# opensearch_knn

January 22, 2024

[ ]:
```
from Bio import Entrez
import json
from torch.utils.data import Dataset
from transformers import AutoTokenizer, AutoModel,␣
↪BertForSequenceClassification, AutoModelForQuestionAnswering
import torch
import scipy
from torch.utils.data import DataLoader
import pickle
from transformers import pipeline
import os
from transformers import AutoModelForCausalLM, AutoTokenizer
from opensearchpy import OpenSearch
import requests
```

```
2024-01-22 11:58:17.954524: I tensorflow/core/util/port.cc:113] oneDNN custom
operations are on. You may see slightly different numerical results due to
floating-point round-off errors from different computation orders. To turn them
off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-01-22 11:58:18.009852: I external/local_tsl/tsl/cuda/cudart_stub.cc:31]
Could not find cuda drivers on your machine, GPU will not be used.
2024-01-22 11:58:18.295416: E
external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:9261] Unable to register
cuDNN factory: Attempting to register factory for plugin cuDNN when one has
already been registered
2024-01-22 11:58:18.295523: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:607] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
2024-01-22 11:58:18.346474: E
external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1515] Unable to
register cuBLAS factory: Attempting to register factory for plugin cuBLAS when
one has already been registered
2024-01-22 11:58:18.458330: I external/local_tsl/tsl/cuda/cudart_stub.cc:31]
Could not find cuda drivers on your machine, GPU will not be used.
2024-01-22 11:58:18.459836: I tensorflow/core/platform/cpu_feature_guard.cc:182]
This TensorFlow binary is optimized to use available CPU instructions in
performance-critical operations.
```

```
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other
operations, rebuild TensorFlow with the appropriate compiler flags.
2024-01-22 11:58:19.690903: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
find TensorRT
```

[ ]:
```python
if torch.cuda.is_available():
  device = "cuda:0"
else:
  device = "cpu"
device
```

[ ]:
```
'cpu'
```

[ ]:
```python
# https://opensearch.org/docs/latest/clients/python-low-level/

host = 'localhost'
port = 9200
auth = ('admin', 'admin')


# Create the client with SSL/TLS enabled, but hostname verification disabled.
client = OpenSearch(
    hosts = [{'host': host, 'port': port}],
    http_compress = True, # enables gzip compression for request bodies
    http_auth = auth,
    use_ssl = True,
    verify_certs = False,
    ssl_assert_hostname = False,
    ssl_show_warn = False,
)
```

### 0.1 Load Data

- only 10.000 results possible this way
- for more results use bash cli: https://www.nlm.nih.gov/dataguide/edirect/install.html#edirect-installation

[ ]:
```python
index_name = 'pub_med_index'
index_body = {
    'settings': {
    'index': {
      'number_of_shards': 4
    },
      'mappings': {
        # Your index mappings here
    }
  }
```

```
}

response = client.indices.create(index_name, body=index_body)
```

[ ]:
```
pubmed_data_path = "/home/chris/University/NLP_project/pubmed_data.json"
pubmed_data_preprocessed_path = "/home/chris/University/NLP_project/
 ↪pubmed_data_preprocessed.json"

if not os.path.exists(pubmed_data_preprocessed_path):
  with open(pubmed_data_path, 'r') as f:
    records = json.loads(f.read())

  records = records["PubmedArticle"]
  preprocessed_records = []
  for idx, article in enumerate(records):
      if (not "Abstract" in article["MedlineCitation"]["Article"].keys()):␣
 ↪continue
      article = {
          "_id": article["MedlineCitation"]["PMID"],
          "title": article["MedlineCitation"]["Article"]["ArticleTitle"],
          "text": " ".
 ↪join(article["MedlineCitation"]["Article"]["Abstract"]["AbstractText"]) #␣
 ↪some abstracts are split in an array
      }
      response
      preprocessed_records.append(article)
  with open(pubmed_data_preprocessed_path, 'w') as f:
    f.write(json.dumps(preprocessed_records))
else:
    with open(pubmed_data_preprocessed_path, 'r') as f:
        preprocessed_records = json.loads(f.read())
```

