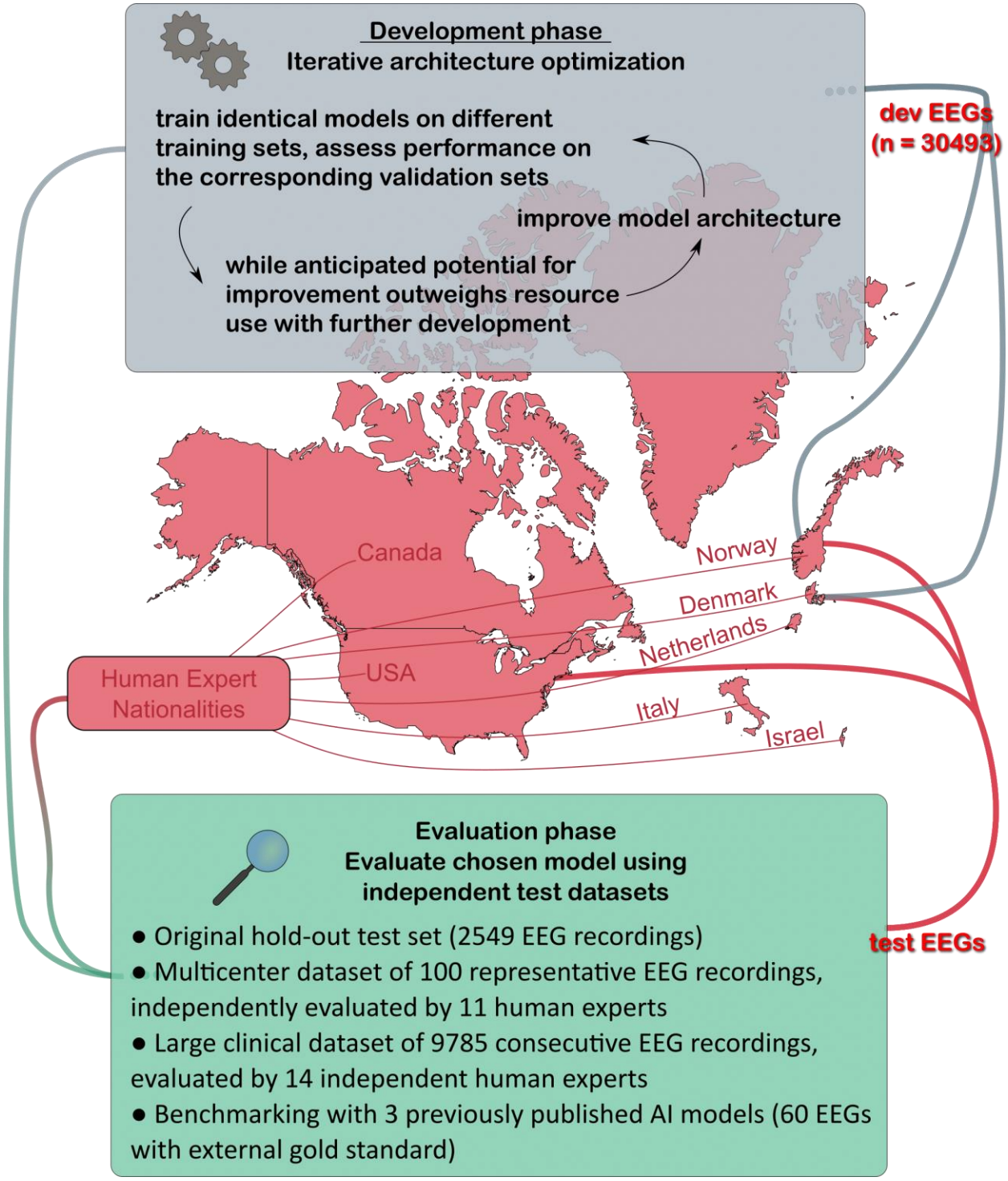
Supplemental Online Content

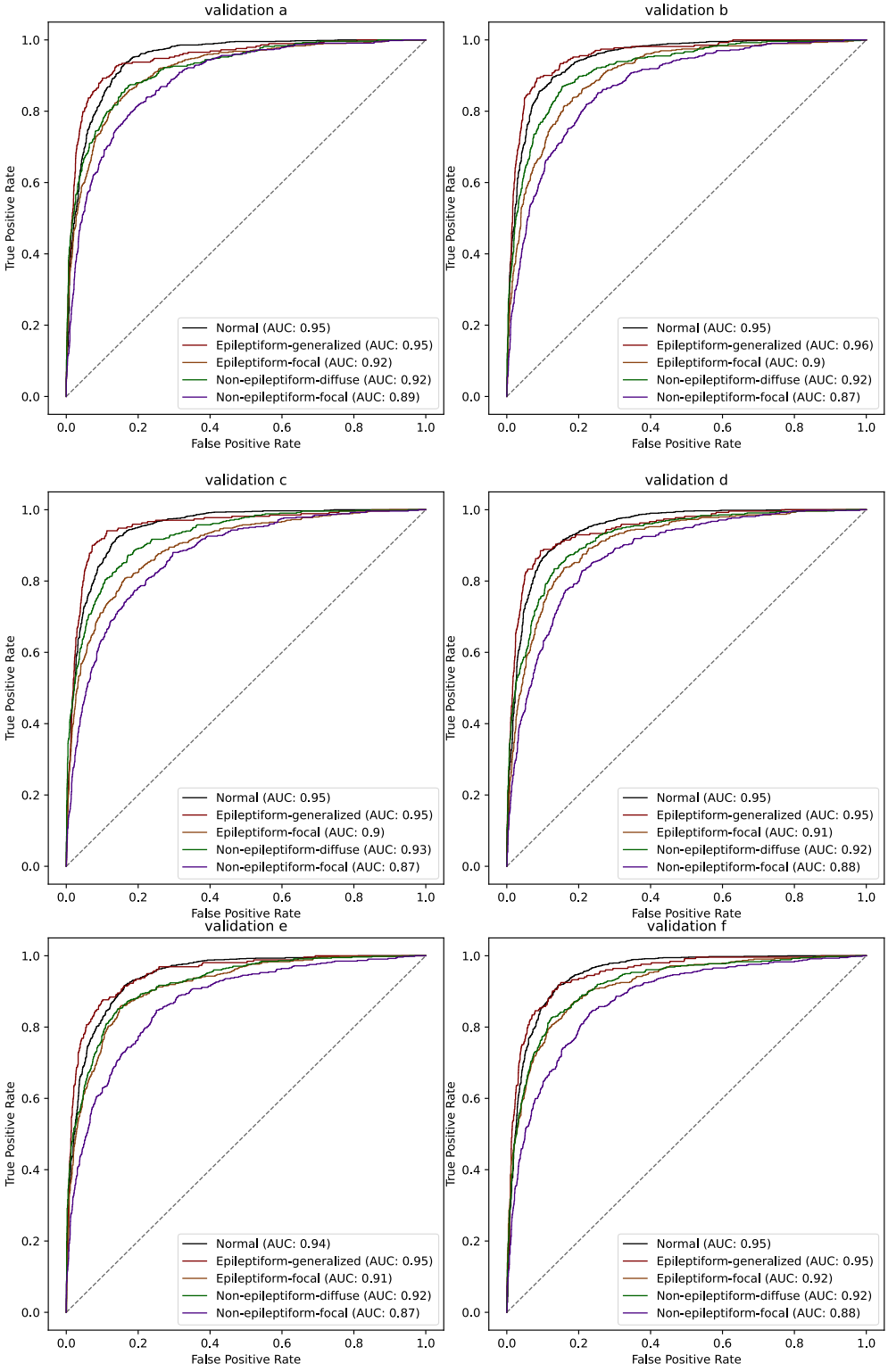
Tveit J, Aurlien H, Plis S, et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA Neurol.* Published online June 20, 2023. doi:10.1001/jamaneurol.2023.1645

This supplemental material has been provided by the authors to give readers additional information about their work.

