Journal of Orthopaedics 35 (2023) 74–78

Contents lists available at ScienceDirect

Journal of Orthopaedics

journal homepage: www.elsevier.com/locate/jor

Development of a machine learning algorithm to identify total and reverse shoulder arthroplasty implants from X-ray images

Eric A. Geng, Brian H. Cho, Aly A. Valliani, Varun Arvind, Akshar V. Patel, Samuel K. Cho, Jun S. Kim *, Paul J. Cagle

Department of Orthopaedic Surgery, Mount Sinai Health System, New York, NY, 10029, USA

ARTICLE INFO

Keywords:
Total shoulder arthroplasty
Machine learning
Mobile application
Digital medicine
Implant identification

ABSTRACT

Introduction: Demand for total shoulder arthroplasty (TSA) has risen significantly and is projected to continue growing. From 2012 to 2017, the incidence of reverse total shoulder arthroplasty (rTSA) rose from 7.3 cases per 100,000 to 19.3 per 100,000. Anatomical TSA saw a growth from 9.5 cases per 100,000 to 12.5 per 100,000. Failure to identify implants in a timely manner can increase operative time, cost and risk of complications. Several machine learning models have been developed to perform medical image analysis. However, they have not been widely applied in shoulder surgery. The authors developed a machine learning model to identify shoulder implant manufacturers and type from anterior-posterior X-ray images.

Methods: The model deployed was a convolutional neural network (CNN), which has been widely used in computer vision tasks. 696 radiographs were obtained from a single institution. 70% were used to train the model, while evaluation was done on 30%.

Results: On the evaluation set, the model performed with an overall accuracy of 93.9% with positive predictive value, sensitivity and F-1 scores of 94% across 10 different implant types (4 reverse, 6 anatomical). Average identification time was 0.110 s per implant.

Conclusion: This proof of concept study demonstrates that machine learning can assist with preoperative planning and improve cost-efficiency in shoulder surgery.

1. Introduction

Total shoulder arthroplasty (TSA) is among the most commonly performed joint replacement procedures in the United States to treat glenohumeral osteoarthritis, humerus fractures, and other related joint diseases and traumas. From 2012 to 2017, the incidence of reverse total shoulder arthroplasty (rTSA) increased from 7.3 cases per 100,000 to 19.3 cases per 100,000, while anatomical total shoulder arthroplasty (aTSA) increased from 9.5 cases per 100,000 to 12.5 cases per 100,000.[1] One study estimated a 122% growth in the overall demand of TSA by 2040, with rTSA projected to increase at a faster rate than aTSA.[2] However, there has also been an increase of revision arthroplasties.[3,4] The most common reasons for shoulder arthroplasty failure include glenoid component failure and rotator cuff/subscapularis tear for aTSA and dislocation/instability and infection for rTSA.[5] There is also considerable variation in failure modes across implant manufacturers.[5]

Indications for revision arthroplasty include component loosening, infection, or trauma.[6] Although shoulder arthroplasties are expected to last 10–15 years, one recent review found a mean time to revision of 3.9 years from the initial procedure.[7] Accurate and timely identification of implant types is critical for preoperative planning. This is because surgeons will need to obtain specific extraction equipment depending on the implant type for proper revision. Surgical technique may also differ depending on the implant. Implants are typically identified through manual image analysis by a surgeon or other medical expert. This can be a time-intensive process and requires detailed knowledge of each implant type. It is especially challenging with more obscure or outdated implants that require more extensive research and labor to properly identify. Additional difficulties arise when patients transition between institutions. Approximately 30% of readmissions following TSA take place at a different hospital than that of the original arthroplasty.[8] The methodology and detail used to document implants is not standardized across hospitals, which complicates the identification process. It has only been in recent years that the United States Food and Drug

* Corresponding author. Department of Orthopedic Surgery, Icahn School of Medicine at Mount Sinai, 425 West 59th Street, 5th floor, New York, NY, 10019, USA.
E-mail address: jun.kim@mountsinai.org (J.S. Kim).

https://doi.org/10.1016/j.jor.2022.11.004
Received 21 June 2022; Received in revised form 16 October 2022; Accepted 7 November 2022
Available online 11 November 2022

