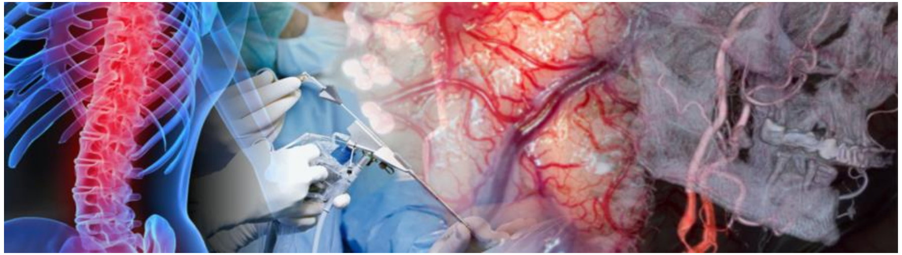
Výzkum a klinické aplikace umělé inteligence - naše zkušenost

Martin Černý
Neurochirurgická a neuroonkologická klinika
Ústřední vojenská nemocnice Praha

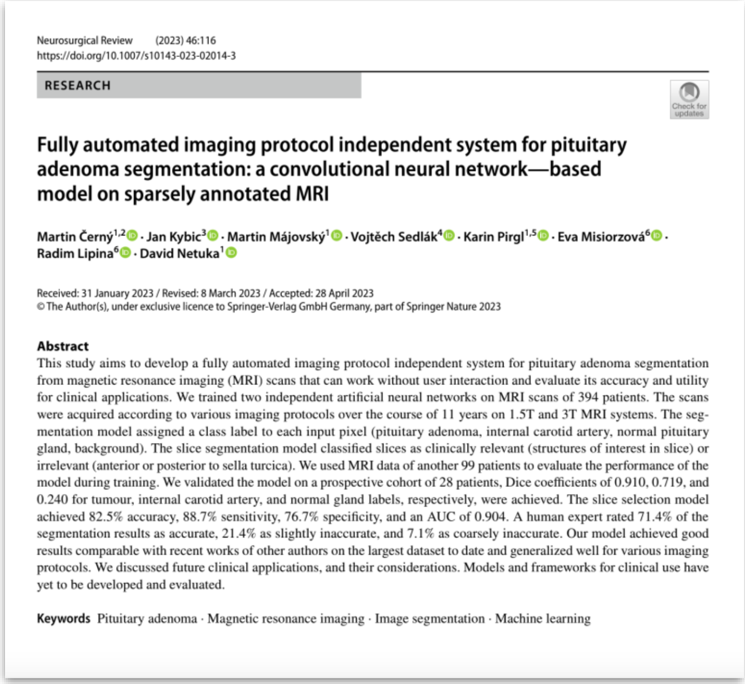
Neurochirurgická a neuroonkologická klinika ÚVN

- přes 3 000 operací ročně
- moderní přístrojové vybavení
- ozařovač CyberKnife
- intraoperační MRI
- rozsáhlá výzkumná aktivita



Témata AI výzkumu na naší klinice

- zpracování obrazových dat
- automatická segmentace
- automatická tvorba ozařovacích plánů
- využití velkých jazykových modelů
- extrakce informací z nestrukturovaných dat
- radiomika


Neurosurgical Review (2023) 46:116
https://doi.org/10.1007/s10143-023-02014-3

RESEARCH

Check for updates

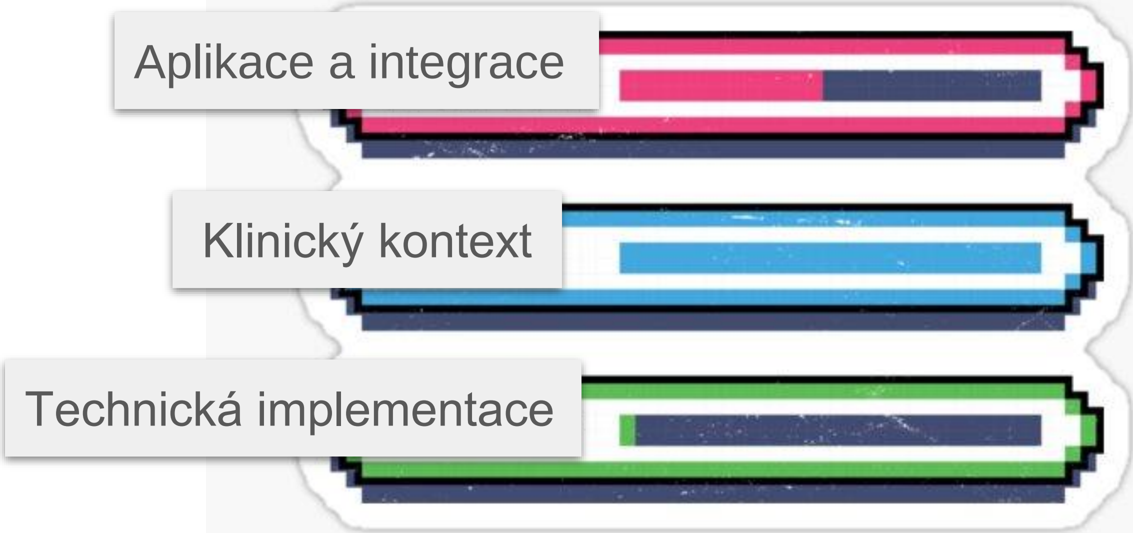
Fully automated imaging protocol independent system for pituitary adenoma segmentation: a convolutional neural network—based model on sparsely annotated MRI

Martin Černý[1,2] · Jan Kybic[3] · Martin Májovský[1] · Vojtěch Sedlák[4] · Karin Pirgl[1,5] · Eva Misiorzová[6] · Radim Lipina[6] · David Netuka[1]

Received: 31 January 2023 / Revised: 8 March 2023 / Accepted: 28 April 2023

Abstract
This study aims to develop a fully automated imaging protocol independent system for pituitary adenoma segmentation from magnetic resonance imaging (MRI) scans that can work without user interaction and evaluate its accuracy and utility for clinical applications. We trained two independent artificial neural networks on MRI scans of 394 patients. The scans were acquired according to various imaging protocols over the course of 11 years on 1.5T and 3T MRI systems. The segmentation model assigned a class label to each input pixel (pituitary adenoma, internal carotid artery, normal pituitary gland, background). The slice segmentation model classified slices as clinically relevant (structures of interest in slice) or irrelevant (anterior or posterior to sella turcica). We used MRI data of another 99 patients to evaluate the performance of the model during training. We validated the model on a prospective cohort of 28 patients, Dice coefficients of 0.910, 0.719, and 0.240 for tumour, internal carotid artery, and normal gland labels, respectively, were achieved. The slice selection model achieved 82.5% accuracy, 88.7% sensitivity, 76.7% specificity, and an AUC of 0.904. A human expert rated 71.4% of the segmentation results as accurate, 21.4% as slightly inaccurate, and 7.1% as coarsely inaccurate. Our model achieved good results comparable with recent works of other authors on the largest dataset to date and generalized well for various imaging protocols. We discussed future clinical applications, and their considerations. Models and frameworks for clinical use have yet to be developed and evaluated.

Keywords Pituitary adenoma · Magnetic resonance imaging · Image segmentation · Machine learning


Aplikace a integrace
Klinický kontext
Technická implementace

Solution-to-a-problem approach

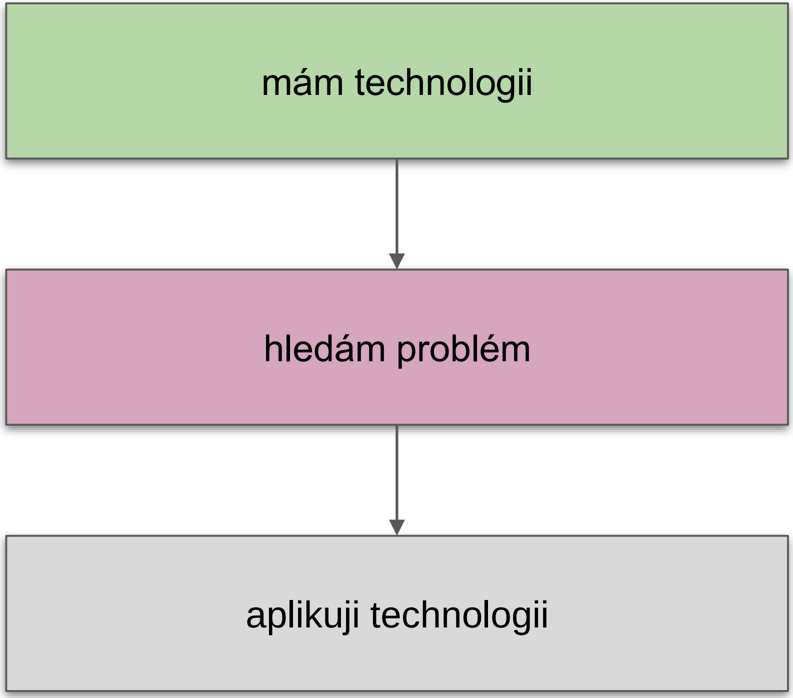
AI

- je aktuální téma
- dobře se publikuje
- dobře se získávají granty
- nadšení z nových technologií
- je cool

Solution-to-a-problem approach

AI

- je aktuální téma
- dobře se publikuje
- dobře se získávají granty
- nadšení z nových technologií
- je cool