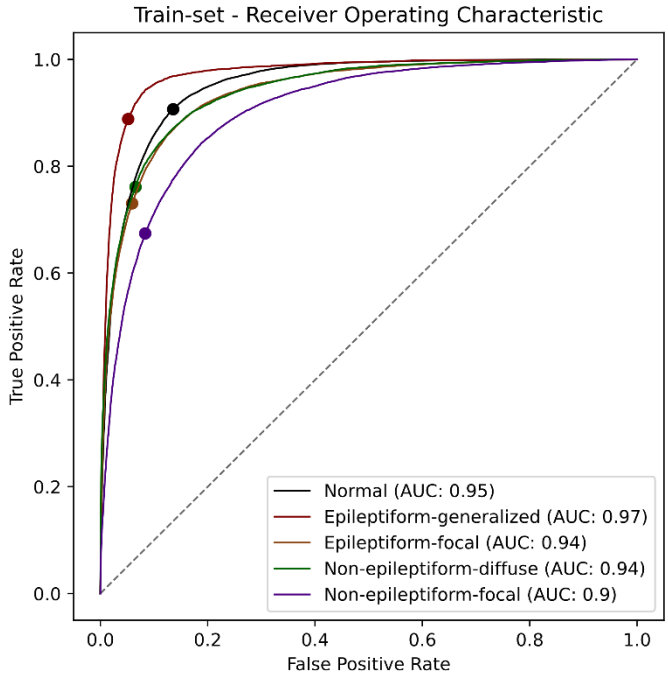
eFigure 1 AI model training and evaluation



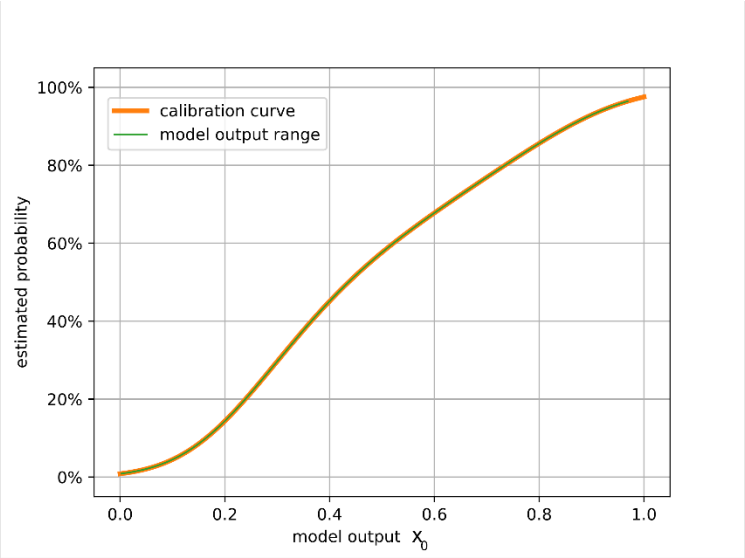
eFigure 2: Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) on the cross-validation datasets for the final model architecture, trained on the corresponding training sets in the cross-validation scheme.



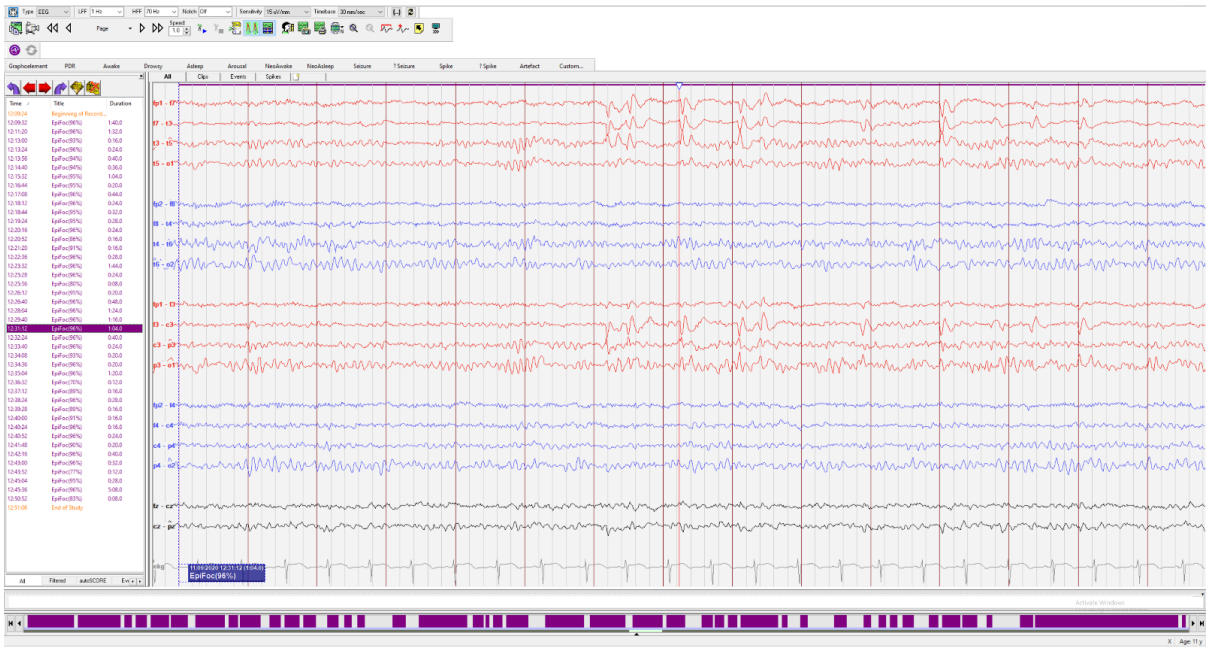
eFigure 3: Receiver Operating Characteristics (ROC) and ROC Area Under the Curve (AUC) for the entire development set of the final model (which was trained on the entire development set). The maximum accuracy thresholds (eTable 2) are indicated by dots placed on the curves.