Administration has mandated the use of Unique Device Identifiers (UDIs), which underscores the issue of device identification in the US healthcare system.[9] To the authors' knowledge, no studies have specifically analyzed the issue of implant identification in TSA. However, a study on orthopedic implant replacement found the median time of identification to be 20 min with approximately 10% of implants being unidentified preoperatively.[10] Failure to accurately identify an implant can lead to increased operative time, complications, and cost.[11] With rising healthcare costs and bundled payment models, it is essential to improve efficiency and maximize time expenditure for shoulder surgeons and staff. Implant identification is an inefficiency in shoulder arthroplasty that can be enhanced to improve the overall cost effectiveness of care.

Machine learning can address the issue of implant identification through automated image classification algorithms. These models excel at analyzing complex image data in comparison to traditional statistical methods. Machine learning has been widely applied in many medical specialties. In orthopedics specifically, models have been developed to classify fractures[12,13] and spinal deformities[14,15] with a high degree of accuracy. Therefore, the authors hypothesized that this technology can be used to identify shoulder implants from radiographs. This is a proof of concept study demonstrating the ability of machine learning to automatically classify reverse and anatomical shoulder implants based on anterior posterior (AP) X-ray images. The technology may be used to assist with preoperative planning, particularly with revision surgeries.

2. Materials and methods

2.1. Image dataset

Institutional review board approval was obtained prior to the start of this study. 696 AP radiographs of shoulder arthroplasties, containing 10 different types of shoulder implants, were obtained from a single institution through the picture archiving and communication system (PACS). Criteria for image selection included AP radiographs of patients aged 18 or older who underwent primary or reverse total shoulder arthroplasty. Images of poor quality or with significant artifacts were excluded from analysis. Implants were identified using operative notes, case implant reports, and review from a fellowship trained shoulder surgeon. The final data set consisted of four reverse and six anatomical implants. The full list of implants is shown in Table 1. Not all implants in use across the United States were available for the model because selection was limited to those used at the single institution. Images were randomly partitioned into 70% train and 30% test datasets. The final model was evaluated on the test dataset, which consisted of images the model had never seen. For preprocessing, images were converted to grayscale and reshaped into square sized images of dimension 224x224.

2.2. Machine learning algorithm

The machine learning algorithm used in this study was a variant of densenet121,[16] a pretrained convolutional neural network (CNN). This type of model has been widely used for computer vision tasks and has demonstrated excellent proficiency in complex image analysis. A CNN scans through an image using filters, detecting key features such as edges and corners. The information is propagated through several layers and then synthesized to reach a classification decision (implant identification in this context). A CNN learns by seeing hundreds of images and their corresponding labels, making incremental improvements over several iterations. A schematic for the model is shown in Fig. 1. 484 implants were used to train the CNN. The model was trained for a maximum of 200 epochs with early stopping if no improvement was seen in validation accuracy for 60 consecutive epochs. The model was then evaluated on 212 test implants that it previously had not been seen before.

Table 1
Performance metrics for individual implants.

Implant	N	PPV	Sensitivity	F1-Score
Arthrex Univers Apex	11	0.92	1.00	0.96
Arthrex Univers II	3	1.00	0.67	0.80
Arthrex Univers Reverse	9	1.00	0.67	0.80
Depuy Delta Xtend Reverse	2	0.50	0.50	0.50
Stryker Reverse	23	1.00	0.87	0.93
Stryker Total	6	0.86	1.00	0.92
Zimmer Bigliani Flatow (Zimmer BF)	61	0.97	0.98	0.98
Zimmer Trabecular Metal Reverse	79	0.93	0.99	0.96
Zimmer Trabecular Metal	12	0.91	0.83	0.87
Zimmer Trabecular Metal Glenoid	6	0.83	0.83	0.83

2.3. Performance metrics

Model performance on the test set was assessed using the following metrics: accuracy, positive predictive value (PPV), sensitivity, and F-1 score. The F-1 score is an average of PPV and sensitivity. A value of 1.0 represents a perfect F1-score. In the problem of classifying shoulder implants, each implant type may only occur a few times per 100 images. This yields a large ratio of true negatives to true positives for a given implant, which inflates accuracy metrics and makes them less useful for evaluating model performance. The PPV and sensitivity metrics address this issue by focusing on the positive classifications, which are important to scrutinize closely. Since both of these metrics are important in a strong classifier, the F-1 score is the most useful metric for model performance despite its less intuitive nature. Metrics were generated for overall and implant-specific performance. Saliency heat maps were also generated, which highlight the most important elements of the implant that best inform model decisions. This provides added transparency to the model's analytical process and helps further validate the model.