Solution-to-a-problem approach

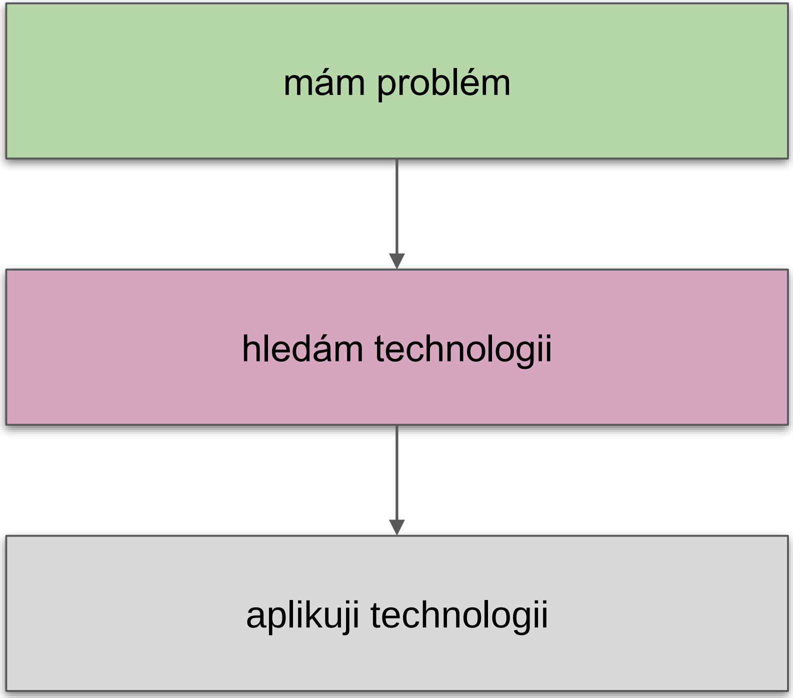
AI

- umožňuje přesnější diagnostiku
- šetří čas
- automatizuje rutinní úkoly
- demokratizuje přístup ke zdravotní péči

Solution-to-a-problem approach

AI

- umožňuje přesnější diagnostiku
- šetří čas
- automatizuje rutinní úkoly
- demokratizuje přístup ke zdravotní péči

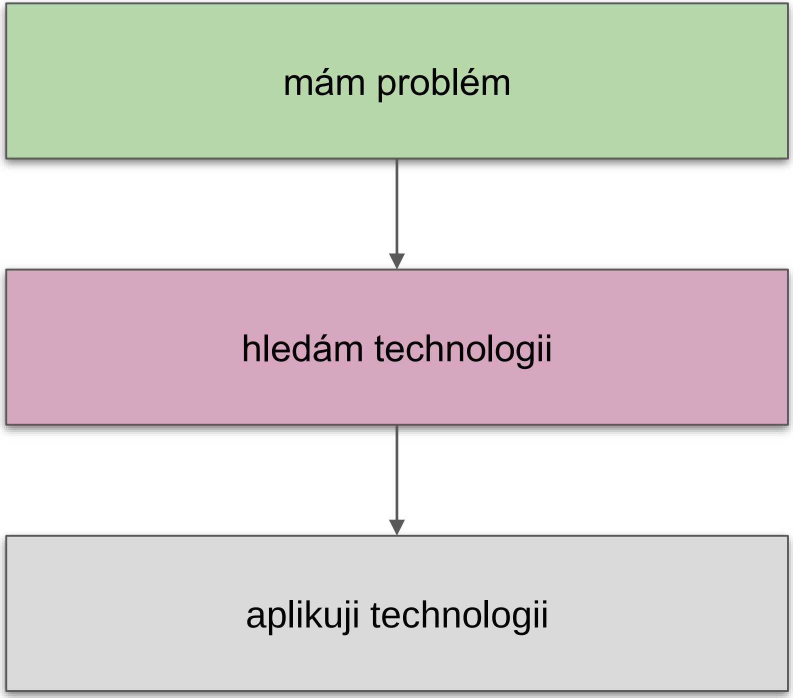


Solution-to-a-problem approach

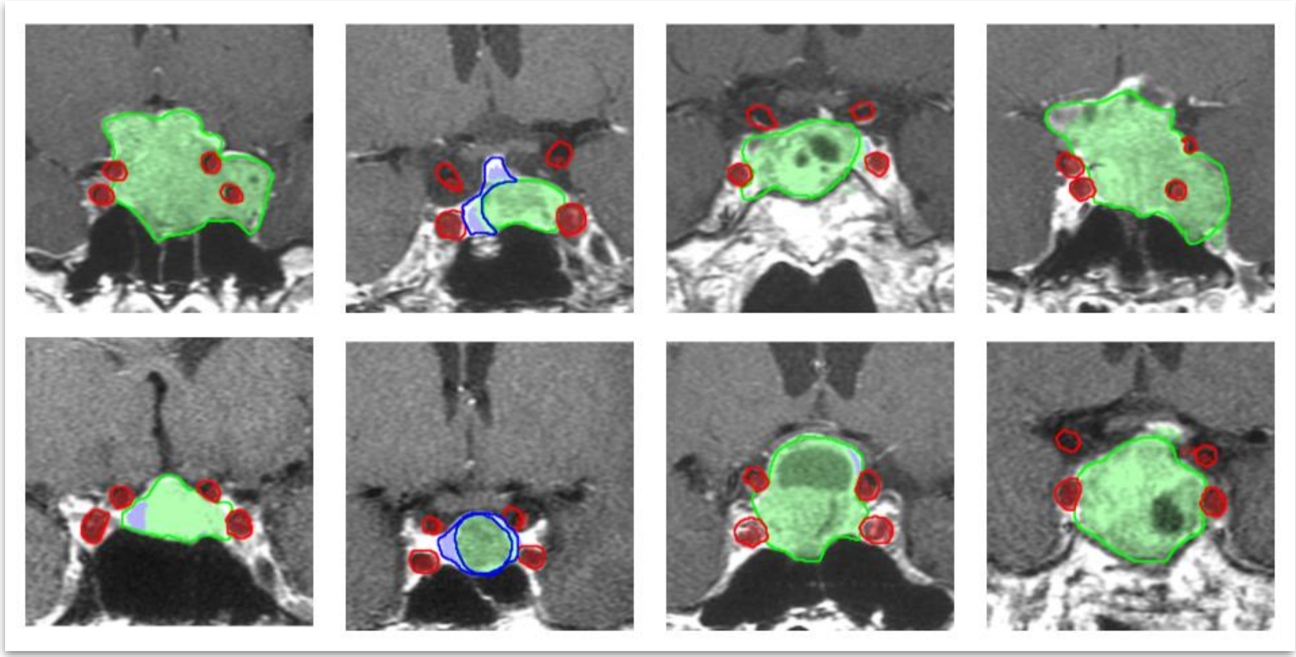
AI

- umožňuje přesnější diagnostiku?
- šetří čas?
- automatizuje rutinní úkoly?
- demokratizuje přístup ke zdravotní péči?

