One of the most common indications for lumbar spinal surgery in the US is lumbar stenosis with degenerative spondylolisthesis. There is enormous practice variation in the surgical treatment of lumbar stenosis with degenerative spondylolisthesis. One of the areas of uncertainty, and therefore controversy, is whether or not to perform a lumbar spinal fusion when surgically decompressing the spinal column to relieve nerve compression symptoms in these patients. One possible interpretation of the recent randomized controlled trials (RCTs) published on this topic is that many patients do not need a fusion given that only 30% of those treated with decompression alone go on to develop delayed instability manifesting as mechanical low-back pain that is often treated with fusion.

The key question is how to initially identify the estimated 30% of the patient population that will develop instability following decompressive laminectomy alone. From registry data in the US, it appears that we performed a fusion on nearly 80% of these patients. Thus, there is an opportunity to tailor the surgical approach by understanding what patient-specific factors may lead to the development of instability so that we can target these patients up front for fusion.

There are two things that can help us approach the opt-
mal answer for individual patients. First, there are clinical registries that house important data, and the most robust contain patient-reported outcomes (PROs) and imaging data. These data are particularly useful when applying advanced computing techniques, specifically artificial intelligence (AI) and the related area of machine learning (ML) to model and predict from real data. Ultimately, these factors can impact decisions by surgeons to develop approaches or make therapeutic recommendations that can be put into practice in real scenarios that may improve surgical practice.

Electronic Medical Data, Administrative Claims Databases, Registries, and Data Warehouses

Multiple administrative claims databases have been used for data analysis. The National Inpatient Sample (NIS) contains data representative of 97% of all patient discharges in the US (https://www.hcup-us.ahrq.gov). The NIS database, which is part of the Healthcare Cost and Utilization Project (HCUP) in partnership with the Agency for Healthcare Research and Quality (AHRQ), is the largest publicly available healthcare database. It can be used to examine trends in healthcare, which is important in understanding variance in treatment for diagnoses such as degenerative spondylolisthesis.

The NIS data can be used to look at not only trends in utilization, but also complication rates and total hospital charges. Norton et al. were able to extract outcome measures such as these as well as demographic data for all patients who had undergone surgery for degenerative spondylolisthesis from 2001 to 2010. Among the 48,911 surgical patients identified using ICD-9 codes, the authors found a complication rate of 22.5%. This complication rate differed significantly by approach and was limited to patients who had undergone a fusion.

State Inpatient Databases (SID) contain claims data that can track individual patients over time. As in the NIS, data in the SID are representative of the population of each respective state, with 49 of the 50 states participating. These databases can be used to study reoperation and readmission rates. A recently published study used SID to quantify the reoperation rates following surgery for lumbar spondylolisthesis. In that study, the rate of reoperation was about 17% at 5 years regardless of whether or not a fusion was utilized when decompression was performed.

While administrative databases contain vast amounts of clinical data, there are limitations to these data sets. As Gollogorsky et al. identified in one practice, the surgeon’s indication for surgery was not always aligned with the diagnosis codes found in the administrative data set. These authors warn about limitations in the accuracy of data extracted by procedure and diagnosis codes that ultimately serve as the source for bigger data sets such as the NIS. Thus, larger data sets should be reviewed with a critical eye when informing policy and the like. Moreover, administrative databases do not contain PROs, so they are not optimal for comparing the effectiveness of surgical procedures.

In response to the increasing need to develop high-quality comparative effectiveness research, clinical registries have also been developed. Recently, the American Association of Neurological Surgeons has developed a quality outcomes database that has examined the outcome of sur-

Lumbar Decompression Alone Versus Lumbar Decompression and Instrumented Fusion

One of the earliest studies on decompression alone or decompression with fusion is a prospective comparative trial published by Herkowitz and Kurz. In their study of 50 patients, 96% of those treated with decompression and posterolateral fusion had satisfactory results at the 3-year follow-up as compared with only 44% of those treated with decompression alone. Six years after the publication of that study, Fischgrund et al. completed and reported the results of an RCT that compared decompression and fusion with or without instrumentation. These investigators found that instrumentation of the spinal column was associated with a higher fusion rate, but that patient outcomes were similar regardless of whether fusion was ultimately achieved. This finding left open the question of whether or not fusion was actually necessary. Then in 2004, Kornblum et al. examined the long-term outcomes in the patients enrolled in both of those trials and concluded that the patients who actually achieved solid fusion had superior clinical outcomes. Nonvalidated assessment tools were used in these studies, making it difficult to draw solid conclusions about comparative patient-reported health outcomes. All of these data and the lack of robust, validated outcome tools underscored the need for modern RCTs.