2.4. Software

Python (v. 3.7) and the open source machine learning library Keras (v. 2.2.2) were used for algorithm development. The code was adapted from a model the authors' previously created to identify hip implants from X-rays.

3. Results

The algorithm demonstrated an overall accuracy of 93.9% for 10 shoulder implants on the test dataset. Likewise, overall PPV, sensitivity, and F-1 score were all 0.94 as well. The algorithm identified implants from 212 test radiographs in 23.37 s, averaging 0.110 s per image. Performance metrics and a confusion matrix for individual implants are shown in Table 1 and Fig. 2, respectively. The top performing implants were Arthrex Univers Apex, Zimmer Trabecular Metal Reverse, and Zimmer Bigliani-Flatow (Zimmer BF) with scores of 0.92 and above for all three metrics. The model achieved a high PPV but low sensitivity with Arthrex Univers II and Arthrex Univers Reverse. Depuy Delta Xtend Reverse was the worst performing implant with a score of 0.50 for all metrics due to data limitations.

Saliency maps (Fig. 3) indicate the areas of highest importance when the model performs classification and provide transparency on the model's "thought" process. This ensures that it is not simply analyzing arbitrary patterns and adds to the validity of the model. The mapping demonstrates that in correct cases the model examines logical features of the implant, such as the neck and head. These are key elements that a surgeon would use to discern the implant type. In incorrect cases, the model tended to focus away from the implant towards the periphery of the image.

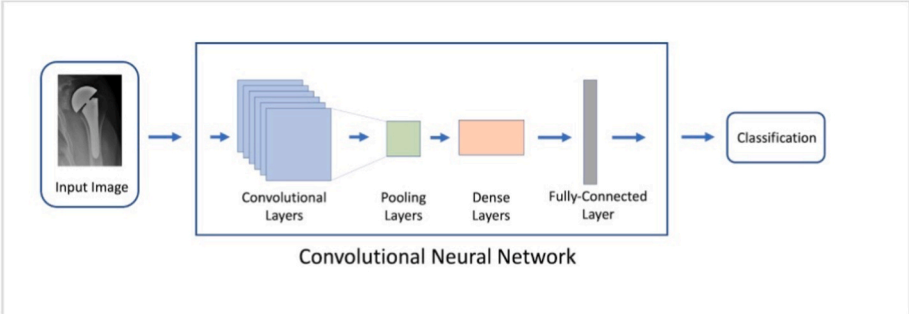

Fig. 1. Simplified schematic representation of the convolutional neural network pipeline used for image classification.