[ ]:
```
class PubMedDataset(Dataset):
    def __init__(self, path):
        with open(path, 'r') as f:
          self.data = json.loads(f.read())

    def __len__(self):
        return len(self.data)

    def __getitem__(self, idx):
        sample = self.data[idx]["text"]
        return sample
```

### 0.2 Embedding

[ ]:
```
tokenizer = AutoTokenizer.from_pretrained('sentence-transformers/
 ↪all-mpnet-base-v2')
model = AutoModel.from_pretrained('sentence-transformers/all-mpnet-base-v2').
 ↪to(device)
```

[ ]:
```
preprocessed_records[0:2]
```

[ ]:
```
[{'_id': '38085539',
  'title': 'High Seebeck Coefficient Inorganic
Ge<sub>15</sub>Ga<sub>10</sub>Te<sub>75</sub> Core/Polymer Cladding Fibers for
Respiration and Body Temperature Monitoring.',
  'text': 'Wearable thermal sensors based on thermoelectric (TE) materials with
high sensitivity and temperature resolution are extensively used in medical
diagnosis, human-machine interfaces, and advanced artificial intelligence.
However, their development is greatly limited by the lack of materials with both
a high Seebeck coefficient and superior anticrystallization ability. Here, a new
inorganic amorphous TE material, Ge<sub>15</sub>Ga<sub>10</sub>Te<sub>75</sub>,
with a high Seebeck coefficient of 1109  V/K is reported. Owing to the large
difference between the glass-transition temperature and initial crystallization
temperature, Ge<sub>15</sub>Ga<sub>10</sub>Te<sub>75</sub> strongly inhibits
crystallization during fiber fabrication by thermally codrawing a precast rod
comprising a Ge<sub>15</sub>Ga<sub>10</sub>Te<sub>75</sub> core and PP polymer
cladding. The temperature difference can be effectively transduced into
electrical signals to achieve TE fiber thermal sensing with an accurate
temperature resolution of 0.03 K and a fast response time of 4 s. It is
important to note that after the 1.5 and 5.5 K temperatures changed repeatedly,
the TE properties of the fiber demonstrated high stability. Based on the Seebeck
effect and superior flexibility of the fibers, they can be integrated into a
mask and wearable fabric for human respiration and body temperature monitoring.
The superior thermal sensing performance of the TE fibers together with their
natural flexibility and scalable fabrication endow them with promising
applications in health-monitoring and intelligent medical systems.'},
 {'_id': '38085351',
  'title': 'Neuro-fuzzy modelling of a continuous stirred tank bioreactor with
ceramic membrane technology for treating petroleum refinery effluent: a case
study from Assam, India.',
  'text': 'A continuous stirred tank bioreactor (CSTB) with cell recycling
combined with ceramic membrane technology and inoculated with Rhodococcus opacus
PD630 was employed to treat petroleum refinery wastewater for simultaneous
chemical oxygen demand (COD) removal and lipid production from the retentate
obtained during wastewater treatment. In the present study, the COD removal
efficiency (COD<sub>RE</sub>) (%) and lipid concentration (g/L) were predicted
using two artificial intelligence models, i.e., an artificial neural network
(ANN) and a neuro-fuzzy neural network (NF-NN) with a network topology of 6-25-2
being the best for NF-NN. The results revealed the superiority of NF-NN over ANN
```

```
in terms of determination coefficient (R<sup>2</sup>), root mean square error
(RMSE), and mean absolute percentage error (MAPE). Three learning algorithms
were tested with NF-NN; among them, the Bayesian regularization backpropagation
(BR-BP) outperformed others. The sensitivity analysis revealed that, if solid
retention time and biomass concentrations were maintained between 35 and 75\xa0h
and 3.0\xa0g/L and 3.5\xa0g/L, respectively, high COD<sub>RE</sub> (93%) and
lipid concentration (2.8\xa0g/L) could be obtained consistently.'}]
```

[ ]:
```
dataset = PubMedDataset(pubmed_data_preprocessed_path)
dataloader = DataLoader(dataset, batch_size=64, shuffle=False)
```