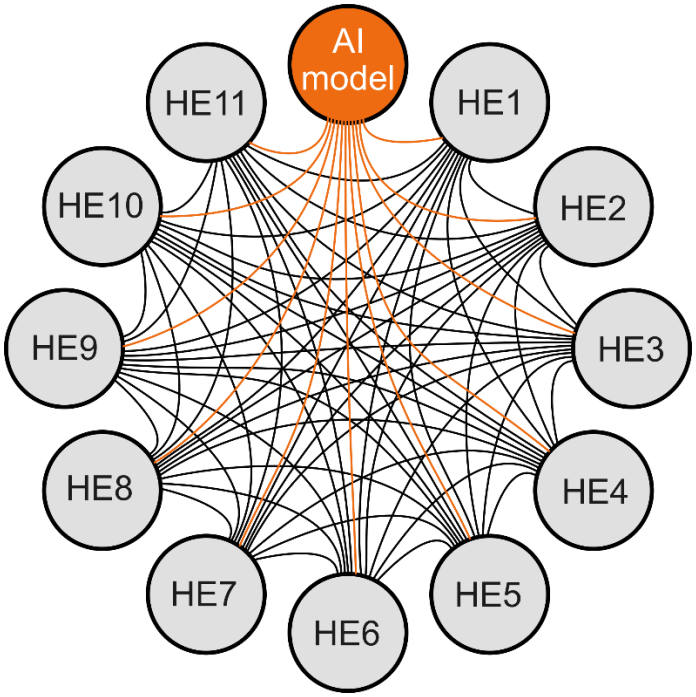
eFigure 4: The calibration curve is an approximate mapping of the model output vs. the estimated probability of the condition. In this plot, the x_0 predictor, which corresponds to the probability of an EEG not containing any abnormalities in a balanced input dataset. The model output range shows the range of all observed model outputs.



eFigure 5. autoSCORE: Integration of SCORE-AI with the Natus NeuroWorks EEG reader. The annotation viewer box to the left lists the epochs of EEG when the model identified abnormal EEG patterns. Selecting an item on the list, navigates the interpreter to the epoch of EEG containing the salient abnormality udentified. The example in the figure shows abnormal focal left centrotemporal epileptiform discharges on EEG that was detected by the model.



eFigure 6: Pairwise comparison strategy. The 11 human experts (HE) give 55 HE-HE pairs and 11 model-HE pairs.

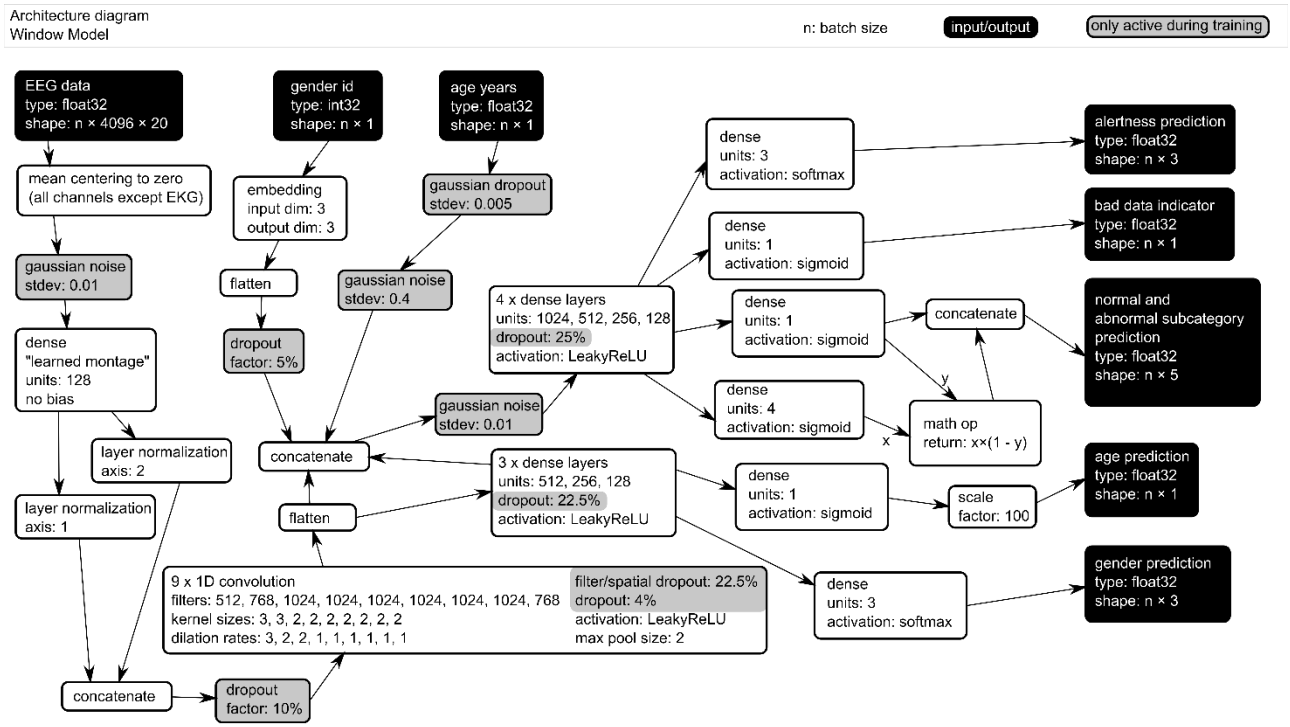


The following comparison strategies were used:

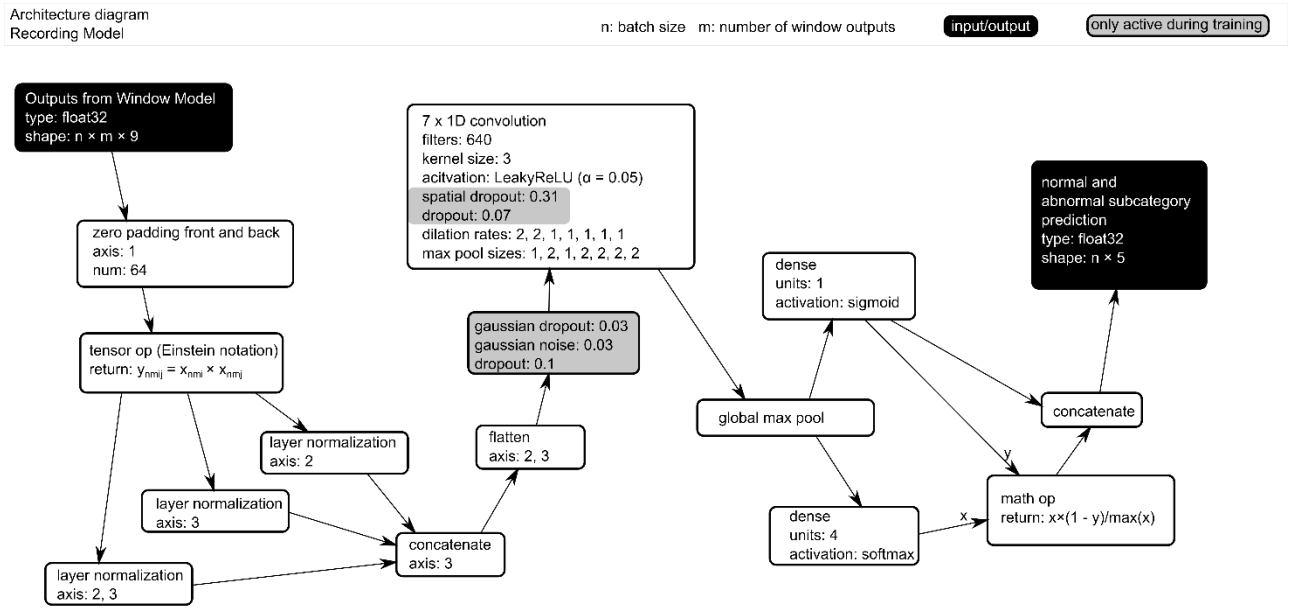
- Average pairwise agreement between pairs of human experts and between the AI model– human-expert pairs.
- Average agreement with human expert majority consensus for normality and each abnormal subcategory individually.

eFigure 7: SCORE-AI model architecture

eFigure 7a: Neural network architecture of the window model.