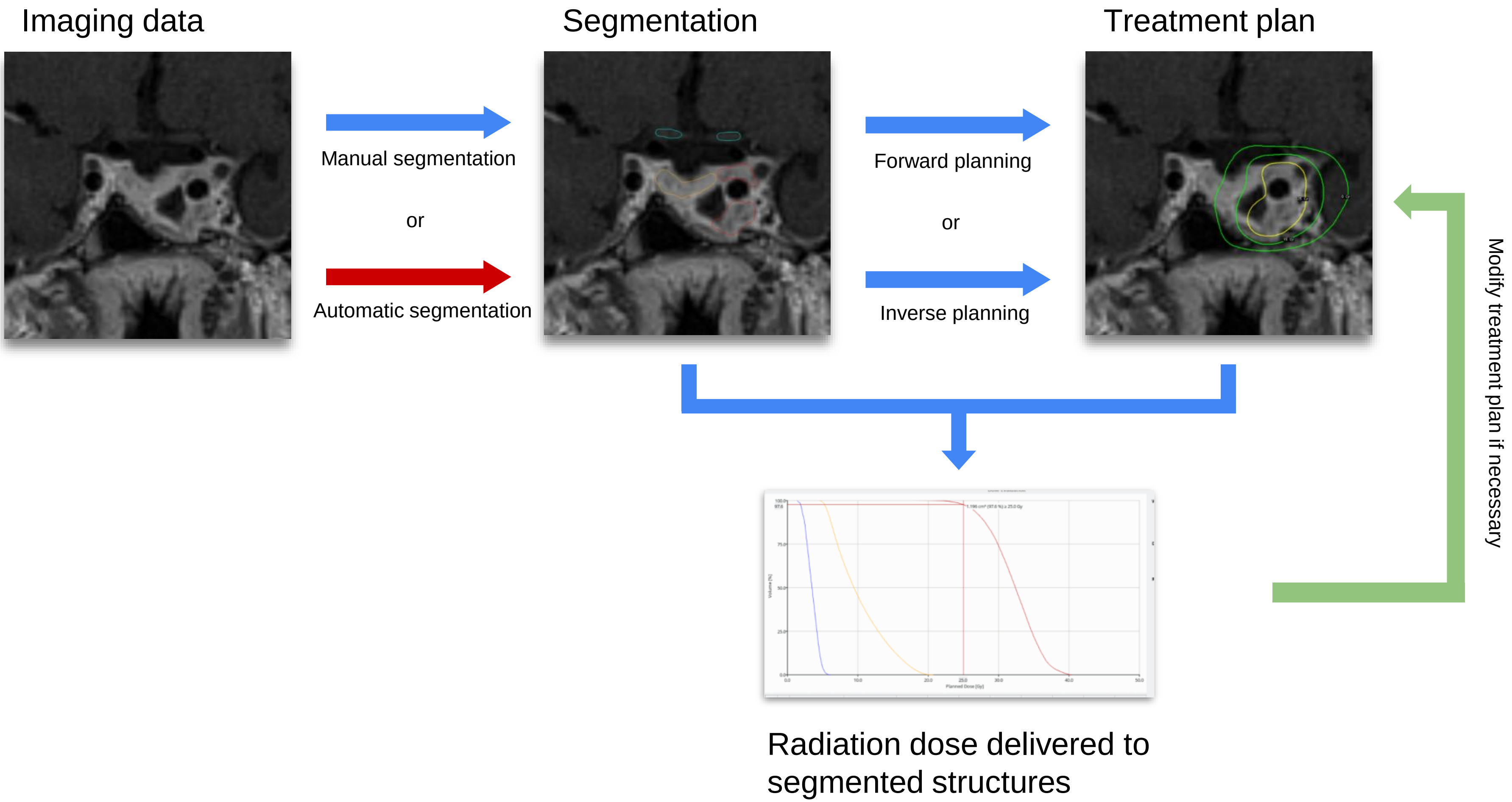
Hypotézy - náš úkol je je ověřit



Automatická segmentace adenomů hypofýzy

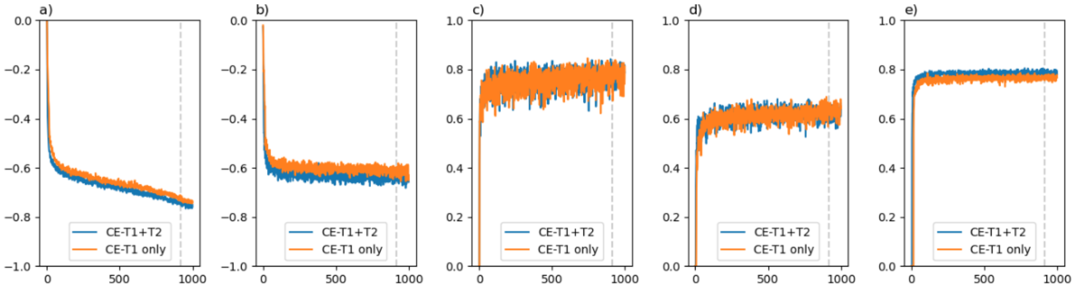


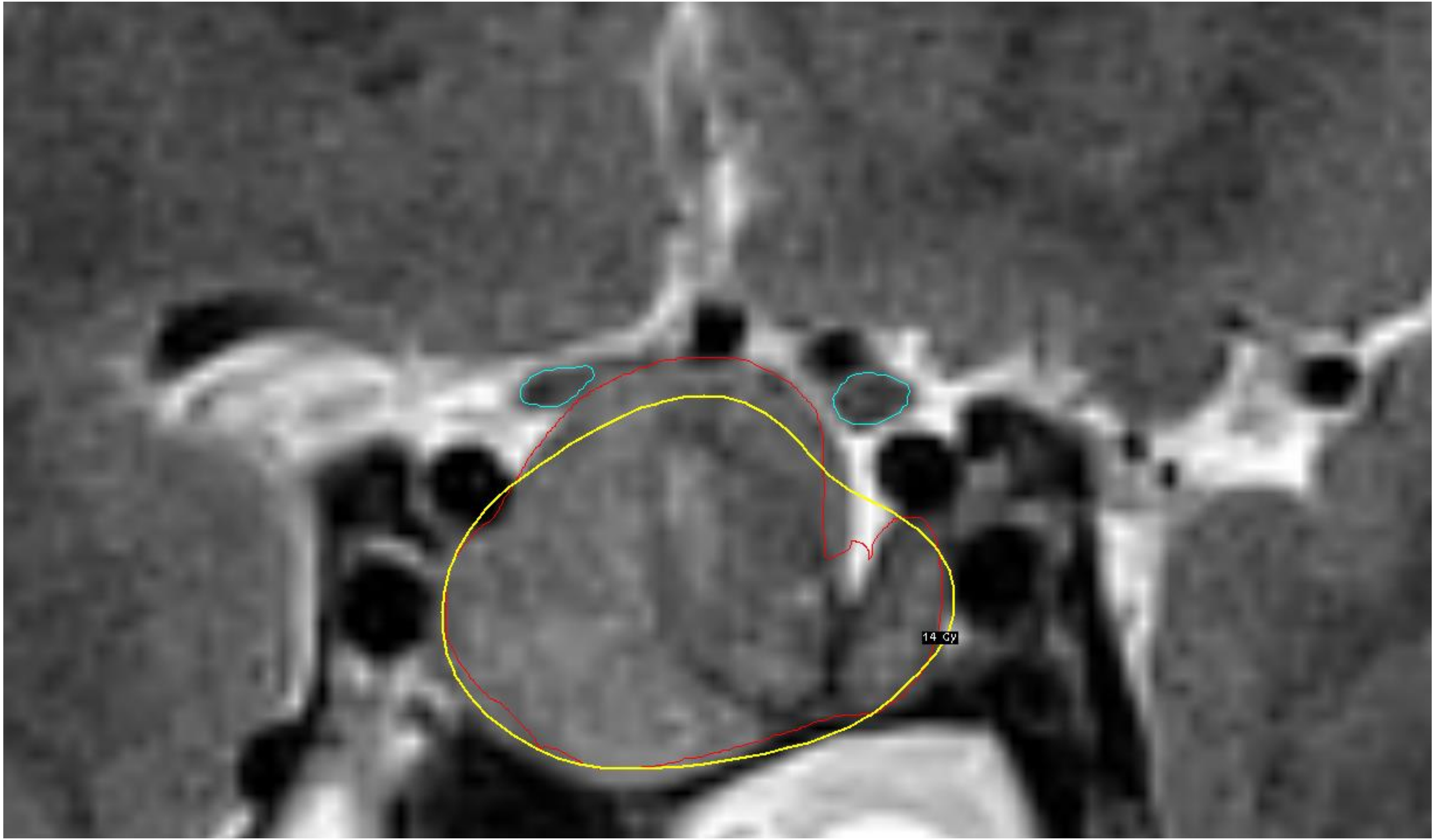
Černý M, Kybic J, Májovský M, et al. Fully automated imaging protocol independent system for pituitary adenoma segmentation: a convolutional neural network-based model on sparsely annotated MRI. Neurosurg Rev. 2023;46(1):116. Published 2023 May 10. doi:10.1007/s10143-023-02014-3


Imaging data
Segmentation
Treatment plan
Manual segmentation
or
Automatic segmentation
Forward planning
or
Inverse planning
Modify treatment plan if necessary
Radiation dose delivered to segmented structures

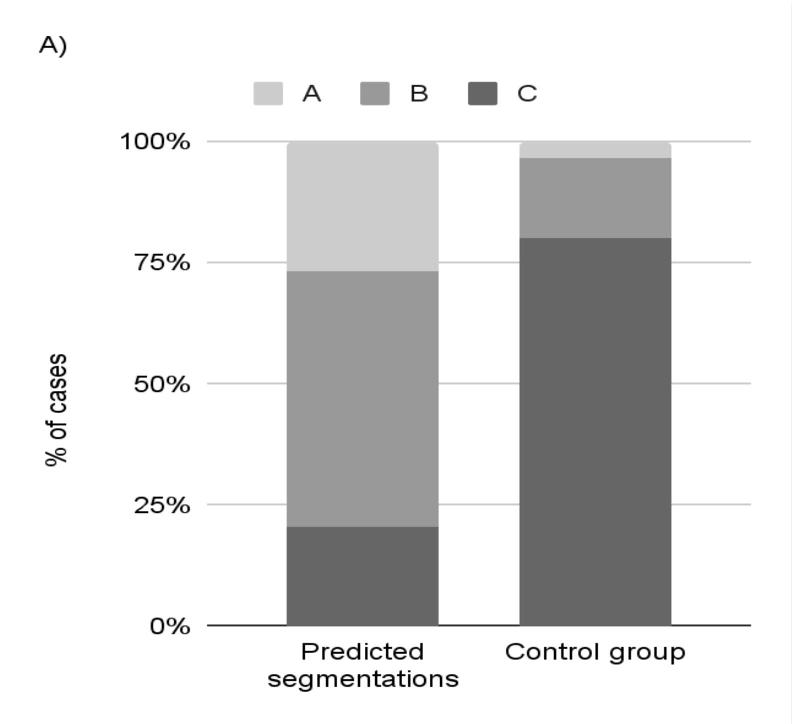
Jak spolehlivá je automatická segmentace?

The model reached the best performance after 915 training epochs and achieved DSC of 82.3%, 63.9% and 79.6% for the tumor, normal gland and optic nerve label classes on the validations dataset. The training took 20 hours and 55 minutes.
The CE-T1 only variant reached the best performance after 947 training epochs and achieved GSC of 79.0%, 63.9% and 76.4% for the respective label classes on the validation dataset. The training took 20 hours and 45 minutes.
Figure 4 illustrates the monitored model performance parameters over the course of the training. Figure 5 presents a comparison of ground truth and predicted segmentations on four examples of slices from the validation dataset. The trained model is publicly available from the main author's Figshare repository [31].




14 Gy

Jak spolehlivá je automatická segmentace?



Jak spolehlivá je automatická segmentace?