[ ]:
```
# why not take cls token?
def mean_pooling(last_hidden_state, attention_mask):
    input_mask_expanded = attention_mask.unsqueeze(-1).expand(last_hidden_state.
 ↪size()).float()
    return torch.sum(last_hidden_state * input_mask_expanded, 1) / torch.
 ↪clamp(input_mask_expanded.sum(1), min=1e-9)
```

[ ]:
```
embeddings = []
with torch.no_grad():
    for i, sample in enumerate(dataloader):
        inputs = tokenizer(sample, return_tensors="pt", padding=True,␣
 ↪truncation=True).to(device)
        out = model(**inputs)
        pooled = mean_pooling(out.last_hidden_state, inputs["attention_mask"]).
 ↪to("cpu")
        embeddings.extend(pooled)
embeddings_stacked = torch.stack(embeddings)
```

[ ]:
```
torch.save(embeddings_stacked, '/home/chris/University/NLP_project/
 ↪pubmed_data_embeddings.bin')
```

[ ]:
```
# Define an index mapping with a custom analyzer
index_mapping = {
  "settings": {
    "index.knn": True
  },
  "mappings": {
    "properties": {
      "title": {
        "type": "text",
        "analyzer": "standard"
      },
      "text": {
        "type": "text",
        "analyzer": "standard"
      },
```

```
      "vector": {
        "type": "knn_vector",
        "dimension": len(embeddings_stacked[0])
      }
    }
  }
}


# Create the index with the custom mapping
index_name = "pub_med_index"
client.indices.create(index=index_name, body=index_mapping)
```

[ ]:
```
{'acknowledged': True, 'shards_acknowledged': True, 'index': 'pub_med_index'}
```

[ ]:
```
actions = [
    ({"index": {"_index": "pub_med_index", "_id":doc["_id"] }},{"title":␣
 ↪doc["title"], "text": doc["text"], "vector": embeddings_stacked[num].
 ↪tolist() })
for num,doc in enumerate(preprocessed_records[:1000])
]
```

[ ]:
```
request = '\n'.join([f'{json.dumps(item, indent=None, separators=(",", ":"))}'␣
 ↪for tpl in actions for item in tpl])
```

[ ]:
```
try:
    response = client.bulk(body=request, refresh=True)
    print("Bulk request successful.")
except Exception as e:
    print(f"Failed to perform bulk request. Error: {e}")
```

```
Bulk request successful.
```

[ ]:
```
question = "What is the influence of alcohol on minors?"

inputs = tokenizer(question, return_tensors="pt", padding=True,␣
 ↪truncation=True).to(device)
query_outputs = mean_pooling(model(**inputs).last_hidden_state,␣
 ↪inputs["attention_mask"]).to("cpu")
print(len(query_outputs[0].tolist()))

# Define the KNN search query
knn_query = {
    "size": 5,
    "_source": ["title", "text"],
    "query": {
        "knn": {
```

```
            "vector": {
                "vector": query_outputs[0].tolist(),
                "k": 5
            }
        }
    }
}

# Perform the KNN search
response = client.search(index=index_name, body=knn_query)
```

