Data Analysis:
Process of Data Extraction

Andrea Párniczky
Pécs, Hungary

DATA EXTRACTION

1. Decision on the aims and variables needed (researcher)

- Aims
- Variables needed
- Time period of data collection

2. Strategic consultation (researcher, consultant, coordinator, IT, statistician)

- Aims, availability of variables, derived variables, affected forms
- Format of database, steps of registry analysis

3. Data extraction

- Internal controlling (IT, registry coordinator, data management)
- Researcher controlling (researcher)

The registry is <u>SUITABLE</u> for analysing

- epidemiology
- risk factors
- course of the disease
- associations

The registry is <u>SUITABLE</u> for

- establishing protocols
- calculate sample sizes for CTs

The registry is <u>NOT SUITABLE</u> for discovering

- causality
- differences between therapies or interventions

DATA EXTRACTION – AIMS – EXAMPLES

1. To understand whether the **components of Metabolic Syndrome** have an independent effect on the outcome of AP.
2. To investigate current clinical practices and develop recommendations that guide clinicians in prescribing **antibiotic treatment** in AP – clinical parameters used in decision making.
3. To determine how **age and comorbidities** modify the outcomes in AP.
4. To assess the past and current role of CRP and WBC in clinical trials on AP (literature review) and to provide evidence from a cohort analysis to guide clinical researchers on the most appropriate **role of CRP and WBC** in future clinical trials.

DATA EXTRACTION – VARIABLES – EXAMPLES

1. MetS: Demographic data, etiology, information on the 4 components to be examined (OB, HT, HL, DM), severity, mortality, complications, LOS.
   (New-onset DM informtaion should be checked in the epicrisis description of the cases. BMI can be calculated - height and weight available, complications should be evaluated.)
2. Antibiotic treatment: 56 parameters, including age, gender, severity, mortality, complications, LOS, details about AB therapy (starting date, type of AB).
   (Type of AB and presence and source of infection should be checked in the available text data as well.
   2012-2017

DATA EXTRACTION – DATABASE – EXAMPLE

	0: BMI under 30 1: BMI 30 or above	0: no 1: yes (Szakács Zsolt CCI információja alapján, EASY-ben külön megnézve)	0: no 1: yes (Regiszterben bejelölt hyperlipidaemia + mért HTG nagyobb, mint 1.7 pirossal)	0: no 1: yes			1:under 18.5 2:18.5-24.99 3:25-29.99 4:30.00 and above			0: no 1: yes	0: no 1: yes	year	1: no data, numeric	999999:no data, 1: male 2: female	999999: no data 1: mild 2: moderate 3: severe	999999: no data 0: no 1: yes	999999: no data, numeric	999999: no data 0: no 1: yes	999999: no data 0: no 1: yes
REGISTRY PARAMETERS													PERSONAL		OUTCOME			COMPLICATIO	
Registry_code	Obesity	Hypertension	Hyperlipidemia	Diabetes	MetS factor combinations	Number of factors	BMI categories 4	Single AP, RAP, CP	Institute	Import	Only_for_epidemiology_and_genetic_analysis	Year_of_admission	Age_at_the_time_of_admission	Gender	Severity	Mortality	Length_of_hospitalization_days	Local_pancreatic_complications	Fluid_collection
1914	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	41	1	2	0	17	1	1
1913	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	75	1	1	0	6	0	0
1891	1	1	0	1	HT+OB+DM	3 factors	4	CP	Hu, Debrec	0	0	2017	82	2	3	0	21	1	1
1890	1	1	1	0	HT+OB+HL	3 factors	4	single AP	Hu, Debrec	0	0	2016	48	2	3	1	7	1	1
1888	0	0	0	0	No MS factors	No factors	3	single AP	Hu, Pécs, F	0	0	2017	52	2	1	0	12	0	0
1887	0	0	0	0	No MS factors	No factors	3	single AP	Hu, Pécs, F	0	0	2017	29	2	1	0	5	0	0
1882	0	0	0	0	No MS factors	No factors	1	single AP	Hu, Debrec	0	0	2017	20	1	1	0	4	0	0
1881	1	0	0	0	OB	1 factor	4	single AP	Hu, Debrec	0	0	2017	87	2	1	0	8	0	0
1880	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	45	1	1	0	9	0	0
1879	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	58	1	1	0	8	1	0
1868	1	0	0	0	OB	1 factor	4	single AP	Hu, Debrec	0	0	2017	21	2	1	0	8	0	0