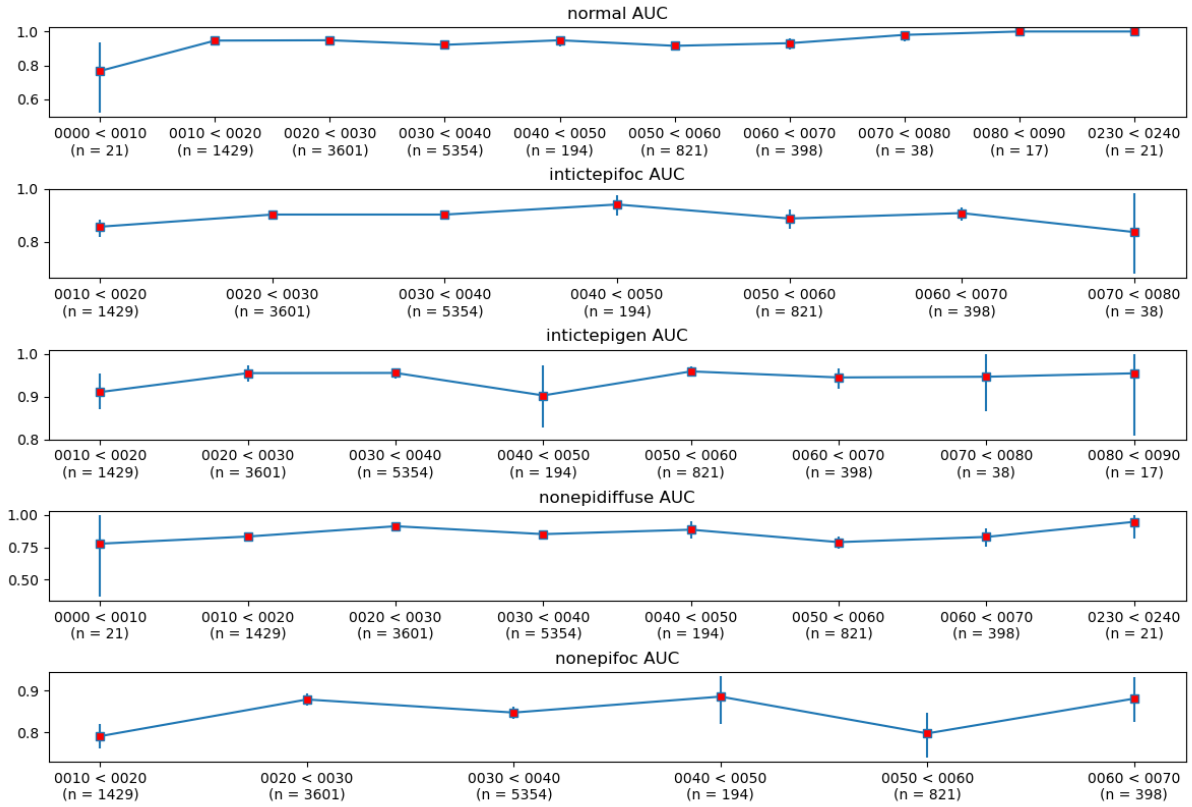


eFigure 7b: Neural network architecture of the recording model.



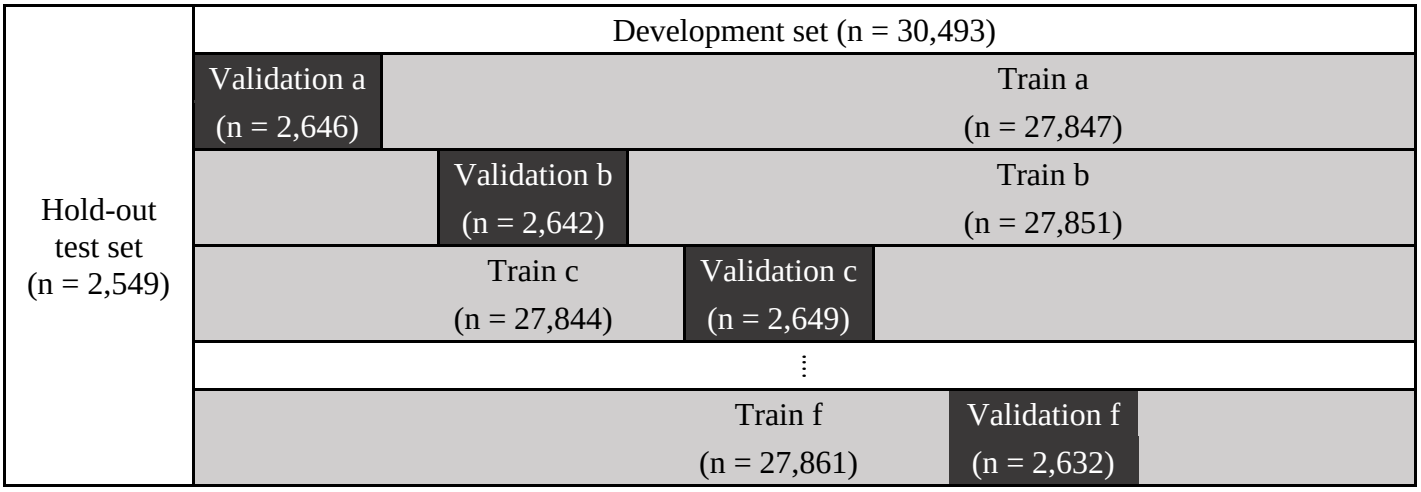
eFigure 8: Area under the ROC curve (AUC) depending on the duration of the EEG recording.

Horizontal scale: 10 minutes EEG recording duration. The blue vertical lines indicate 95% confidence intervals; intictepifoc: focal epileptiform; interictepigen: generalised epileptiform; nonepidiffuse: non-epileptiform diffuse; nonepifoc: non-epileptiform focal abnormality.



eTable 1: A cross validation scheme, used for model development, partitioning the development dataset into training and validation datasets

The partitioning scheme of the data from Haukeland University Hospital and Filadelfia Clinic, where n is the number of EEGs. The hold-out test set and the validation sets contain approximately 50% abnormal and 50% normal EEGs. Different validation sets do not overlap, however, different training and validation sets may overlap. The hold-out test dataset does not overlap with the development dataset and contains different patients. Of the validation dataset, 13,208 EEGs (43.31%) were abnormal. For each validation dataset, we randomly selected 9% of the abnormal EEGs, then added a balanced number of normal EEGs, and kept the rest of the development dataset for training.



eTable 2: Threshold for optimal accuracy based on training dataset (calculated using balanced bootstrap resampling).

subcategory	threshold	SD
Epileptiform-focal	0.3719	0.0128
Epileptiform-generalized	0.2373	0.0098
Non-epileptiform-diffuse	0.3489	0.0314
Non-epileptiform-focal	0.3920	0.0063
Normal	0.6515	0.0106

eTable 3: Performance of the final model on the hold-out test dataset.