In April 2016, two major RCTs were published on the topic of whether or not to fuse the spinal column after decompressing the spinal canal at the level of degenerative spondylolisthesis. The two trials studied different patient populations and came to conclusions that appear different on the surface. The Swedish trial examined a heterogeneous patient population with lumbar stenosis with or without spondylolisthesis with either 1 or 2 levels of disease. On the other hand, the Spinal Laminectomy versus Instrumented Pedicle Screw (SLIP) RCT was focused on a homogeneous patient population and concluded that patients with single-level lumbar stenosis with grade I spondylolisthesis had superior health-related quality of life outcomes at 2, 3, and 4 years when fusion was added to decompression. Moreover, the SLIP study demonstrated over a 4-year follow-up period that fusion plus decompression was associated with a 14% rate of reoperation, which was significantly lower than the 34% rate observed in patients treated with decompression alone.

One way to make sense of the two trials would be to conclude that two-thirds of the patients with grade I spondylolisthesis were stable enough to tolerate decompression without fusion, but that one-third of the patients appeared to develop instability over time. This instability often leads to a reoperation to fuse the spine at the level of listhesis. Future studies should focus on determining which patients are more likely to destabilize after decompression alone in order to select patients more appropriately for upfront fusion when a decompressive laminectomy is performed.
surgery for spondylolisthesis. They have been able to study outcomes in real-world US practice. In a recent study of 441 patients, investigators found that 82% of patients were treated with fusion and decompression and that 1-year outcomes were similar to those reported in the SLIP RCT in the New England Journal of Medicine in 2016.1

Today, most health centers use an electronic medical record. Digitization of the medical record provides an opportunity to harness medical data for analysis in unique ways. Hospitals are utilizing their electronic health record (EHR) systems at a local level to query reports for various indications of interest. Data from the EHR can document patient demographics along with validated PROMIS (Patient-Reported Outcomes in Registries) measures, which can be used to populate clinical registries. In Table 1, we show how we use specialized algorithms to extract key comorbidity data from the EHR in a standardized way. The 9-Item Patient Health Questionnaire (PHQ-9) and the 7-Item Generalized Anxiety Disorder scale (GAD-7) have scores that can be extracted from the EHR as indicators of depression and anxiety, respectively.14,19

In addition, we have been able to develop electronic data extraction techniques to assess opioid utilization in our patients treated for lumbar spondylolisthesis. In Fig. 1, we show a schematic that visualizes the flow of clinical data. Data extracted from the EHR are combined with PRO data in real time.

Furthermore, medical demographic data, including imaging data from picture archiving and communication systems (PACSs), can also be easily de-identified and uploaded into databases for analysis. The digitization of medical data has made it possible to include imaging data in patient registries and to analyze large numbers of patients with their imaging data. The SLIP II Registry is one such example (clinical trial no. NCT03570801; https://clinicaltrials.gov). In this registry, 15 major centers are enrolling 1000 patients with lumbar spondylolisthesis. Patient demographics along with validated PROs are being collected with a follow-up of at least 5 years. In addition, lumbar flexion and extension radiographs along with sagittal T2-weighted MRI with relevant axial images are stored with follow-up dynamic lumbar radiographs 1 and 2 years after treatment for each patient. This will permit a detailed analysis of radiographic imaging data with the goal of identifying additional risk factors for the development of instability. Figure 2 depicts a succinct example of what is presented to the expert panel for the SLIP II Registry.

### Patient-Reported Outcomes in Registries

The importance of the patient perspective in US healthcare is broadly recognized. There is a growing realization that patient experience, satisfaction, and outcome are essential to improving overall quality in spinal care. While process measures and 30-day complications data have helped to focus attention on improving quality in medicine, there are unique aspects to spinal care that demand the monitoring of long-term PROs. The most important aspects of spinal care are the restoration of function and productivity and reduction in pain. Only patients can provide that type of feedback. Thus, there is a need for practical, validated PRO tools in our quest to optimize outcomes within the realm of spinal care.