768

[ ]:
```
response
```

[ ]:
```
{'took': 14,
 'timed_out': False,
 '_shards': {'total': 1, 'successful': 1, 'skipped': 0, 'failed': 0},
 'hits': {'total': {'value': 5, 'relation': 'eq'},
  'max_score': 0.097037144,
  'hits': [{'_index': 'pub_med_index',
    '_id': '38033004',
    '_score': 0.097037144,
    '_source': {'text': 'Alexithymia can be associated with worse addictive
traits, while emotional intelligence is associated with better addictive
outcomes. In Lebanon, the prevalence of cigarette and waterpipe smoking is on
the rise, although people are aware of the associated harms. Also, around 11% of
Lebanese adults have experienced alcohol use disorder (AUD). This study aimed to
assess the association between alexithymia, emotional intelligence, smoking
(cigarette and waterpipe), and AUD among a sample of Lebanese adults. A web-
based cross-sectional study carried out between February and April 2020, during
the lockdown period, enrolled 408 community-dwelling adults. The survey link was
shared on social media to reach participants from all Lebanese
districts/governorates. Taking antidepressants (Beta = 4.37) was significantly
associated with more cigarette dependence, while female gender (Beta = -1.52)
and having a high vs. low monthly income (Beta = 1.02) were significantly
associated with less cigarette dependence. None of the variables, including
alexithymia, were significantly associated with waterpipe dependence. Female
gender (Beta = -0.15) and higher emotional intelligence (Beta = -0.003) were
significantly associated with less AUD, whereas higher alexithymia (Beta =
0.003) was significantly associated with more AUD. This study could demonstrate
a significant association between alexithymia and cigarette smoking and
alexithymia and alcohol consumption. Future research is warranted to investigate
the mediating effect of emotional intelligence and how these results may be used
to meet the needs of alexithymic individuals with addictions.',
     'title': 'Association between alexithymia, emotional intelligence, smoking
addiction, and alcohol use disorder among a sample of Lebanese adults.'}},
```

{'_index': 'pub_med_index',
 '_id': '38058999',
 '_score': 0.095024,
 '_source': {'text': 'To evaluate whether prenatal tobacco exposure (PTE) is related to poorer cognitive performance, abnormal brain morphometry, and whether poor cognitive performance is mediated by PTE-related structural brain differences. The Adolescent Brain Cognitive Development study dataset was used to compare structural MRI data and neurocognitive (NIH Toolbox<sup>®</sup>) scores in 9-to-10-year-old children with (n=620) and without PTE (n=10,989). We also evaluated whether PTE effects on brain morphometry mediated PTE effects on neurocognitive scores. Group effects were evaluated using Linear Mixed Models, covaried for socio-demographics and prenatal exposures to alcohol and/or marijuana, and corrected for multiple comparisons using the false-discovery rate (FDR). Compared to unexposed children, those with PTE had poorer performance (all p-values <0.05) on executive function, working memory, episodic memory, reading decoding, crystallized intelligence, fluid intelligence and overall cognition. Exposed children also had thinner parahippocampal gyri, smaller surface areas in the posterior-cingulate and pericalcarine cortices; the lingual and inferior parietal gyri, and smaller thalamic volumes (all p-values <0.001). Furthermore, among children with PTE, girls had smaller surface areas in the superior-frontal (interaction-FDR-p=0.01), precuneus (interaction-FDR-p=0.03) and postcentral gyri (interaction-FDR-p=0.02), while boys had smaller putamen volumes (interaction-FDR-p=0.02). Smaller surface areas across regions of the frontal and parietal lobes, and lower thalamic volumes, partially mediated the associations between PTE and poorer neurocognitive scores (p-values <0.001). Our findings suggest PTE may lead to poorer cognitive performance and abnormal brain morphometry, with sex-specific effects in some brain regions, in pre-adolescent children. The poor cognition in children with PTE may result from the smaller areas and subcortical brain volumes.',
  'title': 'Prenatal tobacco exposure on brain morphometry partially mediated poor cognitive performance in preadolescent children.'}},
{'_index': 'pub_med_index',
 '_id': '38002647',
 '_score': 0.0889211,
 '_source': {'text': 'The aims of this review are to provide a comprehensive overview of the definition and scope of pharmacoepidemiology, to summarize the study designs and methodologies used in the field, to discuss the future trends in the field and new methodologies to address bias and confounding, and finally to give some recommendations to clinicians interested in pharmacoepidemiologic research. Because drug efficacy and safety from randomized clinical trials do not reflect the real-world situation, pharmacoepidemiological studies on drug safety monitoring and drug effectiveness in large numbers of people are needed by healthcare professionals and regulatory institutions. We aim to highlight the importance of pharmacoepidemiologic research in informing evidence-based medicine and public health policy. The development of new designs and methodologies for the generation of valid evidence, as well as new initiatives to provide guidance and recommendations on how to incorporate real-world