	Sensitivity/TPR	Specificity/TNR	Precision/PPV	NPV	F1-score	Accuracy
Normal	88.0% (86.3%, 89.8%)	89.3% (87.6%, 90.9%)	89.3% (87.6%, 91.0%)	88.0% (86.2%, 89.7%)	88.7% (87.4%, 90.0%)	88.7% (87.4%, 89.9%)
Epi-generalised	82.6% (77.8%, 87.0%)	93.4% (92.4%, 94.4%)	59.2% (54.2%, 64.2%)	97.9% (97.3%, 98.5%)	68.9% (64.6%, 73.0%)	92.3% (91.3%, 93.3%)
Epi-focal	66.7% (62.6%, 70.7%)	91.1% (89.9%, 92.4%)	66.5% (62.4%, 70.5%)	91.2% (90.0%, 92.4%)	66.5% (63.3%, 69.8%)	86.0% (84.7%, 87.4%)
Non-epi-diffuse	76.5% (72.8%, 80.1%)	92.1% (90.9%, 93.2%)	71.2% (67.4%, 74.9%)	93.9% (92.8%, 94.9%)	73.7% (70.7%, 76.6%)	88.9% (87.7%, 90.1%)
Non-epi-focal	69.7% (65.7%, 73.5%)	89.4% (88.1%, 90.8%)	63.0% (59.0%, 67.0%)	92.0% (90.7%, 93.1%)	66.2% (62.8%, 69.3%)	85.4% (84.0%, 86.7%)
Abnormal (mean)	73.9% (71.8%, 75.9%)	91.5% (90.9%, 92.1%)	64.9% (62.8%, 67.0%)	93.8% (93.2%, 94.3%)	68.8% (67.1%, 70.5%)	88.2% (87.6%, 88.8%)

eTable 4: Raw figures of epileptiform findings in previously published dataset of 60 EEGs.

	Our model	**Rater 1**	**Rater 2**	**Rater 3**	**HE Consensus**	**Encevis**	**Persyst**	**Spike-Net**	**Rater Average**
#TP	26	28	27	24	28	29	30	20	79
#TN	27	17	22	24	22	5	1	19	63
#FP	3	13	8	6	8	25	29	11	27
#FN	4	2	3	6	2	1	0	10	11
accuracy	88.3%	75.0%	81.7%	80.0%	83.3%	56.7%	51.7%	65.0%	78.9%
TPR	86.7%	93.3%	90.0%	80.0%	93.3%	96.7%	100.0%	66.7%	87.8%
TNR	90.0%	56.7%	73.3%	80.0%	73.3%	16.7%	3.3%	63.3%	70.0%
FPR	10.0%	43.3%	26.7%	20.0%	26.7%	83.3%	96.7%	36.7%	30.0%
FNR	13.3%	6.7%	10.0%	20.0%	6.7%	3.3%	0.0%	33.3%	12.2%

eTable 5: Training and experience of the human experts who rated the multicenter test dataset of EEGs. In total 14 human experts, from 14 different centers, trained in 12 different institutions, rated the EEGs. Three experts rated only adult EEGs and three experts only pediatric EEGs. Hence, all EEGs were rated by 11 experts. The median number of years with experience in EEG reading was 17, range: 2-45 years.

Rater	Board Certification / Fellowship Training				Institution where trained in EEG	Years of experience
	Neurology	Pediatric neurology	Clinical Neurophysiology	Epilepsy		
Vibeke Arntsen	Yes	No	Yes	No	Trondheim University Hospital	5
Fieke Cox	Yes	No	Yes	No	Leiden University Medical Centre	11
Firas Fahoum	Yes	No	Yes	Yes	Tel Aviv Sourasky Medical Center	10
William B. Gallentine**	Yes	Yes	Yes	Yes	Duke University	15
Elena Gardella	Yes	No	No	Yes	Bellaria Hospital, University of Bologna	27
Cecil Hahn**	Yes	Yes	Yes	Yes	The Hospital for Sick Children, University of Toronto	19
Aatif M. Husain*	Yes	No	Yes	Yes	Duke University	25
Sudha Kessler**	Yes	Yes	Yes	Yes	Columbia University Medical Center Neurologic Institute	15
Fábio A. Nascimento	Yes	No	Yes	Yes	Massachusetts General Hospital	2
Donald L. Schomer*	Yes	No	Yes	Yes	Montreal Neurological Institute	45
Hatice Tankisi	Yes	No	Yes	No	Aarhus University	27
William O. Tatum*	Yes	No	Yes	Yes	University of Pennsylvania	31
Line B. Ulvin	Yes	No	Yes	No	Oslo University Hospital	2
Richard Wennberg	Yes	No	Yes	Yes	Montreal Neurological Institute	27

* Rated the EEGs from adult patients only; **Rated the EEGs from pediatric patients only

eTable 6: Demographic distribution of patients in the multicenter test dataset of EEGs. In this table, the resulting consensus of a normal vs abnormal interpretation is shown.

The dataset included 100 EEGs (61 males, age: 0.8 - 95 years; mean: 34.9 years, median: 25.8 years). In the United States, EEGs were recorded with Xltek NeuroWorks, and in Norway and Denmark with NicoletOne EEG equipment (Natus Neuro, USA).

	Pediatric (< 16 years)		Adult (≥ 16 years)		Total
	normal	Abnormal	normal	abnormal	
EEGs from the US	2	6	6	11	**25**
EEGs from Denmark	4	3	6	7	**20**
EEGs from Norway	9	11	16	19	**55**
Total	**15**	**20**	**28**	**37**	**100**

eTable 7: Results on previously published dataset of 60 EEGs
Significant differences are marked as bold. Raw figures available in eTable 4.

	Accuracy	Sensitivity (TPR)	Specificity (TNR)
SCORE-AI	88.3% (79.2%, 94.9%)	86.7% (72.8%, 96.5%)	90.0% (77.4%, 99.2%)
HE Consensus	83.3% (73.0%, 91.4%)	93.3% (82.5%, 99.8%)	73.3% (56.5%, 87.8%)
Encevis	56.7% (43.9%, 68.6%)	96.7% (88.2%, 99.9%)	16.7% (4.7%, 31.6%)
Persyst	51.7% (39.1%, 64.2%)	100.0% (97.6%, 100.0%)	3.3% (0.1%, 11.9%)
SpikeNet	65.0% (52.9%, 76.3%)	66.7% (49.3%, 82.5%)	63.3% (45.7%, 79.8%)
	Difference from SCORE-AI (significant difference in bold)		
	Accuracy	Sensitivity (TPR)	Specificity (TNR)
HE Consensus - SCORE-AI	-5.0% (-15.7%, 5.6%) p = .18155	6.7% (-6.0%, 20.0%) p = .15767	**-16.7%** **(-33.0%, -0.4%)** **p = .021**
Encevis - SCORE-AI	**-31.7%** **(-45.1%, -17.2%)** **p < .001**	**10.0%** **(0.1%, 21.8%)** **p = .02325**	**-73.3%** **(-87.1%, -55.6%)** **p < .001**
Persyst - SCORE-AI	**-36.7%** **(-50.9%, -20.8%)** **p < .001**	**13.3%** **(2.4%, 26.4%)** **p = .00791**	**-86.7%** **(-95.9%, -71.8%)** **p < .001**
SpikeNet - SCORE-AI	**-23.3%** **(-36.2%, -9.8%)** **p < .001**	**-20.0%** **(-36.9%, -3.1%)** **p = .0106**	**-26.7%** **(-46.5%, -5.8%)** **p = .00658**