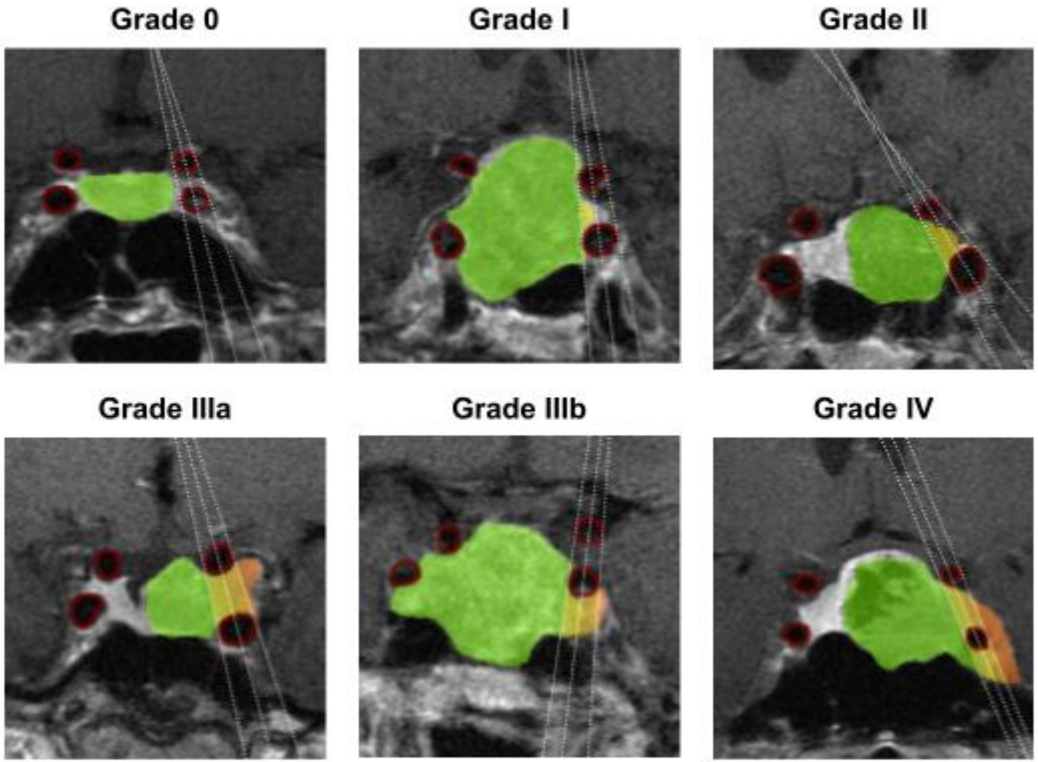
	N	Applicable without modifications	Applicable with minor modifications	Not applicable
Predicted segmentations	146	30 (20.5%)	77 (52.7%)	39 (26.7%)
no previous surgery	40	12 (30%)	19 (47.5%)	9 (22.5%)
previous surgery	106	18 (17%)	58 (54.7%)	30 (28.3%)
no previous SRS	131	28 (21.4%)	73 (55.7%)	30 (22.9$)
previous SRS	15	2 (13.3%)	4 (26.7%)	9 (60%)
functioning	72	8 (11.1%)	33 (45.8%)	31 (43.1%)
non-functioning	74	22 (29.7%)	44 (59.5%)	8 (10.8%)
Control group	146	117 (80.1%)	24 (16.4%)	5 (3.4%)

Table 3 Human expert ratings by model, previous surgery, previous SRS and endocrinological status

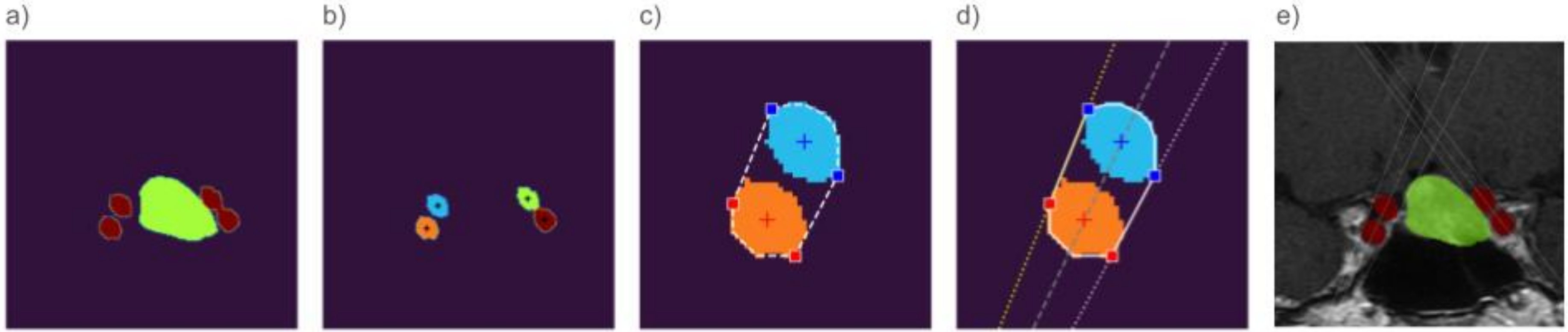
Šetří automatická segmentace čas?

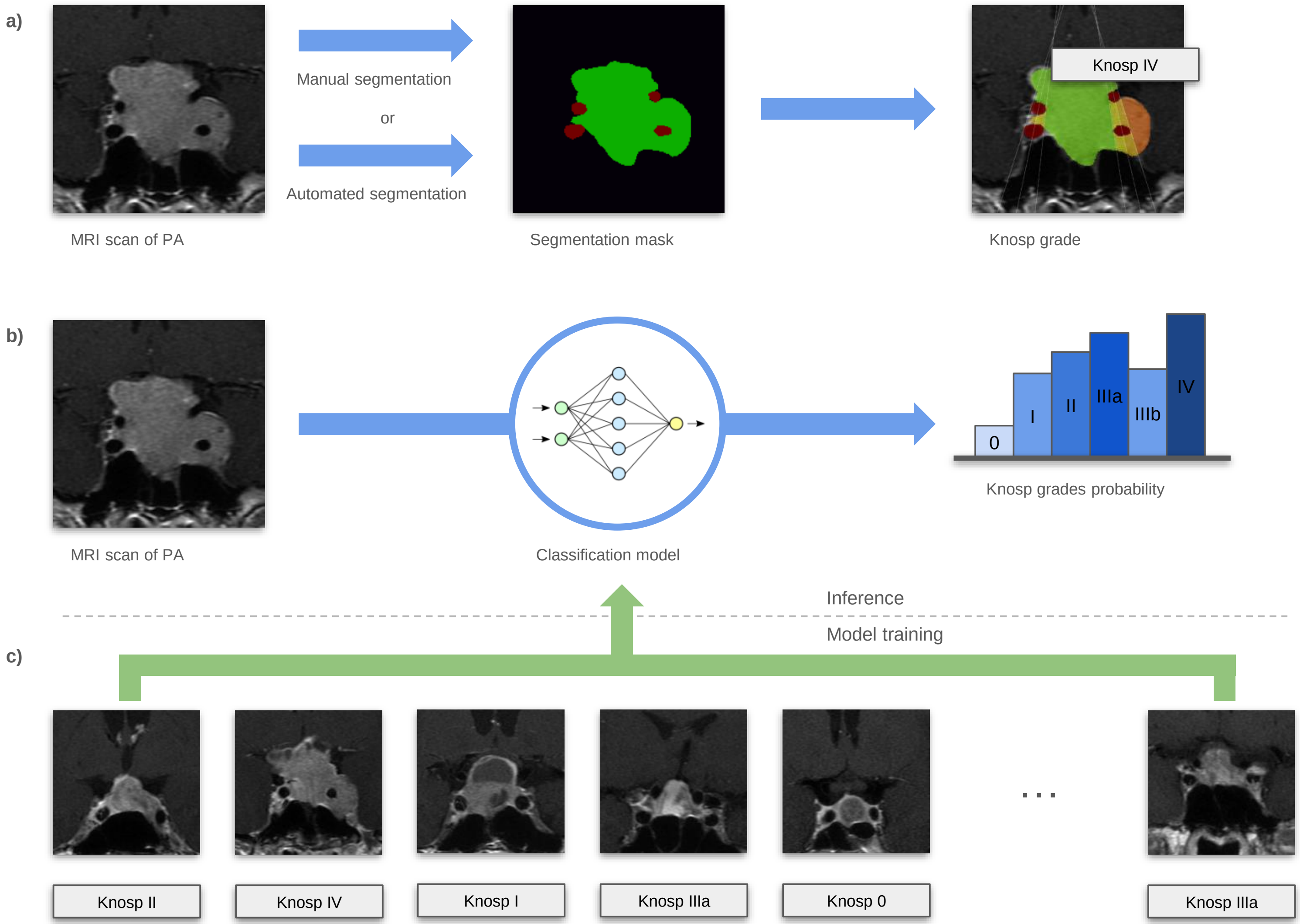
Larger tumor volume, history of a previous radiosurgery and NFPAs were associated with better expert ratings ($p = 0.005$, $p = 0.007$, $p < 0.001$). In the subgroup without previous surgery, although expert ratings were more favorable, the association did not reach statistical significance ($p = 0.074$) (Figure 7). In the subgroup of non-complex cases ($n = 9$), 55.6% of the segmentations were rated as applicable without any manual modifications and no segmentations were rated as non-applicable, the results were non-inferior to the control group ($p = 0.655$). (Figure 6b). Using T2-weighted scans as additional model inputs led to 1.9% improvement in DSC for CN II ($p < 0.001$), there was no significant improvement in DSC for tumor ($p = 0.268$) and normal gland ($p = 0.419$). Manually improving inaccurate segmentations instead of creating them from scratch led to 53.6% reduction of the time burden (3 m 44 s ± 1 m 45 s vs. 1 m 33 s ± 58 s , $p < 0.001$).