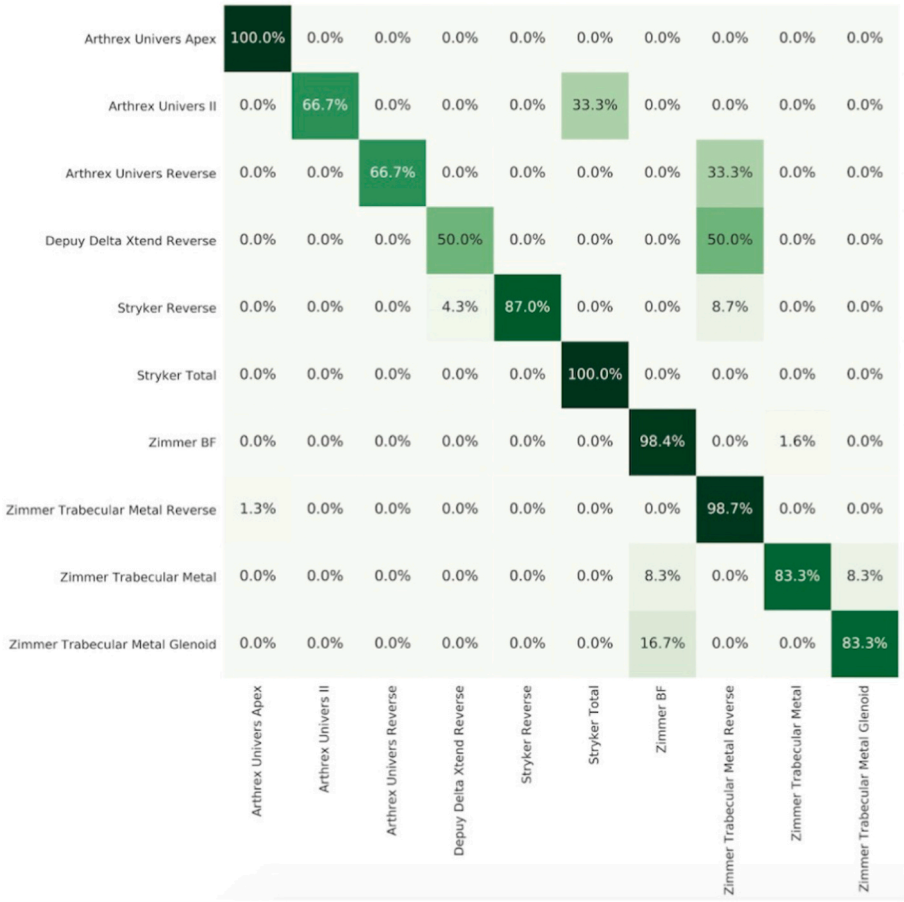

Fig. 2. Normalized confusion matrix of algorithm predictions. Values are expressed as a percentage of true class images. True class is on the y-axis and predicted class is on the x-axis. Darker shading indicates higher values.

4. Discussion

In this study, the authors validated a machine learning algorithm for the automated identification of 10 reverse and anatomical shoulder implants from X-ray images. The model performed strongly with high overall accuracy, sensitivity, PPV, and F1-score. Overall accuracy was approximately 9.4 times greater than random chance (1/10). In addition, the model was also able to accurately classify and distinguish anatomical and reverse implants by manufacturer and implant name. This proof of concept study provides evidence suggesting that a machine learning algorithm may be used to assist shoulder surgeons with implant identification in real-world clinical settings. With a larger dataset and a more diverse array of implant classes, this technology may be further refined for clinical use.

Clinical application of a machine learning algorithm for the identification of shoulder implants could be of great benefit to the busy joint reconstruction surgeon's workflow. The current paradigm for implant identification involves substantial time investment, and the process may require the input of several parties to arrive at a firm conclusion. Wilson et al. estimated that the median identification time for an orthopedic implant was 20 min.[10] By comparison, the algorithm presented in this work identified each implant in 0.110 s on average. Additionally, 87% of surgeons reported using a minimum of three methods to identify an implant.[10] A well-made identification algorithm could become the only required method for identification. The cumulative surgeon time used for implant identification is projected to be over 133,000 h in 2030, which is the equivalent of 275,000 15 min office visits in 2020.[11] Clinical implementation of automated implant identification algorithms

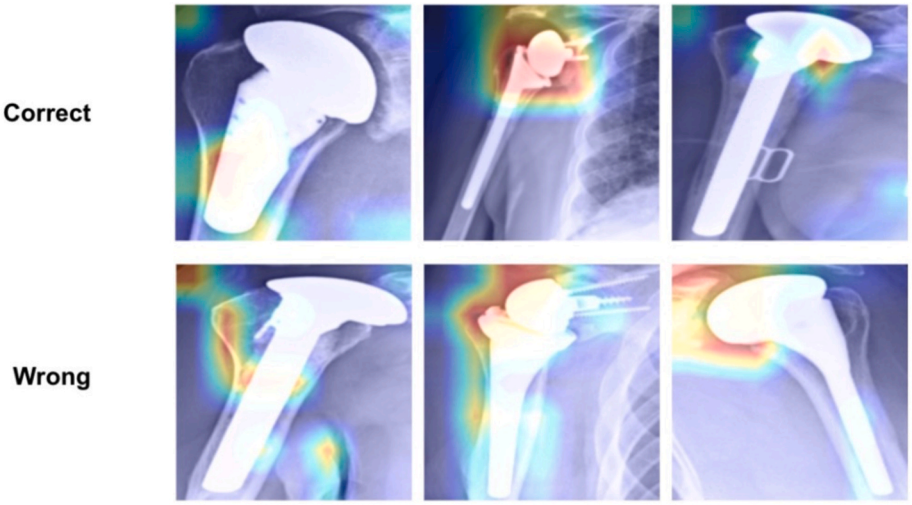

Fig. 3. Saliency mapping showing correct and incorrect identifications.

could virtually eliminate this time investment and allow surgeons to reinvest valuable time elsewhere.