One of the challenges with the collection of PROs in registry efforts to measure quality of care is the burden placed on patients. The completion of validated PRO measures is often time-consuming, and many registries have suboptimal completion rates over time. In addition, many validated PRO measures require licenses and considerable expense for healthcare researchers to use on a routine basis. The NIH-funded Patient-Reported Outcomes Measurement Information System (PROMIS) has been developed and has been validated by multiple groups for both cervical and lumbar spinal disorders.4,18 There is no fee associated with or license required for the use of PROMIS. In addition, PROMIS employs computer adaptive testing (CAT) technology that reduces the number of questions that each patient must complete based on answers to previous questions. One recent study found that PROMIS Phys-

---

**TABLE 1. Clinical definition for data extraction from the EHR**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Indication</th>
<th>Into Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>Hemoglobin A1c</td>
<td>Value (outside normal range + for diabetes)</td>
</tr>
<tr>
<td>Depression</td>
<td>PHQ-9</td>
<td>Value (&gt;10 is + for depression)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>GAD-7</td>
<td>Value (5 = mild, 10 = moderate, 15 = severe)</td>
</tr>
<tr>
<td>Steroids</td>
<td>Immunosuppressives</td>
<td>If indicated, then + for steroids</td>
</tr>
<tr>
<td>Nicotine</td>
<td>Urine test</td>
<td>Nicotine, urine &gt;5.0 ng/ml; cotinine, urine &amp; cotinine, random urine &gt;5.0 ng/ml; anabasine, urine &gt;2.0 ng/ml; nornicotine, urine &gt;2.0 ng/ml</td>
</tr>
<tr>
<td>CAD</td>
<td>Diagnostic code value set†</td>
<td>If indicated, then + for CAD</td>
</tr>
<tr>
<td>Narcotics/opioids</td>
<td>Medication value set†</td>
<td>If indicated, then + for narcotics/opioids</td>
</tr>
<tr>
<td>Postop infection</td>
<td>Prescription of antibiotics w/in 30 days of surgery</td>
<td>If indicated, then + for infection &amp; flagged for data review</td>
</tr>
</tbody>
</table>

* = positive; CAD = coronary artery disease; GAD-7 = 7-Item Generalized Anxiety Disorder scale; PHQ-9 = 9-Item Patient Health Questionnaire.

† Medication value set for narcotics and opioids was created and implemented as a revised list of interest containing Schedule II drugs as defined by the US Drug Enforcement Administration (DEA).
Artificial Intelligence and Medicine

The promise of artificial intelligence (AI) for applications in medicine and healthcare is as widely talked about today as when the concept was first developed in the late 1960s. AI is specifically focused on the use of algorithms and the software that implements these algorithms to approximate human cognitive abilities for the analysis of complex observations and data. In medicine, and healthcare in general, the goal is to develop computer algorithms to approximate conclusions from real medical data, without direct human input.

AI works by modeling and extracting information from a variety of data sources, and then processing it with the goal of being able to give a well-defined and ideally understandable and explainable output to the practicing medical expert. AI leverages machine learning (ML) algorithms to accomplish this goal. ML is essentially an approach that recognizes patterns in data (observations, behaviors, etc.) and creates a model that best represents an equivalent logic of the data. In other words, ML algorithms build a mathematical model from sample data, known as “training data,” to make predictions or decisions without being explicitly programmed to perform the task.2

While the study of AI for medicine and healthcare is not new, there has been a recent upsurge in the excitement for using AI and ML in medicine as well as for the practical systems being developed and deployed that use some recently developed techniques in ML. These approaches rely on large amounts of healthcare data being collected and stored in forms of EHRs that also include rich medical imaging as well as information about medical history, drugs used, diagnosis, and options considered. One prominent example is the Informatics for Integrating Biology and the Bedside, or i2b2, which is a data-rich platform funded by the NIH in 2004.3 One of the advantages of the i2b2 project is that data from the EHR are combined with genomic data.16 The availability of such large amounts of data, now with different ML models and techniques, including convolutional neural networks (CNNs),13 support vector machines (SVMs), and random decision forests, may ultimately inform new algorithms utilized by physician-scientists to advance their practice of data-driven personalized medical care. Most notably, AI and ML have transformed technologies involved in the diagnosis of disease such as cancers, the management of medical records, and precision medicine, attracting interest in these topics from medical or healthcare companies as well as computing companies like Google and IBM.3,15