evidence into the drug development process, are reported on. In addition, we have touched on the implication of artificial intelligence in the management of real-world data. This overview aims to summarize all important aspects to consider when conducting or interpreting a pharmacoepidemiologic study.',
'title': 'Pharmacoepidemiology: An Overview.'}},
{'_index': 'pub_med_index',
'_id': '38009544',
'_score': 0.08775746,
'_source': {'text': "Compare by occurrence-era and age-group how opioid-related deaths (ORDs) and their counterpart evolved in Scotland versus England and Wales during 2006-2020. For Scotland, compare co-implication rates between ORDs and non-ORDs for any benzodiazepine; cocaine; gabapentin/pregabalin; and consider whether co-implication in ORDs depended on opioid-specificity. Cross-tabulations of drug misuse deaths (DMDs) obtained by 3-yearly occurrence-era (2006-2008 to 2018-2020) and age-group (under 25, 25-34, 35-44, 45-54, 55+ years) for England and Wales and subdivided by whether at least one opiate was mentioned on death certificate (DMD-Os or not); and of Scotland's opioid-related deaths (ORDs versus non-ORDs) together with i) co-implication by any benzodiazepine; or cocaine; gabapentin/pregabalin and ii) opioid-specificity of ORDs. ORD is defined by heroin/morphine (H) or methadone (M) or buprenorphine (B) being implicated in DMD. Per era between 2012-2014 and 2018-2020, Scotland's ORDs increased by 54% and non-ORDs by 34%. Increase in DMD-Os in England and Wales was more modest. Cocaine was implicated in 83% of Scotland's 2,690 non-ORDs during 2006-2020; and any benzodiazepine in 53% of 8,409 ORDs. However, in 2018-2020, co-implication rates in 2,926 ORDs (880 non-ORDs) were 81% (33%) for any benzodiazepine; 30% (74%) for cocaine; and 38% (22%) for gabapentin/pregabalin. Co-implication rate in 2018-2020 for any benzodiazepine was lowest at 70% (616/877) for H-ORDs; and, by age-group, at 66% (160/241) for ORDs aged 55+ years. Drug-testing to inform users, shared intelligence between police and public health for earlier detection of changes in supply and monitoring of prescribed daily-dose of methadone are urgent.",
'title': 'Opioid-related deaths and their counterpart by occurrence-era, age-group and co-implicated drugs: Scotland versus England and Wales.'}},
{'_index': 'pub_med_index',
'_id': '37999530',
'_score': 0.08672883,
'_source': {'text': "The knowledge of the effects of organophosphate flame retardants on children's neurodevelopment is limited. The purpose of the present research is to evaluate the association between exposure to organophosphate flame retardants and children's neurodevelopment in two European cohorts involved in the Human Biomonitoring Initiative Aligned Studies. The participants were school-aged children belonging to the Odense Child Cohort (Denmark) and the PCB cohort (Slovakia). In each cohort, the children's neurodevelopment was assessed through the Full-Scale Intelligence Quotient score of the Wechsler Intelligence Scale for Children, using two different editions. The children's urine samples, collected at one point in time, were analyzed for several metabolites of organophosphate flame retardants. The association between