Automatické hodnocení Knospova skóre

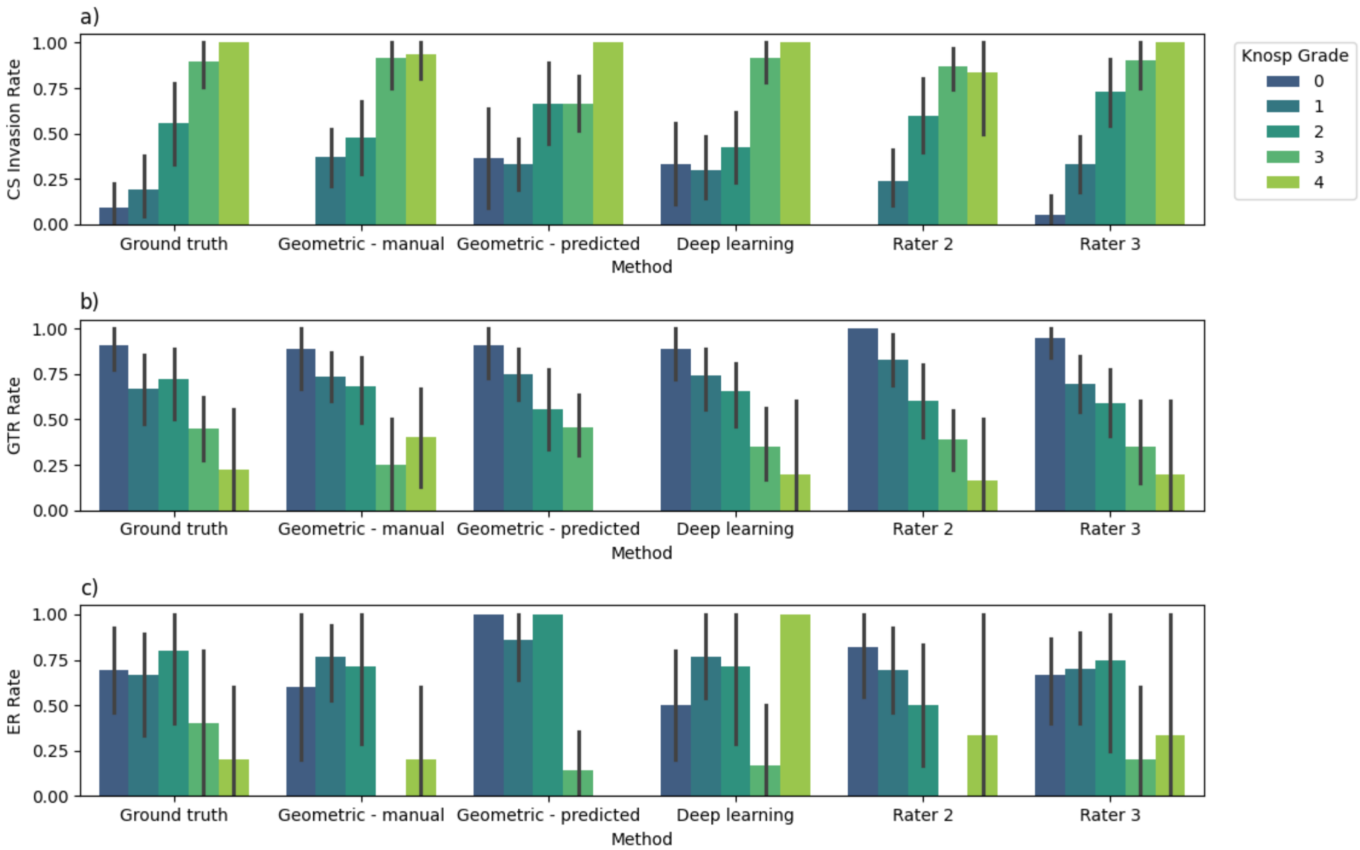


Automatické hodnocení Knospova skóre

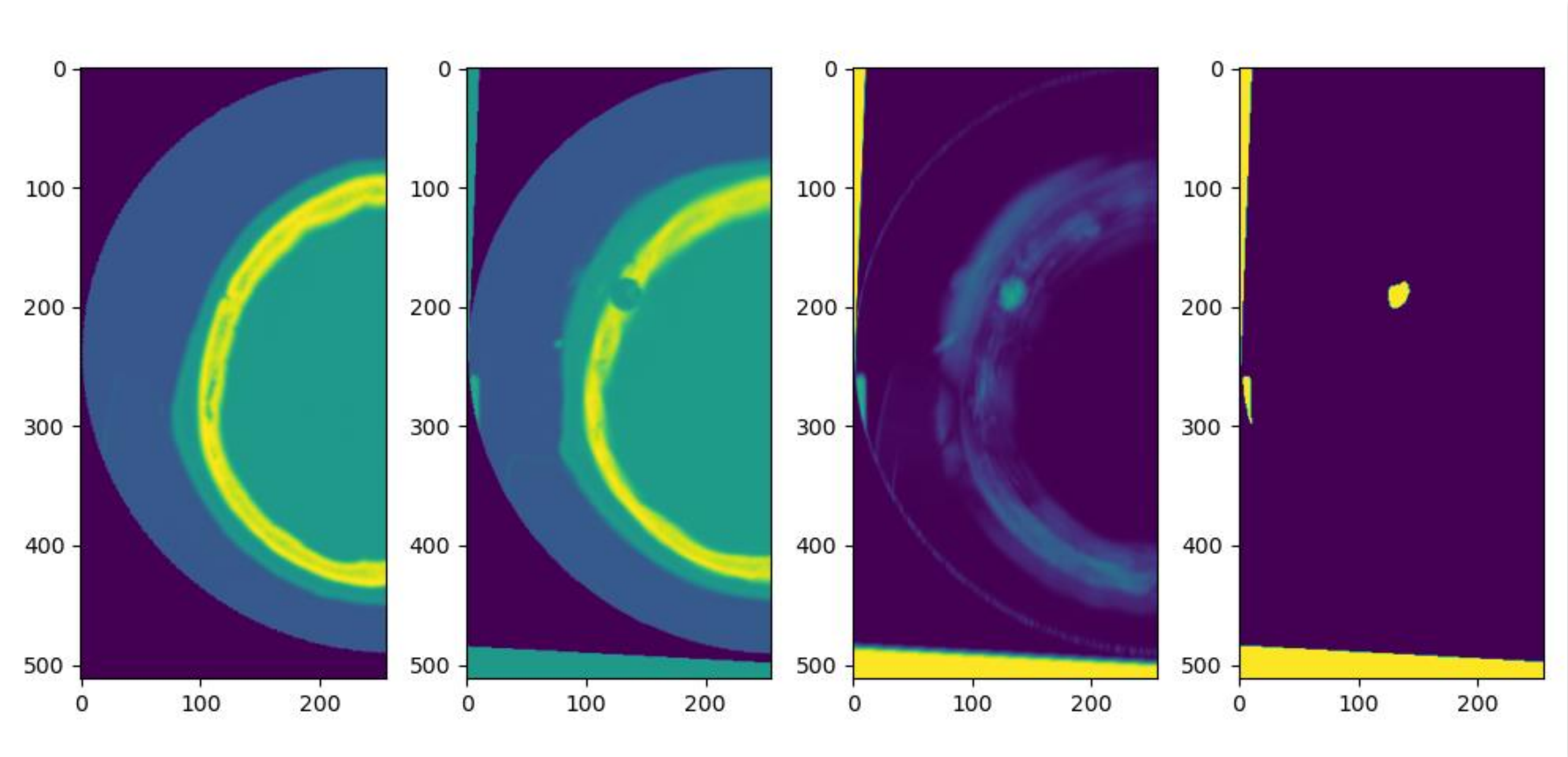


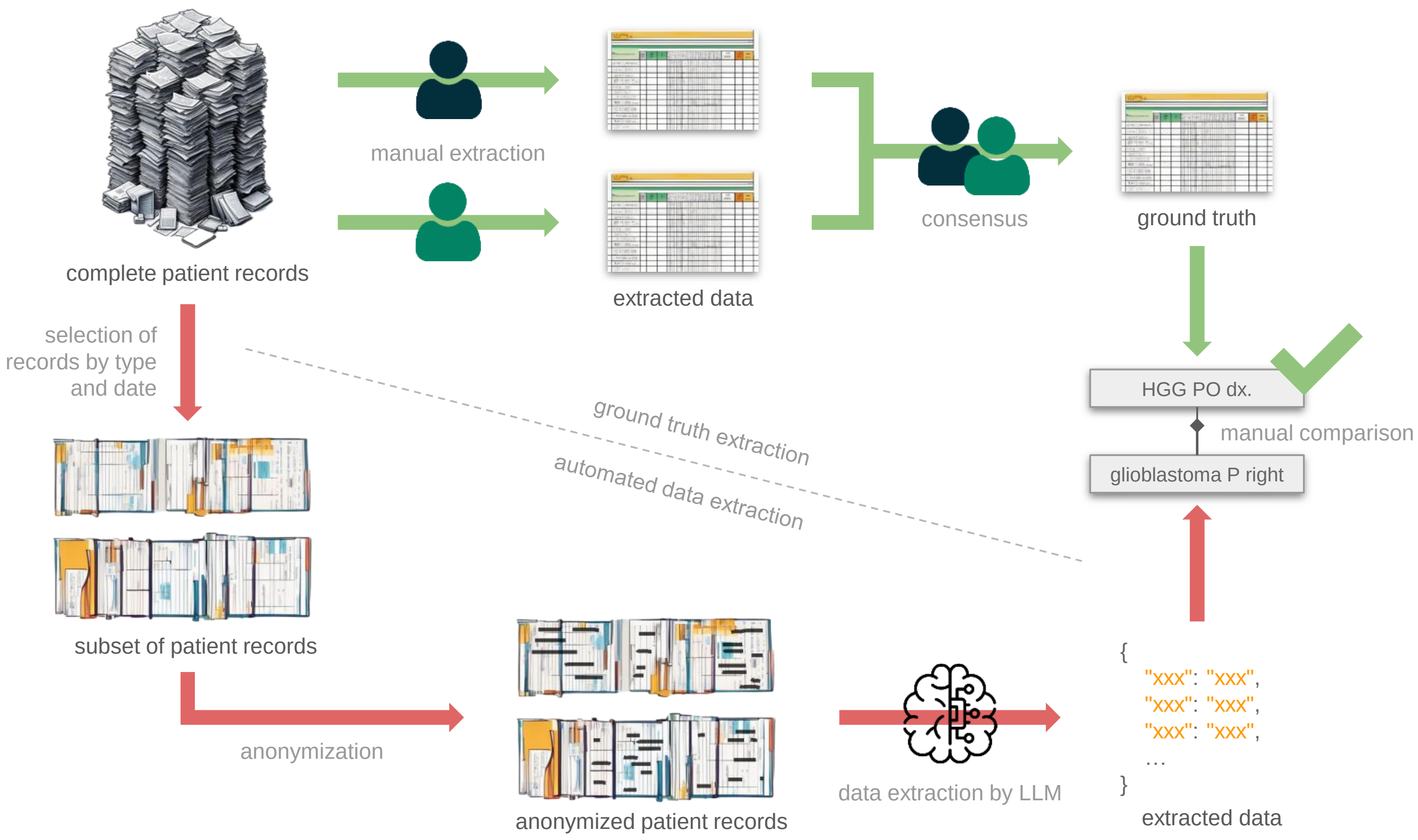

a)
Manual segmentation
or
Automated segmentation
Knosp IV
MRI scan of PA
Segmentation mask
Knosp grade
b)
0
I
II
IIIa
IIIb
IV
Knosp grades probability
MRI scan of PA
Classification model
Inference
Model training
c)
Knosp II
Knosp IV
Knosp I
Knosp IIIa
Knosp 0
Knosp IIIa

Stratifikace rizika - plní automaticky určené Knosp skóre svůj účel?



Auto-anotace obrazových dat




complete patient records
manual extraction
extracted data
consensus
ground truth
selection of records by type and date
ground truth extraction
automated data extraction
HGG PO dx.
manual comparison
glioblastoma P right
subset of patient records
anonymization
anonymized patient records
data extraction by LLM
{
"xxx": "xxx",
"xxx": "xxx",
"xxx": "xxx",
…
}
extracted data

Cca 80% dat v EZZ je nestrukturovaných

Editorial
Healthc Inform Res. 2019 January;25(1):1-2.
https://doi.org/10.4258/hir.2019.25.1.1
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Managing Unstructured Big Data in Healthcare System

Hyoun-Joong Kong
Editorial Taskforce Member of Healthcare Informatics Research, Chungnam National University, Daejeon, Korea

The digital transformation or the 4th industrial revolution which are very recent information technology (IT) agenda make many countries expect the big data to be a source of new economic value that will determine the success and failure of those governments in the future. Due to this trend, the big data industry in the healthcare field has been growing rapidly in recent years and several global IT companies in the United States and Europe are reporting big data use cases in the medical field.

Medical big data refers to large-scale data that is difficult to handle with existing database management systems in a digitalized healthcare environment including medical centers, wearable devices, and social medias. The medical data, which are exploding exponentially, also include large volume of structured and unstructured data as other domains [1].

