. The Wall Street Journal 20 Feb.
16. Johnston M. 2018. The transformation of healthcare with AI and machine learning. InformationWeek October 16.
17. King M.S., Sharp L., Lipsky M.S. 2001. Accuracy of CPT evaluation and management coding by family physicians. Journal of the American Board of Family Practice (14:3), 184.
18. Kühl N., Hirt R., Baier L., Schmitz B., Satzger G. 2021. How to conduct rigorous supervised machine learning in information systems research: The Supervised Machine Learning Reportcard, Communications of the Association for Information Systems (CAIS).
19. Kumetz E. & Goodson J. 2013. The undervaluation of evaluation and management professional services. Chest (144:3), pp. 740-745.
20. Lacity M.C., Scheepers R., Willcocks L.P. 2018. Cognitive Automation as Part of Deakin University's Digital Strategy. MIS Quarterly-Executive (17:2), May.
21. Martin K. 2019 Designing Ethical Algorithms. MIS Quarterly-Executive (18:2) Article 5.
22. Mayer A-S., Strich F., Fiedler M. 2020. Unintended consequences of introducing AI systems for decision making, MIS Quarterly Executive (19:4).

23. Monteiro E. & Parmiggiani E. 2019 Synthetic knowing: The politics of the internet of things. MIS Quarterly (43:1), 141-65.
24. Mousavi R., Raghu T.S., Frey K. 2020 Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums. JMIS (37:4): 1073-1098.
25. Muller O., Junglas I., Debortoli S., vom Brocke J. 2016. Using text analytics to derive customer service management benefits from unstructured data. MIS Quarterly-Executive (15:4), Dec.
26. Newell, S. & Marabelli M. 2015. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. Journal of Strategic Information Systems (24:1): 3-14.
27. Pumplun L, Fecho M., Islam N., Buxman P. 2021. Machine learning systems: How mature is the adoption process in medical diagnostics? Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS), 6317-6326.
28. Reis L., Maier C., Mattke J., Creutzenberg M., Weitzel. 2020. Addressing user resistance would have prevented a healthcare AI project failure, MIS Quarterly Executive (19:4).
29. Rzepka C., Berger B. 2018. User interaction with AI-enabled systems: A systematic review of IS research. Proceedings of the Thirty Ninth International Conference on Information Systems (ICIS), San Francisco, Dec.
30. Schuetz S. & Venkatesh V. 2020. The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. Journal of the Association for Information Systems (21:2), 460-482.
31. Shin D., He S., Lee G.M., Whinston A.B., Centintas S., Lee K0C. 2020. Enhancing social media analysis with visual data analytics: A deep learning approach. MIS Quarterly (44:4), 1459-1492.
32. Topol E. 2019. High performance medicine: The convergence of human and artificial intelligence, Nature Medicine (25), 44-56.
33. Watson H. 2017. Preparing for the cognitive generation of decision support. MIS Quarterly-Executive (16:3), August.
34. Wessel M. & Helmer N. 2020. A crisis of ethics in technology innovation. MIT Sloan Management Review, SMR797, 70-76, Spring.
35. Wiener M., Mahring M., Remus R., Saunders C., Cram W.A. 2019. Moving IS project control research into the digital era: The 'why' of control and the concept of control purposes. Information Systems Research 31 October.
36. Wiens J., Suchi S, Sendak M, et al. 2019. Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine Perspective (25), 1337-1340, Aug.
37. Yin R. 20143 *Case Study Research: Design and Methods* (5th Edition). Sage.
38. Young R.A., Bayles B., Hill J.H., Kumar K.A., Burge S. 2014. Family physicians' opinions on primary care documentation, coding and billing. Family Medicine (46:5), 278-384.
39. Zhang Z., Nandhakumar J.,Hummel J.T., Waardenburg L. 2020. Addressing the key challenges of developing machine learning AI Systems for knowledge-intensive work. MIS Quarterly Executive (19:4), Article 5.
40. Zhang W. & Ram S. 2020. A comprehensive analysis of triggers and risk factors for asthma based on machine learning and large heterogeneous data sources. MIS Quarterly (44:1), 304-349.
41. Zhu H., Santani S., Chen H., Nunamaker J.F. 2020. Human identification for activities of daily living: A deep transfer learning approach. JMIS (37:2): 457-483.