Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Data Management for Digital Health
Revision of Exercise III

Borchert, Dr. Schapranow

Data Management for Digital Health
Winter 2023

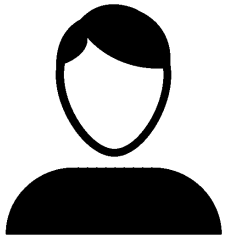
Exercise III
Topics

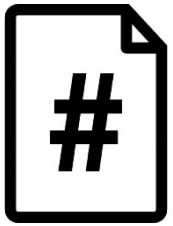
- Clinical Interpretation of Molecular Data (Guest Lecture)
- Natural Language Processing
- Neural Networks, Word Embeddings
- Foundation Models
- Medical Use Case Nephrology
- Clinical Prediction Models

Exercise II
Key Stats

Hasso Plattner Institut

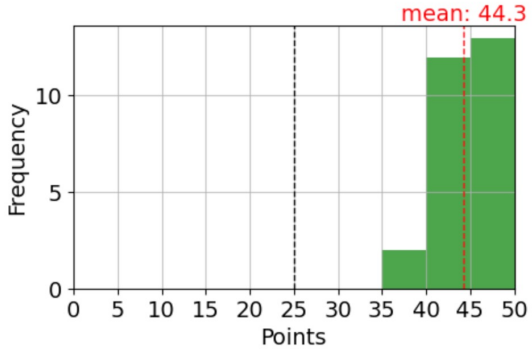
25 Questions 50 Points	26 Students 26 Passed	Average score 44.2 / 88.4%	Average time 92 min



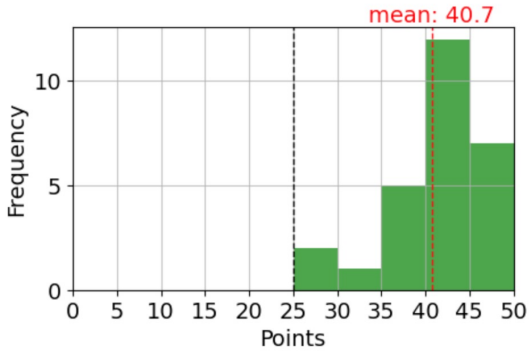


<< 3h

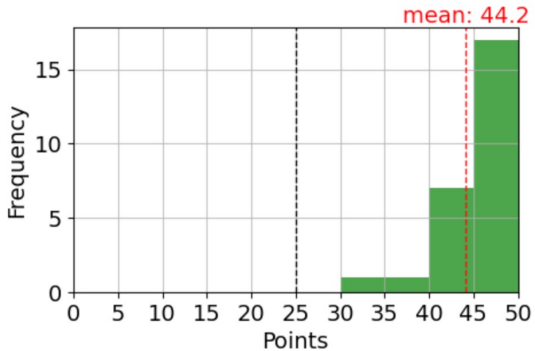
Exercise III
Key Stats



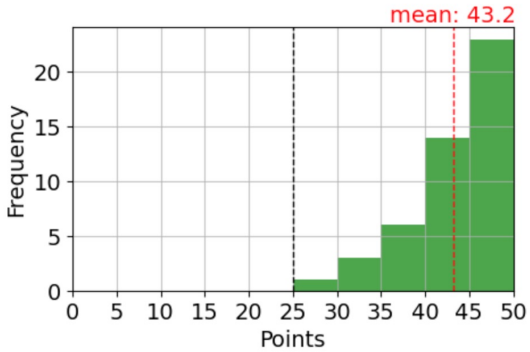
Exercise I (2023)



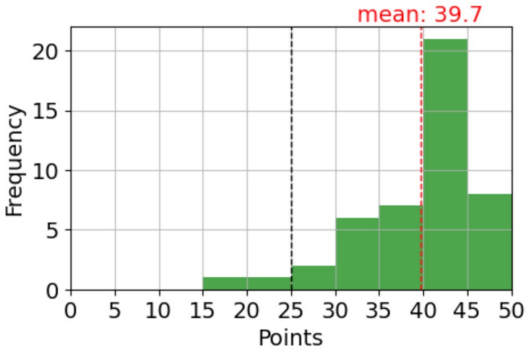
Exercise II (2023)



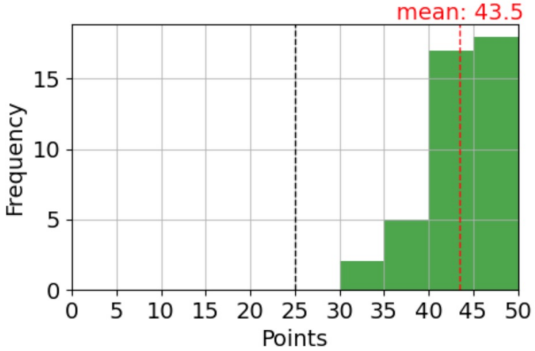
Exercise III (2023)



Exercise I (2022)



Exercise II (2022)



Exercise III (2022)

Q1 Which of the following statements are true for the use of molecular genetic data in course of precision oncology [...]?