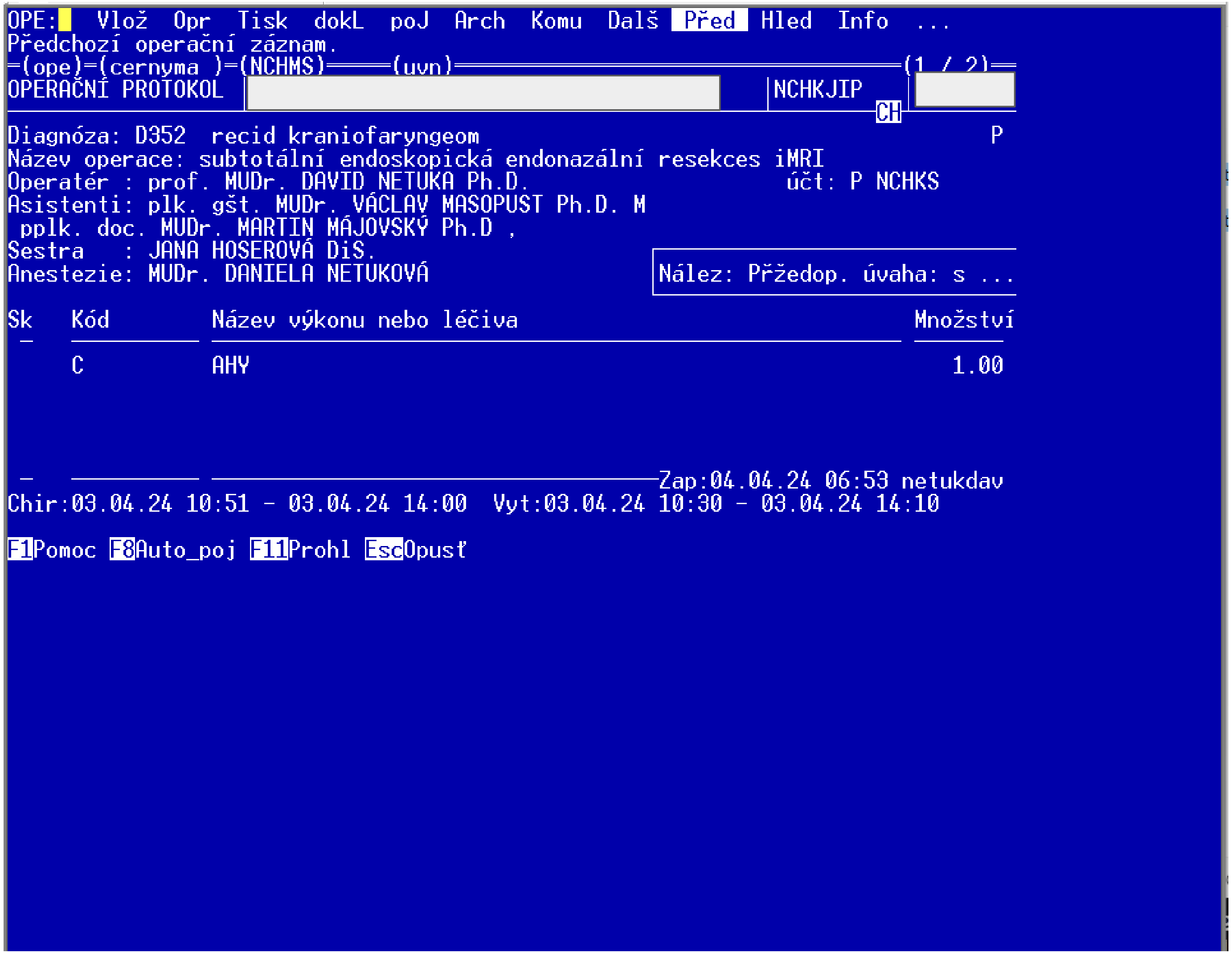
The big problem of healthcare fields is that about 80% of medical data remains unstructured and untapped after it is created (e.g., text, image, signal, etc.) [2]. Since it is hard to handle this type of data for Electronic Medical Record or most hospital information system, it tends to be ignored, unsaved, or abandoned in most medical centers for a long time [3]. Although a lot of data are still created in many hospitals, it is hard to be connected with medical big data research and artificial intelligence industry in healthcare. Therefore, we need to manage those unmanaged unstructured big data medical artificial intelligence which is currently based on machine learning technology.

In many hospitals, time series data are most unmanaged out of many types of unstructured medical data owing to its huge file size despite of the great value of their application. Typical unstructured big data in hospital are as following. The first type of data is medical video data that are recently created explosively from new types of medical imaging devices (e.g., endoscope, laparoscope, surgery robot, capsule endoscope, emergency video camera, thoracoscope, etc.). The second one is biosignal data that have been displayed on screen of patient monitor in operating rooms or intensive care units and wearable health monitoring devices. The third one is audio data that are verbally or nonverbally created from patients pathophysiologically and medical staffs for efficient communication in clinical procedures.

For enhancing the use of these unstructured medical big data, we need to establish the data collection, anonymization, and quality assurance processes. And meta data for each types of unstructured medical data need to be defined, standardized, extracted, and visualized automatically. Then open platform for integration and utilization of the unstructured clinical data should be developed while reflecting these concepts.