neurodevelopment and each organophosphate flame retardant metabolite was explored by applying separate multiple linear regressions based on the approach of MM-estimation in each cohort. In the Danish cohort, the mean ± standard deviation for the neurodevelopment score was 98 ± 12; the geometric mean (95% confidence interval (95% CI)) of bis(1,3-dichloro-2-propyl) phosphate (BDCIPP) standardized by creatinine (crt) was 0.52 µg/g crt (95% CI = 0.49; 0.60), while that of diphenyl phosphate (DPHP) standardized by crt was 1.44 µg/g crt (95% CI = 1.31; 1.58). The neurodevelopment score showed a small, negative, statistically imprecise trend with BDCIPP standardized by crt (  = -1.30; 95%CI = -2.72; 0.11; p-value = 0.07) and no clear association with DPHP standardized by crt (  = -0.98; 95%CI = -2.96; 0.99; p-value = 0.33). The neurodevelopment score showed a negative trend with BDCIPP (  = -1.42; 95% CI = -2.70; -0.06; p-value = 0.04) and no clear association with DPHP (  = -1.09; 95% CI = -2.87; 0.68; p-value = 0.23). In the Slovakian cohort, the mean ± standard deviation for the neurodevelopment score was 81 ± 15; the geometric mean of BDCIPP standardized by crt was 0.18 µg/g crt (95% CI = 0.16; 0.20), while that of DPHP standardized by crt was 2.24 µg/g crt (95% CI = 2.00; 3.52). The association of the neurodevelopment score with BDCIPP standardized by crt was -0.49 (95%CI = -1.85; 0.87; p-value = 0.48), and with DPHP standardized by crt it was -0.35 (95%CI = -1.90; 1.20; p-value = 0.66). No clear associations were observed between the neurodevelopment score and BDCIPP/DPHP concentrations that were not standardized by crt. No clear associations were observed with bis(1-chloro-2-propyl) phosphate (BCIPP) in either cohort, due to the low detection frequency of this compound. In conclusion, this study provides only limited evidence of an inverse association between neurodevelopment and exposure to BDCIPP and DPHP. The timing of exposure and effect modification of other organophosphate flame retardant metabolites and other substances should be the subject of further investigations that address this scientific hypothesis.",
    'title': 'Cognitive Performance and Exposure to Organophosphate Flame Retardants in Children: Evidence from a Cross-Sectional Analysis of Two European Mother-Child Cohorts.'}}]}}

[ ]:
```
response['hits']['hits'][0]['_source']['text']
```

[ ]: 'Alexithymia can be associated with worse addictive traits, while emotional intelligence is associated with better addictive outcomes. In Lebanon, the prevalence of cigarette and waterpipe smoking is on the rise, although people are aware of the associated harms. Also, around 11% of Lebanese adults have experienced alcohol use disorder (AUD). This study aimed to assess the association between alexithymia, emotional intelligence, smoking (cigarette and waterpipe), and AUD among a sample of Lebanese adults. A web-based cross-sectional study carried out between February and April 2020, during the lockdown period, enrolled 408 community-dwelling adults. The survey link was shared on social media to reach participants from all Lebanese districts/governorates. Taking antidepressants (Beta = 4.37) was significantly associated with more cigarette dependence, while female gender (Beta = -1.52) and having a high vs.

low monthly income (Beta = 1.02) were significantly associated with less cigarette dependence. None of the variables, including alexithymia, were significantly associated with waterpipe dependence. Female gender (Beta = -0.15) and higher emotional intelligence (Beta = -0.003) were significantly associated with less AUD, whereas higher alexithymia (Beta = 0.003) was significantly associated with more AUD. This study could demonstrate a significant association between alexithymia and cigarette smoking and alexithymia and alcohol consumption. Future research is warranted to investigate the mediating effect of emotional intelligence and how these results may be used to meet the needs of alexithymic individuals with addictions.'

[ ]:
```
response['hits']['hits'][1]['_source']['text']
```

[ ]: 'To evaluate whether prenatal tobacco exposure (PTE) is related to poorer cognitive performance, abnormal brain morphometry, and whether poor cognitive performance is mediated by PTE-related structural brain differences. The Adolescent Brain Cognitive Development study dataset was used to compare structural MRI data and neurocognitive (NIH Toolbox<sup>®</sup>) scores in 9-to-10-year-old children with (n=620) and without PTE (n=10,989). We also evaluated whether PTE effects on brain morphometry mediated PTE effects on neurocognitive scores. Group effects were evaluated using Linear Mixed Models, covaried for socio-demographics and prenatal exposures to alcohol and/or marijuana, and corrected for multiple comparisons using the false-discovery rate (FDR). Compared to unexposed children, those with PTE had poorer performance (all p-values <0.05) on executive function, working memory, episodic memory, reading decoding, crystallized intelligence, fluid intelligence and overall cognition. Exposed children also had thinner parahippocampal gyri, smaller surface areas in the posterior-cingulate and pericalcarine cortices; the lingual and inferior parietal gyri, and smaller thalamic volumes (all p-values <0.001). Furthermore, among children with PTE, girls had smaller surface areas in the superior-frontal (interaction-FDR-p=0.01), precuneus (interaction-FDR-p=0.03) and postcentral gyri (interaction-FDR-p=0.02), while boys had smaller putamen volumes (interaction-FDR-p=0.02). Smaller surface areas across regions of the frontal and parietal lobes, and lower thalamic volumes, partially mediated the associations between PTE and poorer neurocognitive scores (p-values <0.001). Our findings suggest PTE may lead to poorer cognitive performance and abnormal brain morphometry, with sex-specific effects in some brain regions, in pre-adolescent children. The poor cognition in children with PTE may result from the smaller areas and subcortical brain volumes.'