Other studies have attempted to develop a machine learning algorithm to identify shoulder implants. Urban et al. trained a machine learning model to classify shoulder implants from X-rays on the basis of four manufacturers, achieving an overall accuracy of 80%.[18] The authors' algorithm, in contrast, identifies more implants with greater accuracy. This added granularity provides greater clinical utility as several implant types exist per manufacturer. These have been produced over varying decades and will likely require differing equipment and procedural techniques. The authors' model additionally demonstrates the ability to classify reverse implants, which constituted 4 of the 10 implants in the sample. This is particularly significant moving forward because rTSA is outgrowing aTSA in terms of utilization.[2] For patients 85 and older, rTSA is expected to grow by 120%, while aTSA is projected to decline by 20%.[2] Therefore, shoulder surgeons will increasingly encounter reverse implants, making it critical for an algorithm to accurately classify both reverse and anatomical implants. Yi et al. also developed a model to classify five shoulder implant models from X-ray images with high accuracy and area under the curve (AUC).[19] However, they trained five separate binary classifiers (one for each implant type), thereby limiting clinical utility at scale. Their model was able to distinguish whether an implant was reverse or anatomical, but not by specific manufacture or type. The authors' model improves upon that of Yi et al. First, this model demonstrates strong proficiency in classifying up to 10 implant types, including reverse implants. Second, the model is a single multi-classifier model, which is more memory efficient and therefore more readily deployable in clinical practice. Machine learning algorithms are computationally expensive and efficiency of design must be taken into account for real world application. Efficient algorithms enable deployment on smaller platforms such as consumer computers or smartphones. Despite the presented algorithm's strong performance in implant identification, several limitations remain. First, class imbalance was an issue with certain implants being underrepresented in the dataset. Model performance suffered with these implants in particular. A smaller sample size results in fewer examples for the algorithm to learn from, which ultimately results in lower accuracy and potentially biases predictions toward majority classes. This was accounted for in part by implementing class weights while training the model, where certain implant types are weighted more or less heavily depending on their prevalence in their dataset. However, the gold standard solution would simply be to increase the number of minority classes. This should be an aim for future studies. Second, this study was limited to X-rays from a single institution. This limits the selection of images, as not all implant types in the United States are used. Factors such as X-ray quality and technique also vary across institutions. Future studies should consider the possibility of using multi-institutional data from a wide selection of regions to enhance implant diversity in their dataset. There are, however, significant challenges to broad implant data acquisition. Radiographs are typically siloed by institutions due to HIPAA requirements. In addition, imaging data is generally memory intensive, so storage and transfer of large quantities poses a significant logistical and computational burden. Ideally, the finalized version of this software for clinical use should include the majority of, if not all, implants used across the United States. This study demonstrates the potential of machine learning in automating implant identification, and this initial model may be further developed for real time clinical use.

5. Conclusion

Rapid identification of shoulder implants is a critical part for preoperative planning in shoulder arthroplasties and revisions. In this work, the authors' present a machine learning algorithm that is capable of identifying shoulder implants from AP plain films with high accuracy and speed. This work improves upon previous studies by classifying more implant types, including reverse implants, with equal or greater performance. While this is a proof of concept study, the technology may be augmented with the addition of more implant types and ultimately be leveraged to promote cost-effective care in shoulder arthroplasty procedures.

Funding/sponsorship

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Informed consent (patient/guardian)

N/A.