Even if machine learning technologies with high accuracy

Příklad

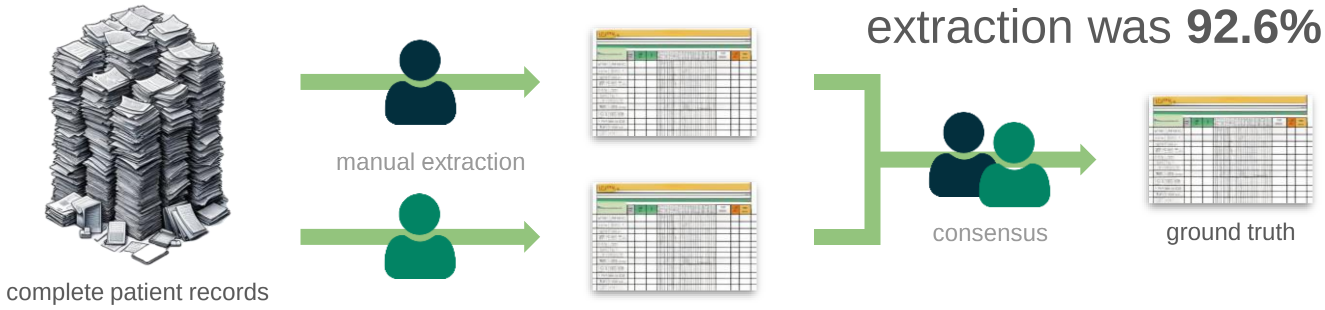

OPE: Vlož Opr Tisk dokL poJ Arch Komu Dalš Před Hled Info ...
Předchozí operační záznam.
=(ope)=(cernyma)=(NCHMS)====(uvn)====(1 / 2)==
OPERAČNÍ PROTOKOL
NCHKJIP
CH
Diagnóza: D352 recid kraniofaryngeom
P
Název operace: subtotální endoskopická endonazální resekces iMRI
Operatér : prof. MUDr. DAVID NETUKA Ph.D.
účt: P NCHKS
Asistenti: plk. gšt. MUDr. VÁCLAV MASOPUST Ph.D. M
pplk. doc. MUDr. MARTIN MÁJOVSKÝ Ph.D ,
Sestra : JANA HOSEROVÁ DiS.
Anestezie: MUDr. DANIELA NETUKOVÁ
Nález: Přžedop. úvaha: s ...
Sk
Kód
Název výkonu nebo léčiva
Množství
C
AHY
1.00
Zap:04.04.24 06:53 netukdav
Chir:03.04.24 10:51 - 03.04.24 14:00 Vyt:03.04.24 10:30 - 03.04.24 14:10
F1Pomoc F8Auto_poj F11Prohl EscOpusť

Manuální extrakce 35 definovaných datových bodů

Hospitalization	Admission diagnosis (text), admission diagnosis ICD code (text), hospitalization purpose (therapeutic/diagnostic), surgery (text), further patient course (discharge/transfer), number of days in ICU (number), revision surgery (text, if underwent a revision).
Neurological state on admission	GCS (number), speech impairment (yes/no), motor impairment (no/UE monoparesis, LE monoparesis, hemiparesis, paraparesis, quadriparesis, cranial nerve paresis), sensory impairment (yes/no), UMN signs (yes/no), pain intensity (number, VAS scale), pain localization (text), visual impairment (yes/no).
Neurological state at the 1st postoperative day	Same as on admission
Neurological state at discharge	Same as on admission
Therapy	Corticotherapy (yes/no), algotherapy (no/peroral/intravenous/parenteral opioids), antibiotics (no/peroral/intravenous, excluding perioperative prophylaxis), antibiotics (number of days, excluding perioperative prophylaxis).

EHR types collected for the time period of the hospitalization	
Clinical summaries	A brief summary of the patient, usually one or two paragraphs, includes relevant patient history, presenting symptoms, performed procedures and imaging exams, at least one at admission, possibly multiple such summaries during the hospitalization
History of present illness	Usually one or two paragraphs, includes the nature and onset of symptoms, their duration, severity, and progression, as well as factors that exacerbate or alleviate them
Neurological assessment on admission	An examination report detailing consciousness, orientation, speech, cranial nerve functions, motor functions, sensory system functions, reflexes, gait, cerebellar functions and meningeal signs
Consultation notes	Reports from interdisciplinary consultations requested during the hospitalization
Progress notes	Daily notes reporting any changes in patient signs and symptoms and performed procedures
Prescription orders	Daily drug prescriptions
Therapeutic orders	Prescriptions for non-pharmacological therapy, includes physiotherapy, ergotherapy, wound management, catheters and dietary measures
Nursing assessment	Includes vital signs, physical and mental status assessment, fluid intakes and outputs, pain assessment, patient education and communication and safety measures
EHR types collected for the time period of two years before the hospitalization	
Neurosurgery and neurology outpatient consult notes	Including the patient's medical history, neurological findings, diagnosis, treatment plan, and follow-up instructions, relevant to their neurological condition
Emergency department visit summaries	Typically including the reason for the visit, initial assessment, interventions provided, results of diagnostic tests, final diagnosis, and any follow-up plans or referrals
Radiology reports	Detailed findings from imaging studies, including information about the technique used, area imaged, findings, and any potential abnormalities or diagnostic clues relevant to the patient's condition

The mean length of the concatenated patient documentation was 906.0 ± 158.7 words rate of agreement in data extraction was **92.6%**



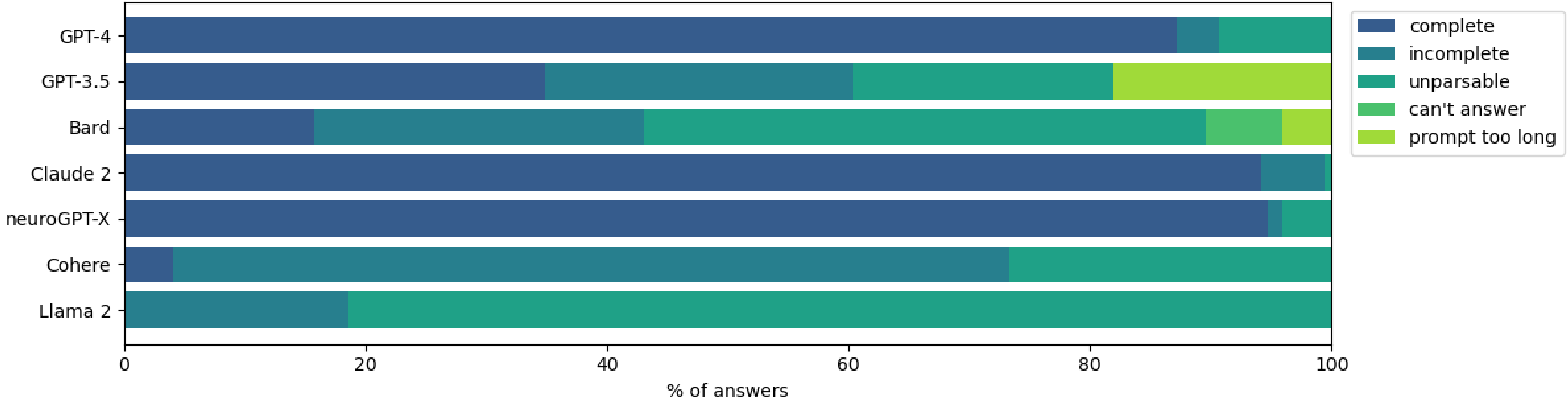
Model prompting

Model	Developer	Published	Maximum input length	# of parameters	Availability	Operation	Domain specificity	Date of prompting
GPT-4 [24]	Open AI	April 2023	8,192 tokens ≈ 4,000 words	175 billion	Subscription based	Cloud-based	General purpose	January 19th 2024
GPT-3.5 [25]	Open AI	November 2021	4,096 tokens ≈ 2,000 words	175 billion	Free access	Cloud-based	General purpose	January 19th 2024
Bard [26, 27]	Google	March 2023	Unpublished	137 billion	Free access	Cloud-based	General purpose	January 19th 2024
Claude 2 [28]	Anthropic	March 2023	100,000 tokens ≈ 60,000 words	200 billion	Limited free access	Cloud-based	General purpose	January 19th 2024
neuroGPT-X [29]	Hotchkiss Brain Institute	October 2023	4,096 tokens ≈ 2,000 words	175 billion	Proprietary	Cloud-based	Vestibular schwannoma	January 20th 2024
LLaMA-2 [30]	Meta	July 2023	2,048 tokens ≈ 1,000 words	70 billion	Open-source	Local	General purpose	February 14th 2024
Cohere [31]	Cohere	Details on the model not published			Proprietary	Cloud-based	General purpose	February 21st 2024