HE: human expert

eAppendix 1

Python 3.8.5
Installed packages in the dev environment

Package	Version
-rapt	1.11.2
absl-py	0.11.0
alabaster	0.7.12
altgraph	0.17
anaconda-client	1.7.2
anaconda-navigator	1.10.0
anaconda-project	0.8.3
argh	0.26.2
argon2-cffi	20.1.0
asn1crypto	1.4.0
astroid	2.4.2
astropy	4.0.2
astunparse	1.6.3
async-generator	1.10
atomicwrites	1.4.0
attrs	20.3.0
autopep8	1.5.4
azure-common	1.1.26
azure-core	1.11.0
azure-mgmt-core	1.2.2
azure-mgmt-storage	17.0.0
azure-storage-blob	12.8.0
Babel	2.8.1
backcall	0.2.0
backports.functools-lru-cache	1.6.1
backports.shutil-get-terminal-size	1.0.0
backports.tempfile	1.0
backports.weakref	1.0.post1
bcrypt	3.2.0
beautifulsoup4	4.9.3
bitarray	1.6.1
bkcharts	0.2
bleach	3.2.1
bokeh	2.2.3
boto	2.49.0
Bottleneck	1.3.2
brotlipy	0.7.0
cachetools	4.2.1
certifi	2020.6.20

cffi	1.14.3
chardet	3.0.4
click	7.1.2
cloudpickle	1.6.0
clyent	1.2.2
colorama	0.4.4
comtypes	1.1.7
conda-package-handling	1.7.2
conda-verify	3.4.2
contextlib2	0.6.0.post1
cryptography	3.1.1
cycler	0.10.0
Cython	0.29.21
cytoolz	0.11.0
dask	2.30.0
decorator	4.4.2
defusedxml	0.6.0
diff-match-patch	20200713
distributed	2.30.1
docutils	0.16
entrypoints	0.3
et-xmlfile	1.0.1
fastcache	1.1.0
filelock	3.0.12
flake8	3.8.4
Flask	1.1.2
flatbuffers	1.12
fsspec	0.8.3
future	0.18.2
gast	0.4.0
gevent	20.9.0
glob2	0.7
google-auth	1.25.0
google-auth-oauthlib	0.4.2
google-pasta	0.2.0
greenlet	0.4.17
grpcio	1.34.1
h5py	3.1.0
HeapDict	1.0.1
html5lib	1.1
idna	2.10
imageio	2.9.0
imagesize	1.2.0
importlib-metadata	2.0.0
iniconfig	1.1.1
intervaltree	3.1.0

ipykernel	5.3.4
ipython	7.19.0
ipython-genutils	0.2.0
ipywidgets	7.5.1
isodate	0.6.0
isort	5.6.4
itsdangerous	1.1.0
jdcal	1.4.1
jedi	0.17.1
Jinja2	2.11.2
joblib	0.17.0
json5	0.9.5
jsonschema	3.2.0
jupyter	1.0.0
jupyter-client	6.1.7
jupyter-console	6.2.0
jupyter-core	4.6.3
jupyterlab	2.2.6
jupyterlab-pygments	0.1.2
jupyterlab-server	1.2.0
keras-nightly	2.5.0.dev2021032900
Keras-Preprocessing	1.1.2
keyring	21.4.0
kiwisolver	1.3.0
lazy-object-proxy	1.4.3
libarchive-c	2.9
llvmlite	0.34.0
locket	0.2.0
lxml	4.6.1
Markdown	3.3.3
MarkupSafe	1.1.1
matplotlib	3.3.2
mccabe	0.6.1
menuinst	1.4.16
mistune	0.8.4
mkl-fft	1.2.0
mkl-random	1.1.1
mkl-service	2.3.0
mock	4.0.2
more-itertools	8.6.0
mpmath	1.1.0
msgpack	1.0.0
msrest	0.6.21
multipledispatch	0.6.0
navigator-updater	0.2.1
nbclient	0.5.1

nbconvert	6.0.7
nbformat	5.0.8
nest-asyncio	1.4.2
networkx	2.5
nltk	3.5
nose	1.3.7
notebook	6.1.4
numba	0.51.2
numexpr	2.7.1
numpy	1.19.5
numpydoc	1.1.0
oauthlib	3.1.0
olefile	0.46
openpyxl	3.0.5
opt-einsum	3.3.0
packaging	20.4
pandas	1.1.3
pandocfilters	1.4.3
paramiko	2.7.2
parso	0.7.0
partd	1.1.0
path	15.0.0
pathlib2	2.3.5
pathtools	0.1.2
patsy	0.5.1
pefile	2021.5.24
pep8	1.7.1
pexpect	4.8.0
pickleshare	0.7.5
Pillow	8.0.1
pip	20.2.4
pkginfo	1.6.1
pluggy	0.13.1
ply	3.11
prometheus-client	0.8.0
prompt-toolkit	3.0.8
protobuf	3.14.0
psutil	5.7.2
py	1.9.0
pyasn1	0.4.8
pyasn1-modules	0.2.8
pycodestyle	2.6.0
pycosat	0.6.3
pycparser	2.20
pycurl	7.43.0.6
pydocstyle	5.1.1

pyEDFlib	0.1.20
pyflakes	2.2.0
Pygments	2.7.2
pyinstaller	4.3
pyinstaller-hooks-contrib	2021.1
pylint	2.6.0
PyNaCl	1.4.0
pyodbc	4.0.0-unsupported
pyOpenSSL	19.1.0
pyparsing	2.4.7
pyreadline	2.1
pyrsistent	0.17.3
PySocks	1.7.1
pytest	0.0.0
python-dateutil	2.8.1
python-jsonrpc-server	0.4.0
python-language-server	0.35.1
pytz	2020.1
PyWavelets	1.1.1
pywin32	227
pywin32-ctypes	0.2.0
pywinpty	0.5.7
PyYAML	5.3.1
pyzmq	19.0.2
QDarkStyle	2.8.1
QtAwesome	1.0.1
qtconsole	4.7.7
QtPy	1.9.0
regex	2020.10.15
requests	2.24.0
requests-oauthlib	1.3.0
rope	0.18.0
rsa	4.7
Rtree	0.9.4
ruamel-yaml	0.15.87
scikit-image	0.17.2
scikit-learn	0.23.2
scipy	1.5.2
seaborn	0.11.0
Send2Trash	1.5.0
setuptools	50.3.1.post20201107
simplegeneric	0.8.1
singledispatch	3.4.0.3
sip	4.19.13
six	1.15.0
snowballstemmer	2.0.0

sortedcollections	1.2.1
sortedcontainers	2.2.2
soupsieve	2.0.1
Sphinx	3.2.1
sphinxcontrib-applehelp	1.0.2
sphinxcontrib-devhelp	1.0.2
sphinxcontrib-htmlhelp	1.0.3
sphinxcontrib-jsmath	1.0.1
sphinxcontrib-qthelp	1.0.3
sphinxcontrib-serializinghtml	1.1.4
sphinxcontrib-websupport	1.2.4
spyder	4.1.5
spyder-kernels	1.9.4
SQLAlchemy	1.3.20
statsmodels	0.12.0
svgwrite	1.4.1
sympy	1.6.2
tables	3.6.1
tblib	1.7.0
tensorboard	2.5.0
tensorboard-data-server	0.6.1
tensorboard-plugin-wit	1.8.0
tensorflow	2.5.0
tensorflow-estimator	2.5.0
termcolor	1.1.0
terminado	0.9.1
testpath	0.4.4
threadpoolctl	2.1.0
tifffile	2020.10.1
toml	0.10.1
toolz	0.11.1
tornado	6.0.4
tqdm	4.50.2
traitlets	5.0.5
typing-extensions	3.7.4.3
ujson	4.0.1
unicodecsv	0.14.1
urllib3	1.25.11
watchdog	0.10.3
wcwidth	0.2.5
webencodings	0.5.1
Werkzeug	1.0.1
wheel	0.35.1
widgetsnbextension	3.5.1
win-inet-pton	1.1.0
win-unicode-console	0.5

wincertstore	0.2
wrapt	1.12.1
xlrd	1.2.0
XlsxWriter	1.3.7
xlwings	0.20.8
xlwt	1.3.0
xmltodict	0.12.0
yapf	0.30.0
zict	2.0.0
zipp	3.4.0
zope.event	4.5.0
zope.interface	5.1.2

eAppendix 2

Study protocol.