Hasso Plattner Institut

- ✓ Clinical relevance of genetic variants is assessed through knowledge bases like CIViC or OnkoKB.
- ✓ Today, whole genome sequencing is applied only for a limited number of severe ill patients, e.g., when enrolled in the DKTK MASTER program.
- ✗ A biological rationale constitutes the highest level of evidence, e.g., using molecular pathways only.
- ✗ Clinical interpretation of molecular data is already fully automated through general-purpose large language models.

Frequently missed

Frequent incorrect answer

Q1 Which of the following statements are true for the use of molecular genetic data in course of precision oncology […]?

Hasso Plattner Institut

NGS (accredited/validated)
- Oncomine Focus/Precision DNA Assay
- Oncomine Focus/Precision RNA Assay
- ColonLung Panel V2
- Cancer Hotspot Panel
- Myeloid Panel (Custom)
- (B-cell) Lymphoma Panel
- Oncomine cfDNA (Liquid Biopsies)
- Breast cfDNA Panel (Liquid Biopsies)
- BRCA1/2 Panel
- Tumor Mutational Burden (1.7 Mbases)
- Molecular Health 600+ Panel (3 Mbases); NextSeq
- Oncomine Comprehensive Assay V4 (500+) Panel
- TSO500 (DNA/RNA) Panel
- Ig/TCR Clonality Panel
- Archer RNA Panel

IHC/FISH
- nTRK screening
- TMB
- Other Targets (e.g. HER2,
- Fusion Gene validation

Other
- e.g. EPIC (Methylom)

DKTK MASTER
- WES/WGS
- RNASeq

ExLiquid
- ctDNA

Functional Analyses

Single cell analyses

Evidence-based Assessment of Clinical Cancer Therapies

Hasso Plattner Institut

- The higher the GoR and LoE, the more evidence about a success of the desired treatment is available

GoR	LoE	Type of Study
A	1a	Systematic review of (homogeneous) randomized controlled trials
A	1b	Individual randomized controlled trials (with narrow confidence intervals)
B	2a	Systematic review of (homogeneous) cohort studies of "exposed" and "unexposed" subjects
B	2b	Individual cohort study / low-quality randomized control studies
B	3a	Systematic review of (homogeneous) case-control studies
B	3b	Individual case-control studies
C	4	Case series, low-quality cohort or case-control studies
D	5	Expert opinions based on non-systematic reviews of results or mechanistic studies

https://guides.library.stonybrook.edu/evidence-based-medicine/levels_of_evidence

Q1 Which of the following statements are true for the use of molecular genetic data in course of precision oncology [...]?

Gleiche Tumorentität	**m1A**	In der **gleichen Tumorentität** wurde der prädiktive Wert des Biomarkers oder die klinischeWirksamkeit in einer **Biomarker-stratifizierten Kohorte** einer adäquat gepowerten **prospektiven Studie** oder **Metaanalyse** gezeigt.
	m1B	In der **gleichen Tumorentität** wurde der prädiktive Wert des Biomarkers oder die klinische Wirksamkeit in einer **retrospektiven Kohorte** oder **Fall-Kontroll-Studie** gezeigt.
	m1C	Ein oder mehrere **Fallberichte** in der **gleichen Tumorentität**.
Andere Tumorentität	**m2A**	In einer **anderen Tumorentität** wurde der prädiktive Wert des Biomarkers oder die klinische Wirksamkeit in einer **Biomarker-stratifizierten Kohorte** einer adäquat gepowerten **prospektiven Studie** oder **Metaanalyse** gezeigt.
	m2B	In einer **anderen Tumorentität** wurde der prädiktive Wert des Biomarkers oder die Klinische Wirksamkeit in einer **retrospektiven Kohorte** oder **Fall-Kontroll-Studie** gezeigt.
	m2C	Unabhängig von der Tumorentität wurde beim Vorliegen des Biomarkers eine **klinische Wirksamkeit** in einem oder mehreren **Fallberichten** gezeigt.
In vitro oder Tiermodell	**m3**	**Präklinische Daten** (*in vitro*-/*in vivo*-Modelle, funktionelle Untersuchungen) zeigen eine Assoziation des Biomarkers mit der Wirksamkeit der Medikation, welche durch eine wissenschaftliche Rationale gestützt wird.
Biologische Rationale	**m4**	Eine **wissenschaftliche, biologische Rationale** legt eine Assoziation des Biomarkers mit der Wirksamkeit der Medikation nahe, welche bisher **nicht durch (prä)klinische Daten** gestützt wird.