Building a Spine Database for ML

The goal in our current effort is to bring the full force of AI and ML to the problem of lumbar spinal surgery, specifically lumbar stenosis with degenerative spondy-
lolisthesis. We are motivated to explore this because of the medical need, but on the technology forefront, we see some explicit reasons that suggest the viability of this exploration given the 1) availability of data and 2) expert opinions. The SLIP II Registry will capture key elements of clinical data, radiographic data, and expert reviews for each case in order to assess an expert panel’s judgment regarding the benefit of adding fusion to a decompression or performing a decompression alone. The study will collect data on 1000 patients from 15 sites in the US with 10–15 expert reviews per case. In Fig. 3, we show an example of the expert panel review from the SLIP II study. The baseline images for this patient were presented in Fig. 2. These images were sent to the SLIP II expert panel for review.

One possible technique for judging if a pattern exists is to determine if several people “see” the same thing when presented with a data set. For example, if a group of experts examined the same imaging and demographic data for a patient with grade I spondylolisthesis and believed that there were signs of instability, there might be an opportunity to analyze these data using modern computing methodology to help identify the pattern being recognized. In the Cervical Spondylotic Myelopathy Surgical (CSM-S) study, each patient had an expert panel review to determine if there was clinical equipoise to justify randomization to either ventral or dorsal surgery. In that trial, 269 patients were enrolled and 163 patients were randomized. The patients who were not randomized were also included in a separate cohort for analysis. There were 84 cases (roughly one-third of cases) in which 80% or more of the experts voted for a particular strategy. These patients were not
randomized. We examined the 1-year outcomes of cases in which the patients and surgeons were aligned with the majority opinion versus when they chose to do the opposite of what the majority had judged to be optimal. In Fig. 4, the data for 61 patients with 1-year follow-up are represented.

In the SLIP II study, patients are randomized to have an expert panel review of their imaging data. Thus, we will have data on 500 patients with expert panel reviews. Based on our preliminary data, it is likely that one-third of cases will have a “pattern” that will be recognized by a significant majority of the experts. We believe that these types of cases will inform our AI colleagues as they develop predictive algorithms. The imaging data are being stored in Extensible Markup Language (XML) format in order to be analyzable by the ML teams.

Developing AI for Lumbar Spondylolisthesis

In the previous section, we highlighted the collection of data underway and the challenges associated with the analysis of said data that suggest a need for automated classification of this data. The availability of expert annotation makes this a classic supervised ML problem, though as we continue to add data, we will also consider explorations into semi-supervised and unsupervised learning methods.

Supervised ML Using Support Vector Machines

A classic method for supervised ML is the use of support vector machines (SVMs), which provide models with associated learning algorithms that analyze data and support classification and regression analysis. An SVM algorithm analyzes training examples, marked as belonging to one or the other of two categories (e.g., recommended for fusion or not recommended for fusion), and builds a model that assigns new examples to one category or the other (Fig. 5). This is the simplest case of a nonprobabilistic binary linear classifier. Overall, an SVM model is a representation of the examples from the training data as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New incoming examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Exploration of Deep Learning Methods

Our initial goal will be to employ classic methods like...
SVMs for our task of automated classification. However, as our data size increases, with needs for improved accuracy, we will also consider exploring deep learning methods. These methods have recently had much success, specifically as they are a broader set of algorithms and not constrained to be task specific. The hierarchical representation usually employed by the deep learning algorithms allows for a more robust feature of learning from data, as multiple levels (it is this transition from higher to lower levels that makes these algorithms “deep”). Deep learning algorithms have showcased results comparable to, if not better than, human experts in many domains, including bioinformatics and drug discovery. However, deep learning methods are data and annotation hungry; that is, lots of annotated data are required—something we will address in the later parts of our study as we collect more rich data (Fig. 6).\(^7\)

References


Disclosures

Dr. Ghogawala has received support from the Alan L. and Jacqueline B. Stuart Spine Research Center and has ownership in GN2.0-Nidus, Ms. Dunbar owns stock in GN2.0-Nidus, and Dr. Essa is an employee of Google.

Author Contributions

Conception and design: all authors. Acquisition of data: Ghogawala. Analysis and interpretation of data: Ghogawala. Drafting the article: all authors. Critically revising the article: all authors. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Ghogawala. Study supervision: Ghogawala.

Correspondence

Zoher Ghogawala: Lahey Hospital & Medical Center, Burlington, MA. zoher.ghogawala@lahey.org.