[ ]:
```
response['hits']['hits'][2]['_source']['text']
```

[ ]: 'The aims of this review are to provide a comprehensive overview of the definition and scope of pharmacoepidemiology, to summarize the study designs and methodologies used in the field, to discuss the future trends in the field and new methodologies to address bias and confounding, and finally to give some recommendations to clinicians interested in pharmacoepidemiologic research.

Because drug efficacy and safety from randomized clinical trials do not reflect the real-world situation, pharmacoepidemiological studies on drug safety monitoring and drug effectiveness in large numbers of people are needed by healthcare professionals and regulatory institutions. We aim to highlight the importance of pharmacoepidemiologic research in informing evidence-based medicine and public health policy. The development of new designs and methodologies for the generation of valid evidence, as well as new initiatives to provide guidance and recommendations on how to incorporate real-world evidence into the drug development process, are reported on. In addition, we have touched on the implication of artificial intelligence in the management of real-world data. This overview aims to summarize all important aspects to consider when conducting or interpreting a pharmacoepidemiologic study.'

[ ]:
```
response['hits']['hits'][3]['_source']['text']
```

[ ]: "Compare by occurrence-era and age-group how opioid-related deaths (ORDs) and their counterpart evolved in Scotland versus England and Wales during 2006-2020. For Scotland, compare co-implication rates between ORDs and non-ORDs for any benzodiazepine; cocaine; gabapentin/pregabalin; and consider whether co-implication in ORDs depended on opioid-specificity. Cross-tabulations of drug misuse deaths (DMDs) obtained by 3-yearly occurrence-era (2006-2008 to 2018-2020) and age-group (under 25, 25-34, 35-44, 45-54, 55+ years) for England and Wales and subdivided by whether at least one opiate was mentioned on death certificate (DMD-Os or not); and of Scotland's opioid-related deaths (ORDs versus non-ORDs) together with i) co-implication by any benzodiazepine; or cocaine; gabapentin/pregabalin and ii) opioid-specificity of ORDs. ORD is defined by heroin/morphine (H) or methadone (M) or buprenorphine (B) being implicated in DMD. Per era between 2012-2014 and 2018-2020, Scotland's ORDs increased by 54% and non-ORDs by 34%. Increase in DMD-Os in England and Wales was more modest. Cocaine was implicated in 83% of Scotland's 2,690 non-ORDs during 2006-2020; and any benzodiazepine in 53% of 8,409 ORDs. However, in 2018-2020, co-implication rates in 2,926 ORDs (880 non-ORDs) were 81% (33%) for any benzodiazepine; 30% (74%) for cocaine; and 38% (22%) for gabapentin/pregabalin. Co-implication rate in 2018-2020 for any benzodiazepine was lowest at 70% (616/877) for H-ORDs; and, by age-group, at 66% (160/241) for ORDs aged 55+ years. Drug-testing to inform users, shared intelligence between police and public health for earlier detection of changes in supply and monitoring of prescribed daily-dose of methadone are urgent."