Q4 [...] search for “FOLFIRI antineoplastic chemotherapy regimen”. [...] What can you infer from the SNOMED CT ontology?

Hasso Plattner Institut

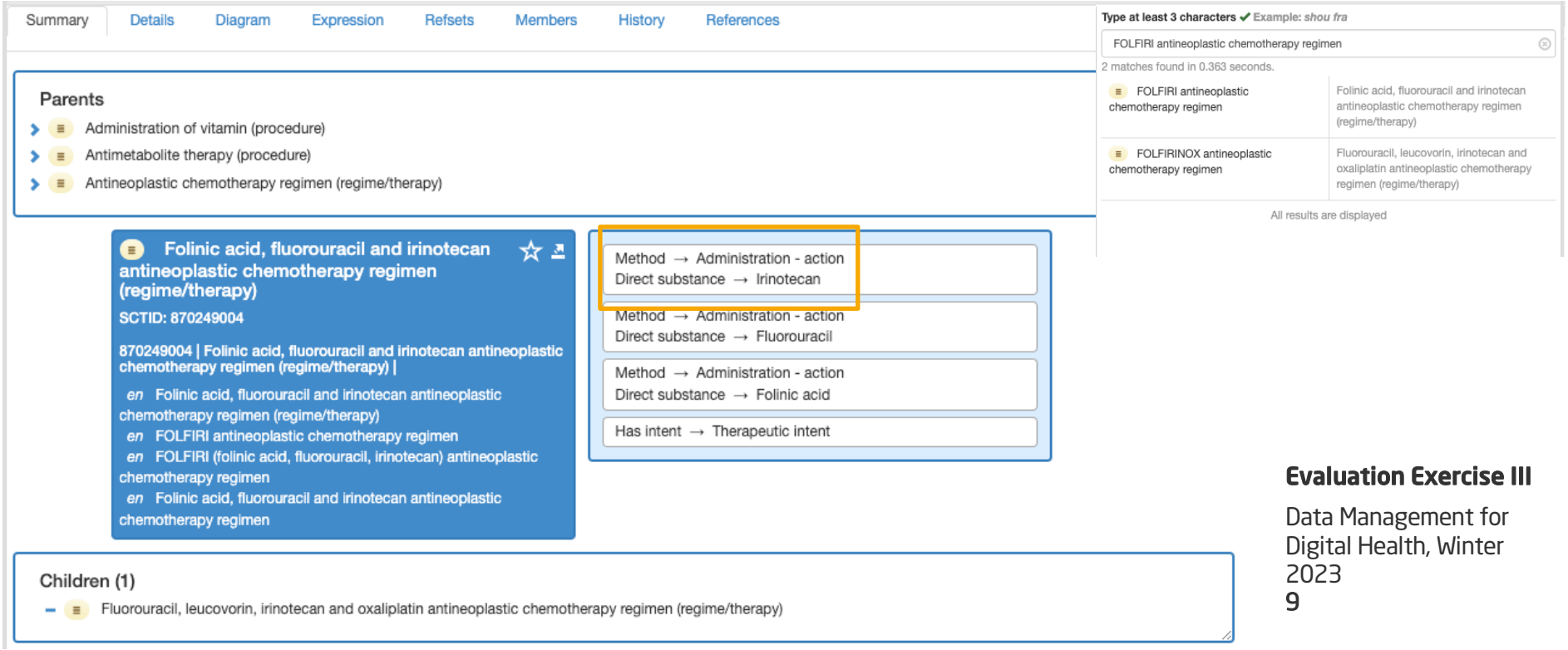
- ✓ The concept is connected to its included substances through the “Direct Substance” attribute.
- ✗ The concept only has a single parent, as SNOMED CT is a classification system.
- ✗ The concept is connected to its included substances through the “Method” attribute.
- ✗ “FOLIFIRINOX” is a preferred synonym for the concept.

Frequently missed

Frequent incorrect answer

Q4 [...] search for "FOLFIRI antineoplastic chemotherapy regimen". [...] What can you infer from the SNOMED CT ontology?

Hasso Plattner Institut



Evaluation Exercise III

Data Management for Digital Health, Winter 2023

Q14 Which statements are true about few-shot prompting in the context of large language models (LLMs)? Select all that apply:

Hasso Plattner Institut

- ✓ Few-shot prompting is a type of in-context learning, where provided examples are used as context for the LLM predictions.
- ✗ Few-shot prompting requires a fixed set of examples in the prompt to be used for each execution on a new input sample.
- ✗ As opposed to zero-shot prompting, few-shot prompting requires parameter updates to the underlying model.
- ✗ Few-shot prompting prevents LLM hallucinations.