Institutional ethical committee approval

This study was approved by the Institutional Review Board at the [*Omitted]*

CRediT authorship contribution statement

Eric A. Geng: Conceptualization, Writing – original draft, Software, Formal analysis. **Brian H. Cho:** Conceptualization, Writing – original draft, Software, Formal analysis. **Aly A. Valliani:** Writing – review &

editing, Software. **Varun Arvind:** Writing – review & editing, Software. **Akshar V. Patel:** Data curation. **Samuel K. Cho:** Conceptualization, Supervision. **Jun S. Kim:** Conceptualization, Supervision. **Paul J. Cagle:** Conceptualization, Supervision.

Declaration of competing interest

None.

References

1 Best MJ, Aziz KT, Wilckens JH, McFarland EG, Srikumaran U. Increasing incidence of primary reverse and anatomic total shoulder arthroplasty in the United States. *J Shoulder Elbow Surg.* 2020. https://doi.org/10.1016/j.jse.2020.08.010. Published online August 25.

2 Rabinowitz J, Kothandaraman V, Lin J, Li X, Friedman RJ, Eichinger JK. Utilization of shoulder arthroplasty in the United States – an analysis of current trends and future predictions. *Semin Arthroplasty: JSES.* 2020;30(3):200–209.

3 Schwartz BE, Savin DD, Youderian AR, Mossad D, Goldberg BA. National trends and perioperative outcomes in primary and revision total shoulder arthroplasty: trends in total shoulder arthroplasty. *Int Orthop.* 2015;39(2):271–276.

4 Day JS, Lau E, Ong KL, Williams GR, Ramsey ML, Kurtz SM. Prevalence and projections of total shoulder and elbow arthroplasty in the United States to 2015. *J Shoulder Elbow Surg.* 2010;19(8):1115–1120.

5 Somerson JS, Hsu JE, Neradilek MB, Matsen 3rd FA. Analysis of 4063 complications of shoulder arthroplasty reported to the US Food and Drug Administration from 2012 to 2016. *J Shoulder Elbow Surg.* 2018;27(11):1978–1986.

6 Ravi V, Murphy RJ, Moverley R, Derias M, Phadnis J. Outcome and complications following revision shoulder arthroplasty : a systematic review and meta-analysis. *Bone Jt Open.* 2021;2(8):618–630.

7 Knowles NK, Columbus MP, Wegmann K, Ferreira LM, Athwal GS. Revision shoulder arthroplasty: a systematic review and comparison of North American vs. European outcomes and complications. *J Shoulder Elbow Surg.* 2020;29(5):1071–1082.

8 Schairer WW, Zhang AL, Feeley BT. Hospital readmissions after primary shoulder arthroplasty. *J Shoulder Elbow Surg.* 2014;23(9):1349–1355.

9 Barlas S. New FDA medical device rule imposes minimal burden on hospitals: facilities able to scan unique device Identifiers will benefit. *Pharm Therapeut.* 2013;38 (12):720.

10 Wilson NA, Jehn M, York S, Davis 3rd CM. Revision total hip and knee arthroplasty implant identification: implications for use of Unique Device Identification 2012 AAHKS member survey results. *J Arthroplasty.* 2014;29(2):251–255.

11 Wilson N, Broatch J, Jehn M, Davis 3rd C. National projections of time, cost and failure in implantable device identification: consideration of unique device identification use. *Health.* 2015;3(4):196–201.

12 Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiology: Artif Intell.* 2020;2(2), e190023.

13 Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med.* 2019;2:31.

14 Yang J, Zhang K, Fan H, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. *Commun Biol.* 2019;2:390.

15 Galbusera F, Niemeyer F, Wilke HJ, et al. Fully automated radiological analysis of spinal disorders and deformities: a deep learning approach. *Eur Spine J.* 2019;28(5): 951–960.

16 Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. arXiv [csCV]. Published online http://arxiv.org/abs/1608.06993. Accessed August 25, 2016.

18 Urban G, Porhemmat S, Stark M, Feeley B, Okada K, Baldi P. Classifying shoulder implants in X-ray images using deep learning. *Comput Struct Biotechnol J.* 2020;18: 967–972.

19 Yi PH, Kim TK, Wei J, et al. Automated detection and classification of shoulder arthroplasty models using deep learning. *Skeletal Radiol.* 2020;49(10):1623–1632.

17 AIdentyfy surgical planning. https://apps.apple.com/us/app/aidentyfy-surgical-planning/id1524478136. Accessed December 3, 2020.