Evaluation of autoSCORE: an artificial intelligence based algorithm for EEG classification versus human experts

Document Name	Study Protocol
Public Title	Evaluation of autoSCORE: an artificial intelligence-based algorithm for EEG classification versus human experts
Scientific Title	Accuracy of EEG classification by autoSCORE algorithm compared with human experts
Acronym	autoSCORE
Document Number	24084-02
Version	4.0
Release Date	28 Feb 2022

Abbreviations

Abbreviation	Explanation
EEG	Electroencephalography
HE	Human expert
HUS	Haukeland University Hospital, Norway
OUS	Oslo University Hospital, Norway
FEH	Filadelfia Epilepsy Hospital, Denmark
SCORE	Standardized Computer-based Organized Reporting of EEG
VM	Virtual machine

STUDY DESCRIPTION

Source of Monetary and Material Support
The study is funded by Holberg-EEG AS (Fjøsangerveien 70 A, 5068 Bergen, Norway)
Web: holbergeeg.com
Phone: +47 926 44 261

Principal Investigator
Professor Sándor Beniczky (Danish Epilepsy Centre and Aarhus University Hospital, Denmark) takes responsibility for initiating and managing the study.

Contact for Public and Scientific Queries
Professor Sándor Beniczky, MD, PhD,
Danish Epilepsy Centre and Aarhus University
Address: Epilepsihospitalet Filadelfia, Visby Allé 5, 4293 Dianalund, Denmark
Phone: +4526981536
Email: sbz@filadelfia.dk

Countries of Recruitment
Denmark, Norway, USA.

Problems Studied
Electroencephalography (EEG) in patients suspected for epilepsy, seizures, impaired consciousness or altered cognition.

INTERVENTIONS

Background

Electroencephalography (EEG) measures electric brain activity using electrodes attached to the scalp. This is used in clinical practice to investigate brain disease, most commonly epilepsy, coma, and dementia. The clinical interpretation of EEGs is until now mainly based on expert visual analysis (Tatum IV et al. 2016), and there are indications that EEG reviewers are under increasing time pressures (Ng and Gillis 2017; Brogger et al. 2018). The interrater agreement assessing EEG studies is only moderate (Van Donselaar et al., 1992; Stroink et al., 2006). Holberg EEG has initiated the development of an EEG decision support tool based on deep learning techniques with the purpose of assisting the process of EEG interpretation and increase the interrater agreement. Hospital partners at Haukeland University Hospital (Norway), Filadelfia Epilepsy Hospital (Denmark), and Oslo University Hospital have for many years used SCORE-EEG software developed by Holberg EEG to assess and tag EEG in a standardized way, and at the same time produce a large database of tagged EEGs. This database is used to train an algorithm (autoSCORE) to automatically assess EEGs. autoSCORE will be trained to separate normal from abnormal EEGs. When autoSCORE assesses the EEG as abnormal it will further sub-classify abnormalities into one or more of the subgroups focal epileptiform

abnormality, generalized epileptiform abnormality, focal non-epileptiform abnormality, and diffuse non-epileptiform abnormality.

Objective

To evaluate the accuracy of autoSCORE in distinguishing between normal and abnormal EEG recordings, and classifying the abnormal EEG recordings into the four major clinical categories: focal-epileptiform, generalized-epileptiform, diffuse-slowing (non-epileptiform), focal-slowing (non-epileptiform).
In this diagnostic accuracy study, index-test is autoSCORE, and reference standard is assessment of the routine EEG recordings by HEs. In the phase-3 part of the study, reference standard is the majority consensus of a panel of 11 HEs. In the phase-4 part of the study, reference standard is the clinical assessment of the EEGs, as part of the routine, by HEs at a centre which did not participate in the development of autoSCORE.

Methods

- **Inclusion and exclusion criteria**:
    - *Inclusion*: Routine clinical EEG recordings in patients referred to EEG on suspicion of epilepsy or seizures, and patients referred to EEG on for diagnostic work-up in patients with impaired consciousness or cognitive impairment.
    - *Exclusion*: Patients younger than 3 months, and critically ill patients with rhythmic or periodic EEG patterns.
- **Index test**: AutoSCORE analysis of the EEG recordings. The analysis is fully automated and blinded to all other data. The algorithm and the detection threshold values are fixed (pre-defined according to the previous development process). No iterations are allowed.
- **EEGs in phase-3**: The EEGs to be included into this study have not been part of the training dataset to develop the autoSCORE algorithm. The routine clinical EEGs are recorded at HUS, FEH and at Mayo Clinics. The distribution in this representative validation dataset should be as follows:

	Pediatric <16 years	**Adult >16 years**	**Row sum**
Normal EEG	15	28	43
Abnormal EEG	20	37	57
Column sum	35	65	100

With the above described distribution, 75 EEGs will be randomly selected from the independent test-datasets of 3.000 EEGs from HUS and FEH and 25 EEGs from the independent test-dataset of 140 EEGs from the Mayo Clinics. All EEGs will be anonymized by the Hospitals before they are transmitted to Holberg.

Security of assessments will be assured by restricting access of the HE to their own Excel sheet for storing their assessments, which they can edit with the data in the predetermined columns. Once complete, the HE will sign the Excel sheet and send to Holberg EEG for placement on the SharePoint site. Holberg EEG is blinded to the HE assessments, until the autoSCORE results are documented for all EEGs.

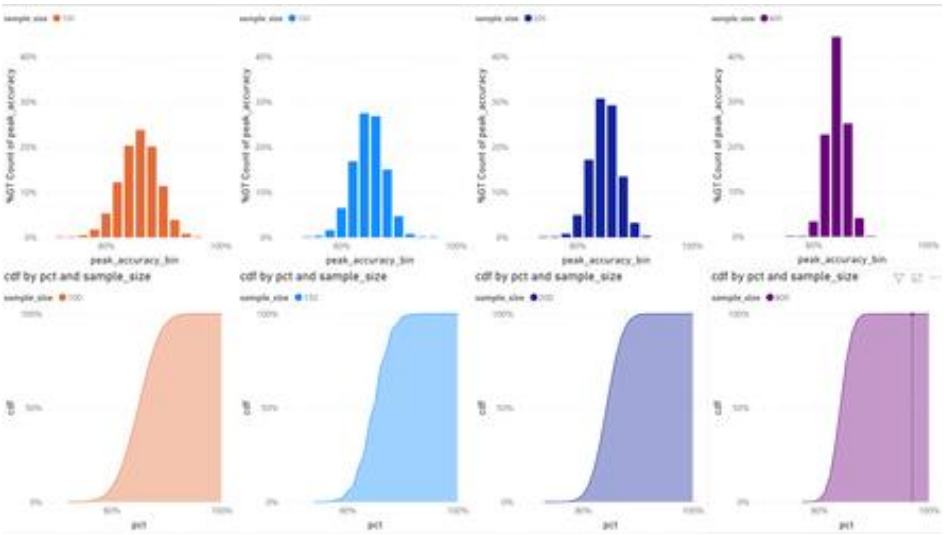
- **Reference standard in phase-3**: Majority consensus of a panel of HEs, who assess independently 100 routine clinical EEGs. Each EEG is assessed by 11 HEs, who will make the following decisions:
    - EEG is normal or abnormal
    - If the EEG is abnormal, HEs assess if one or more of the following categories of abnormality is present:
        - focal-epileptiform abnormality
        - generalized epileptiform abnormality
        - focal-slowing (non-epileptiform) abnormality
        - diffuse-slowing (non-epileptiform) abnormality