[ ]:
```
response['hits']['hits'][4]['_source']['text']
```

[ ]: "The knowledge of the effects of organophosphate flame retardants on children's neurodevelopment is limited. The purpose of the present research is to evaluate the association between exposure to organophosphate flame retardants and children's neurodevelopment in two European cohorts involved in the Human Biomonitoring Initiative Aligned Studies. The participants were school-aged children belonging to the Odense Child Cohort (Denmark) and the PCB cohort

(Slovakia). In each cohort, the children's neurodevelopment was assessed through the Full-Scale Intelligence Quotient score of the Wechsler Intelligence Scale for Children, using two different editions. The children's urine samples, collected at one point in time, were analyzed for several metabolites of organophosphate flame retardants. The association between neurodevelopment and each organophosphate flame retardant metabolite was explored by applying separate multiple linear regressions based on the approach of MM-estimation in each cohort. In the Danish cohort, the mean ± standard deviation for the neurodevelopment score was 98 ± 12; the geometric mean (95% confidence interval (95% CI)) of bis(1,3-dichloro-2-propyl) phosphate (BDCIPP) standardized by creatinine (crt) was 0.52 µg/g crt (95% CI = 0.49; 0.60), while that of diphenyl phosphate (DPHP) standardized by crt was 1.44 µg/g crt (95% CI = 1.31; 1.58). The neurodevelopment score showed a small, negative, statistically imprecise trend with BDCIPP standardized by crt (  = -1.30; 95%CI = -2.72; 0.11; p-value = 0.07) and no clear association with DPHP standardized by crt (  = -0.98; 95%CI = -2.96; 0.99; p-value = 0.33). The neurodevelopment score showed a negative trend with BDCIPP (  = -1.42; 95% CI = -2.70; -0.06; p-value = 0.04) and no clear association with DPHP (  = -1.09; 95% CI = -2.87; 0.68; p-value = 0.23). In the Slovakian cohort, the mean ± standard deviation for the neurodevelopment score was 81 ± 15; the geometric mean of BDCIPP standardized by crt was 0.18 µg/g crt (95% CI = 0.16; 0.20), while that of DPHP standardized by crt was 2.24 µg/g crt (95% CI = 2.00; 3.52). The association of the neurodevelopment score with BDCIPP standardized by crt was -0.49 (95%CI = -1.85; 0.87; p-value = 0.48), and with DPHP standardized by crt it was -0.35 (95%CI = -1.90; 1.20; p-value = 0.66). No clear associations were observed between the neurodevelopment score and BDCIPP/DPHP concentrations that were not standardized by crt. No clear associations were observed with bis(1-chloro-2-propyl) phosphate (BCIPP) in either cohort, due to the low detection frequency of this compound. In conclusion, this study provides only limited evidence of an inverse association between neurodevelopment and exposure to BDCIPP and DPHP. The timing of exposure and effect modification of other organophosphate flame retardant metabolites and other substances should be the subject of further investigations that address this scientific hypothesis."

[ ]:
```
pipe_qa = pipeline("question-answering", model="deepset/roberta-base-squad2")
summarizer = pipeline("summarization", model="facebook/bart-large-cnn")
```

```
config.json:   0%|          | 0.00/571 [00:00<?, ?B/s]
model.safetensors:   0%|          | 0.00/496M [00:00<?, ?B/s]
tokenizer_config.json:   0%|          | 0.00/79.0 [00:00<?, ?B/s]
vocab.json:   0%|          | 0.00/899k [00:00<?, ?B/s]
merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
special_tokens_map.json:   0%|          | 0.00/772 [00:00<?, ?B/s]
```

```
config.json:   0%|          | 0.00/1.58k [00:00<?, ?B/s]
model.safetensors:   0%|          | 0.00/1.63G [00:00<?, ?B/s]
generation_config.json:   0%|          | 0.00/363 [00:00<?, ?B/s]
vocab.json:   0%|          | 0.00/899k [00:00<?, ?B/s]
merges.txt:   0%|          | 0.00/456k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/1.36M [00:00<?, ?B/s]
```

[ ]:
```
text = response['hits']['hits'][0]['_source']['text']
context = summarizer(text, max_length=100, min_length=50,␣
↪do_sample=False)[0]["summary_text"]
```

[ ]:
```
pipe_qa({"context": context, "question": question})
```

[ ]:
```
{'score': 0.0009504778427071869,
 'start': 35,
 'end': 57,
 'answer': 'worse addictive traits'}
```