Frequently missed

Frequent incorrect answer

Few-Shot Prompting / In-context Learning

```
Your task is to determine the smoking status
of the person described in a clinical note.

Here are some examples:
Input: "Smoker until 1999“
Output: ex-smoker

Input: "SOCIAL HISTORY: Widowed since 1972,
no tobacco, no alcohol, lives alone.“
Output: non-smoker

Input: "He is a heavy smoker and drinks 2-3
shots per day at times."
Output: current smoker

Input:
"Social History: No alcohol use and quit
tobacco greater than 25 years ago with a 10-
pack year smoking history.”
Output:
```

Prompt Template

Few-Shot Examples

Input

- In-context learning (ICL) refers to the model's ability to adapt and respond based on the immediate context provided within a prompt.
- Few-shot prompting is a type of ICL where the model is given a few examples to illustrate the task
- Examples can be fixed or based on the input

Q17 Please select all appropriate statements about the Glomerular Filtration Rate (GFR) as discussed in class.

Hasso Plattner Institut

- ✓ GFR can be used as a good indicator for the kidney function and its cleaning quality.
- ✓ Amongst others, personal serum creatine, sex, and age can be used to estimate GFR.
- ✗ Amongst others, renal blood flow (RBF) and renal plasma flow (RPF) can be used to estimate GFR.
- ✗ GFR is defined as the quotient of renal plasma flow (RPF) and urine flow rate (UFR).

Frequently missed

Frequent incorrect answer

Kidney Diseases: Glomerular Filtration Rate

- Measuring GFR (mGFR) is complex → Estimated Glomerular Filtration Rate (eGFR) is good approximation for the kidney's function and cleaning quality
- eGFR depends on multiple factors, e.g. age, sex, ethnicity, etc.
- eGFR = *141 × min(Scr/κ, 1)α × max(Scr/κ, 1)−1.209 × 0.993 Age × 1.018 [if female] _ 1.159 [if African American]* [1]
  - with Scr = serum creatinine,
  - κ is 0.7 for females, 0.9 for males,
  - α is -0.329 for females, -0.411 for males,
  - the minimum of Scr/κ or 1, and
  - the maximum of Scr/κ or 1.

[1] Levey AS et al.; CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May 5;150(9):604-12. doi: 10.7326/0003-4819-150-9-200905050-00006. Erratum in: Ann Intern Med. 2011 Sep 20;155(6):408. PMID: 19414839; PMCID: PMC2763564.

Q23 During training of your CPM, you figure out that the model performance lacks behind your expectations. Which of the following aspect can be helpful to address this?

Hasso Plattner Institut

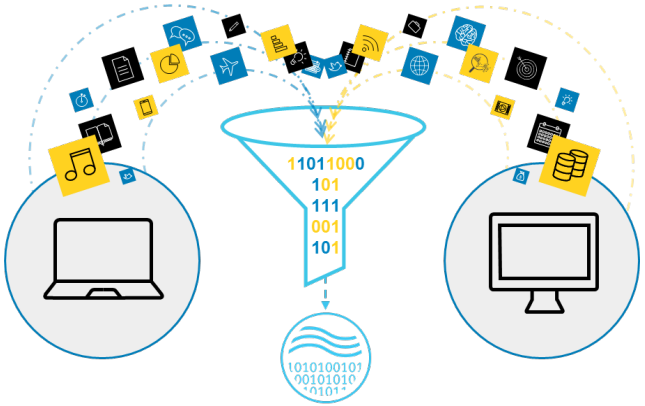
- ✓ Conduct imputation of missing values to improve prediction quality, e.g., use mean values for missing eGFR measurements.
- ✓ Consider feature engineering to improve model accuracy, e.g., select most important features for the clinical endpoint of interest and remove less important features for model training.
- ✗ Ask for additional training data from the same hospitals from 30 years ago.
- ✗ Perform min-max and variance scaling for all features of the complete training dataset.

Frequently missed

Frequent incorrect answer

3. Data Preparation: Data Transformation

- Scaling: Data may contain attributes with a mixtures of scales, but some ML methods require attributes to have the same scale
- Decomposition: Features may represent a complex concept that may be more useful to a ML method when split into its parts, e.g. data, zip code, etc.
- Aggregation: Features that might be aggregated into a single feature



https://blog.dellemc.com/en-us/digital-transformation-just-got-easier-with-analytic-insights/