  HEs are blinded to autoSCORE.
- **EEGs in phase-4**: 9,398 consecutive EEG recordings from OUH, fulfilling inclusion and exclusion criteria. These recordings have not been used to train the algorithm, and this centre did not participate in development of the algorithm.
- **Reference standard in phase-4**: clinical EEG assessment of the recordings, by HEs evaluating these EEGs as part of the patients´ routing diagnostic workup. The HE assessment is blinded to autoSCORE. Fourteen HEs contributed to the clinical EEG assessment of the EEGs included into phase-4.
- **Benchmarking**: Currently there isn´t any commercially available or published algorithm which provides a comprehensive, fully automated assessment of routine, clinical EEG recordings, comparable with autoSCORE. However, the ENCEVIS software (FDA approved) has a functionality for automated detection of epileptiform discharges. This corresponds to a combination of two of the four categories in the classification of EEG abnormalities (focal-epileptiform and generalized-epileptiform). We will compare the accuracy of autoSCORE and ENCEVIS to identify these combined classes.
- **Outcome measures**:
    - *Primary outcome measures*: diagnostic accuracy parameters, according to the STARD criteria. We will calculate: sensitivity, specificity, accuracy, positive predictive value, negative predictive value and F1-score, for the EEGs in the phase-3 part of the study.
    - *Secondary outcome measures*: Inter-test agreement (autoSCORE vs. HE) in the phase-4 part of the study.

Sample Size

Simulations showed the random distribution of measured accuracy for sample sizes of 100, 150, 200 and 400 recordings in the training dataset. The simulation is based on binary classification. For sub-classification, similar results can be expected of a similar level of accuracy reached (if it is less accurate, then the random variation increases). The diagram illustrates this.

For an expected sensitivity of 75%, specificity of 90%, with a confidence interval of ±5%, we needed at least 85 EEGs (Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20:453–458).



Tools and procedures

Excel has been selected for use by the participants as it is easily accessible and generally well understood. SharePoint has been selected as it is an easily managed tool that meets the needs of accessibility while maintaining the maintaining the integrity of the study. For each human expert, a Virtual Machine is set up to host the NeuroWorks EEG software (version 9.2.0.6628-54426). The number of human experts need to be at least seven. Previous studies on inter-rater variability in EEG showed that majority consensus of a panel of human experts does not change significantly beyond seven raters. There will be an even distribution of HEs from North America and Europe. All the HEs are board certified in Clinical Neurophysiology, or hold specialty competence within Clinical Neurophysiology or Neurology including EEG reading competence.

Instructions for Human Experts

The HE will get instructions for how to:

- Open the virtual machine (VM) where the necessary infrastructure is set up for each individual HE.
- How to operate the EEG software
- Subgroup definitions
- How to report the assessments of each EEG in an Excel sheet installed at the VM.
- How to send a screenshot of the finalized Excel sheet to Holberg when all EEGs are assessed.

EEG Data Provision
All EEGs have been provided to the study under a legal contract with the relevant institution, which have been responsible for anonymization of the data, which has removed the need for individual patient consent.

Data Evaluation

1. The Excel sheet has been set up with data validation to ensure that only relevant data are inserted.
2. The Excel sheet has been configured to prevent editing by HE of cells that have already been prefilled by Holberg.
3. After HE has finalized all their assessments in the Excel sheet, they are instructed to take a screenshot and send this to Holberg.
4. The HEs will also be send a wet signed copy of the final assessment sheet.
5. The SharePoint and dedicated inbox will be monitored by the Clinical & RA Manager.

Overall trial start date

June 1st, 2021.

Ethics Review
IRB and data safety approval.
Reference number: "Sagsnr. 0100256". Date: July 7th 2020
Contact details: Pernille Worm (legal counsel, DPO) Direktionssekretariatet, Kolonivej 1, st., 4293 Dianalund. Phone: 58264200. Email: pwo@filadelfia.dk

IPD sharing statement

Individual clinical trial participant-level data (IPD) will be shared upon request.
Contact: Professor Sandor Beniczky, Danish Epilepsy Centre and Aarhus University, Denmark (Visby Allé 5, 4293 Dianalund, Denmark; Phone:+4526981536; Email: sbz@filadelfia.dk).
Type of data: For the phase-3 dataset the anonymised EEG, Diagnostic Gold standard; Demographics (age, gender), output of the algorithm will be available upon request.
Data will be available upon request, for 10 years from the publication, for scientific non-commercial use. As the dataset is de-identified, there is no need for consent from the participants.

eReferences

1. Beniczky, Sándor, Harald Aurlien, Jan C Brøgger, Anders Fuglsang-Frederiksen, António Martins-da-Silva, Eugen Trinka, Gerhard Visser, Guido Rubboli, Helle Hjalgrim, and Hermann Stefan. 2013. "Standardized computer-based organized reporting of EEG: SCORE." Epilepsia 54 (6): 1112-1124.
2. Beniczky, Sándor, Harald Aurlien, Jan C Brøgger, Lawrence J Hirsch, Donald L Schomer, Eugen Trinka, Ronit M Pressler, Richard Wennberg, Gerhard H Visser, and Monika Eisermann. 2017. "Standardized computer-based organized reporting of EEG: SCORE–second version." Clinical Neurophysiology 128 (11): 2334-2346.
3. Brogger, Jan, Tom Eichele, Eivind Aanestad, Henning Olberg, Ina Hjelland, and Harald Aurlien. 2018. "Visual EEG reviewing times with SCORE EEG." Clinical neurophysiology practice 3: 59-64.
4. Ng, Marcus C, and Kara Gillis. 2017. "The state of everyday quantitative EEG use in Canada: a national technologist survey." Seizure 49: 5-7.
5. Stroink, Hans, Robbert-Jan Schimsheimer, Al W de Weerd, Ada T Geerts, Willem F Arts, Els A Peeters, Oebele F Brouwer, A Boudewijn Peters, and Cees A van Donselaar. 2006. "Interobserver reliability of visual interpretation of electroencephalograms in children with newly diagnosed seizures." Developmental medicine and child neurology 48 (5): 374-377.
6. Tatum IV, William O, Olga Selioutski, Juan G Ochoa, Heidi Munger Clary, Janna Cheek, Frank W Drislane, and Tammy N Tsuchida. 2016. "American clinical neurophysiology society guideline 7: guidelines for EEG reporting." The Neurodiagnostic Journal 56 (4): 285-293.
7. van Donselaar, Cees A, Robert-Jan Schimsheimer, Ada T Geerts, and August C Declerck. 1992. "Value of the electroencephalogram in adult patients with untreated idiopathic first seizures." Archives of neurology 49 (3): 231-237.
8. HEEG-autoSCORE-IU-01, Intended Use Record autoSCORE Revision 3.0
9. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20:453–458