Response formatting

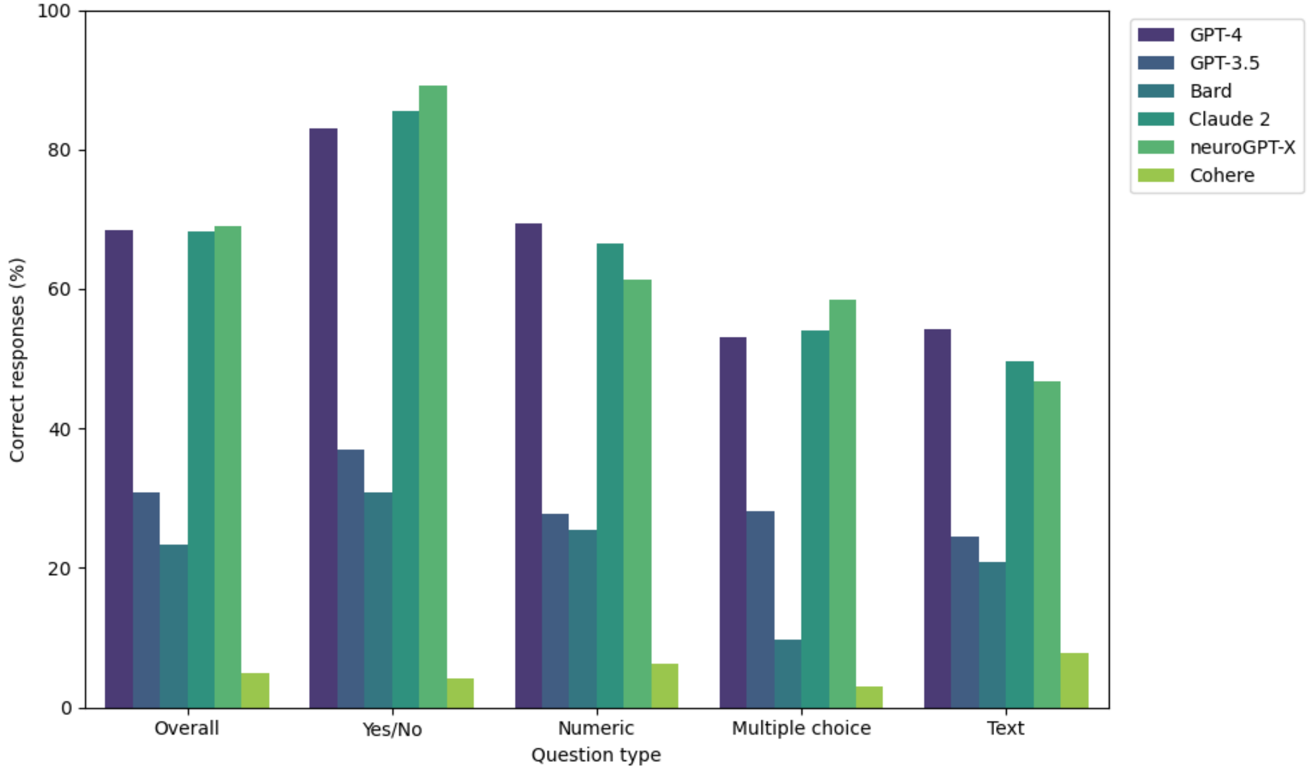

```
{
    "diagnoza_prijeti": "Neurinom L2",
    "mkn_kod": "D33.2",
    "typ_hospitalizace": "terapeuticka",
    "operace": "resekce susp. ependymomu filum terminale",
    "ukonceni_hospitalizace": "dimise",
    "dny_na_ICU": 0,
    "revize": "",
    "GCS_vstupne": 15,
    "porucha_reci_vstupne": false,
    "porucha_hybnosti_vstupne": "ne",
    "porucha_citi_vstupne": false,
    "iritacni_jevy_vstupne": false,
```

Response formatting - results



Model	Complete	Incomplete	Unparsable	Can't answer	Prompt too long
GPT-4	150 (87.2)	6 (3.5)	16 (9.3)	0 (0.0)	0 (0.0)
GPT-3.5	60 (34.9)	44 (25.6)	37 (21.5)	0 (0.0)	31 (18.0)
Bard	27 (15.7)	47 (27.3)	80 (46.5)	11 (6.4)	7 (4.1)
Claude 2	162 (94.2)	9 (5.2)	1 (0.6)	0 (0.0)	0 (0.0)
neuroGPT-X	163 (94.8)	2 (1.2)	7 (4.1)	0 (0.0)	0 (0.0)
LLaMA-2	0 (0.0)	32 (18.6)	140 (81.4)	0 (0.0)	0 (0.0)
Cohere	7 (4.1)	119 (69.2)	46 (26.7)	0 (0.0)	0 (0.0)

Data extraction accuracy



Role of domain-specific models

JNS

CLINICAL ARTICLE

J Neurosurg 140:1041–1053, 2024

neuroGPT-X: toward a clinic-ready large language model

Edward Guo,[1,2] Mehul Gupta, MD,[1] Sarthak Sinha, BSc,[1] Karl Rössler, MD,[3] Marcos Tatagiba, MD,[4] Ryojo Akagami, MD,[5] Ossama Al-Mefty, MD,[6] Taku Sugiyama, MD, PhD,[7] Philip E. Stieg, PhD, MD,[8] Gwynedd E. Pickett, MD,[9] Madeleine de Lotbiniere-Bassett, MD,[1,2] Rahul Singh, BSc,[1,2] Sanju Lama, MD, PhD,[1,2] and Garnette R. Sutherland, MD[1,2]

[1]Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [2]Department of Clinical Neurosciences, Project neuroArm, Hotchkiss Brain Institute University of Calgary, Calgary, Alberta, Canada; [3]Department of Neurosurgery, Medical University of Vienna, Vienna, Austria; [4]Department of Neurosurgery, Tubingen University, Tubingen, Germany; [5]Department of Surgery, University of British Columbia, Vancouver, British Columbia, Canada; [6]Department of Neurosurgery, Harvard School of Medicine, Boston, Massachusetts; [7]Department of Neurosurgery, Hokkaido University Graduate School of Medicine, Sapporo, Japan; [8]Department of Neurosurgery, Weill Cornell Medicine/NewYork-Presbyterian Hospital, New York, New York; and [9]Department of Surgery, Dalhousie University, Halifax, Nova Scotia, Canada

OBJECTIVE The objective was to assess the performance of a context-enriched large language model (LLM) compared with international neurosurgical experts on questions related to the management of vestibular schwannoma. Furthermore, another objective was to develop a chat-based platform incorporating in-text citations, references, and memory to enable accurate, relevant, and reliable information in real time.

METHODS The analysis involved 1) creating a data set through web scraping, 2) developing a chat-based platform called neuroGPT-X, 3) enlisting 8 expert neurosurgeons across international centers to independently create questions (n = 1) and to answer (n = 4) and evaluate responses (n = 3) while blinded, and 4) analyzing the evaluation results on the management of vestibular schwannoma. In the blinded phase, all answers were assessed for accuracy, coherence, relevance, thoroughness, speed, and overall rating. All experts were unblinded and provided their thoughts on the utility and limitations of the tool. In the unblinded phase, all neurosurgeons provided answers to a Likert scale survey and long-answer questions regarding the clinical utility, likelihood of use, and limitations of the tool. The tool was then evaluated on the basis of a set of 103 consensus statements on vestibular schwannoma care from the 8th Quadrennial International Conference on Vestibular Schwannoma.

RESULTS Responses from the naive and context-enriched Generative Pretrained Transformer (GPT) models were consistently rated not significantly different in terms of accuracy, coherence, relevance, thoroughness, and overall performance, and they were often rated significantly higher than expert responses. Both the naive and content-enriched GPT models provided faster responses to the standardized question set than expert neurosurgeon respondents ($p < 0.01$). The context-enriched GPT model agreed with 98 of the 103 (95%) consensus statements. Of interest, all expert surgeons expressed concerns about the reliability of GPT in accurately addressing the nuances and controversies surrounding the management of vestibular schwannoma. Furthermore, the authors developed neuroGPT-X, a chat-based platform designed to provide point-of-care clinical support and mitigate the limitations of human memory. neuroGPT-X incorporates features such as in-text citations and references to enable accurate, relevant, and reliable information in real time.

CONCLUSIONS The present study, with its subspecialist-level performance in generating written responses to complex neurosurgical problems for which evidence-based consensus for management is lacking, suggests that context-enriched LLMs show promise as a point-of-care medical resource. The authors anticipate that this work will be a springboard for expansion into more medical specialties, incorporating evidence-based clinical information and developing expert-level dialogue surrounding LLMs in healthcare.

https://thejns.org/doi/abs/10.3171/2023.7.JNS23573

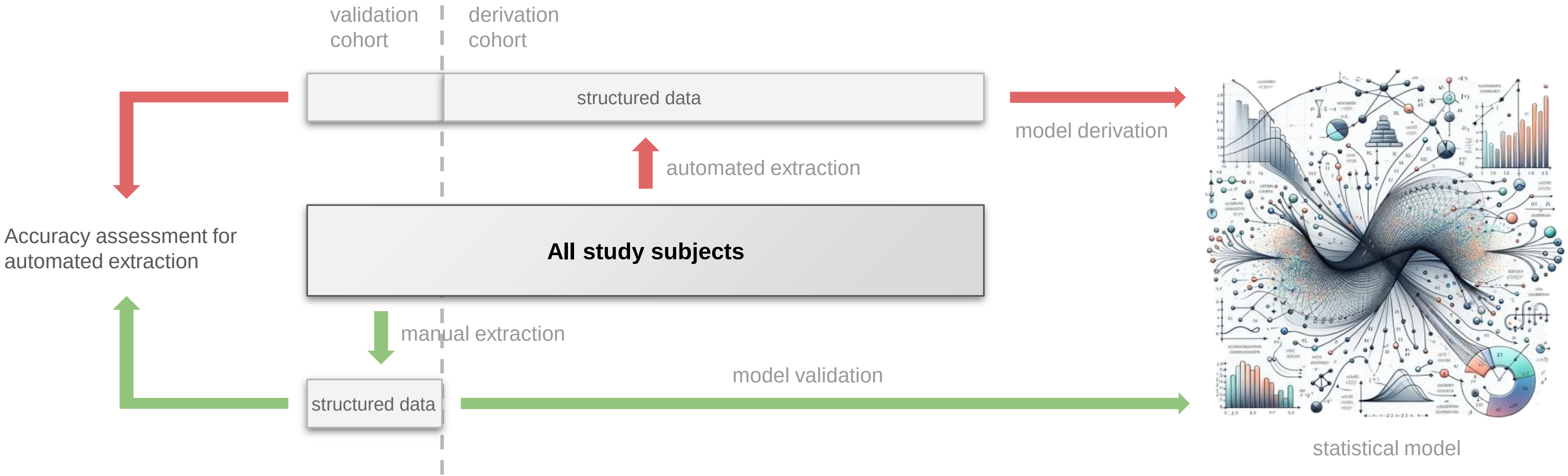
Role of domain-specific models

Fine-tuned models

vs

Context-enriched models

Is data extraction by LLMs good enough?



Děkuji za pozornost a těším se na další setkávání a spolupráce