DATA EXTRACTION – DATABASE – EXAMPLE

	0: BMI under 30 1: BMI 30 or above	0: no 1: yes (Szakács Zsolt CCI információ alapján, EASY-ben külön megnézve)	0: no 1: yes (Regisztert en bejelölt hyperlipidemia + mért HTG nagyobb, mint 1.7 pirossal)	0: no 1: yes			1:under 18.5 2:18.5-24.99 3:25-29.99 4:30.00 and above			0: no 1: yes	0: no 1: yes	year	1: no data, numeric	999999:no data, 1: male 2: female	999999: no data 1: mild 2: moderate 3: severe	999999: no data 0: no 1: yes	999999: no data, numeric	999999: no data 0: no 1: yes	999999: no data 0: no 1: yes
REGISTRY PARAMETERS													PERSONAL		OUTCOME			COMPLICATIO	
Registry_code	Obesity	Hypertension	Hyperlipidemia	Diabetes	Metś factor combinations	Number of factors	BMI categories 4	Single AP, RAP, CP	Institute	Import	Only_for_epidemiology_and_genetic_analysis	Year_of_admission	Age_at_the_time_of_admission	Gender	Severity	Mortality	Length_of_hospitalization_days	Local_pancreatic_complications	Fluid_collection
1914	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	41	1	2	0	17	1	1
1913	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	75	1	1	0	6	0	0
1891	1	1	0	1	HT+OB+DM	3 factors	4	CP	Hu, Debrec	0	0	2017	82	2	3	0	21	1	1
1890	1	1	1	0	HT+OB+HL	3 factors	4	single AP	Hu, Debrec	0	0	2016	48	2	3	1	7	1	1
1888	0	0	0	0	No MS factors	No factors	3	single AP	Hu, Pécs, F	0	0	2017	52	2	1	0	12	0	0
1887	0	0	0	0	No MS factors	No factors	3	single AP	Hu, Pécs, F	0	0	2017	29	2	1	0	5	0	0
1882	0	0	0	0	No MS factors	No factors	1	single AP	Hu, Debrec	0	0	2017	20	1	1	0	4	0	0
1881	1	0	0	0	OB	1 factor	4	single AP	Hu, Debrec	0	0	2017	87	2	1	0	8	0	0
1880	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	45	1	1	0	9	0	0
1879	0	0	0	0	No MS factors	No factors	2	single AP	Hu, Debrec	0	0	2017	58	1	1	0	8	1	0
1868	1	0	0	0	OB	1 factor	4	single AP	Hu, Debrec	0	0	2017	21	2	1	0	8	0	0

Always keep the original database UNCHANGED!!!

Technical controlling:

- All needed parameters are in the database in the appropriate format?
- Source of errors if any? Missing values? Text information?

Researcher controlling:

- Are all requested parameters included?
- Are there any extreme or life-incompatible values?

DATA ANALYSIS – FORMING GROUPS – EXAMPLES

Total cohort	Obesity (n=1257)		Hypertension (n=1127)		Hyperlipidemia (n=1036)		Diabetes (n=1257)	
	NON-OB	OB	NON-HT	HT	NON-HL	HL	NON-DM	DM
1257	886 70.5%	371 29.5%	451 40.0%	676 60.0%	687 66.3%	349 33.7%	1051 83.6%	206 16.4%

	GROUPS		n	%
1	noAB		122	12.7%
2	noAB-suspINF		122	12.7%
	noAB		244	25.4%
3	prevAB		120	12.5%
4a	AB-noBACT	no bact culture	420	43.7%
4b		neg bact culture	102	10.6%
5	AB-pozBACT		76	7.9%
	AB		718	74.6%
			962	100%

Should be considered:

- representativeness of the total analysed population vs. total cohort
- group sizes
- availability, quality of data in the groups

DATA ANALYSIS

1. Clinical question(s)
2. Forming groups
3. Representativeness
4. Data availability, data quality
5. Analysis of main outcomes in the groups
6. Hypothesis, statistical analysis for visible differences
7. Discussion on the findings
8. Decision on additional analyses
9. Hypotheses, detailed statistical analyses

These two will be explaned in the next presentation by our STATISTICIAN

COMMON MISTAKES

- Looking for **causality** or **therapy, intervention differences**.
- **Not focusing** on the main question (analysing everything ☺).
- **Not evaluating** appropriately biases, limitations, data availability.

TAKE HOME MESSAGE

- Specify the **clinical question**, **aims** and **variables** appropriately!
- **Check** your database!
- Form your **groups** properly!
- Examine **representativeness, data availability** and **quality**!
- Always keep the **original database** unchanged!

Thank you for your attention!

www.tm-centre.org

Data analysis: Statistics

Lilla Hanák
Centre for Translational Medicine

2nd October, 2019

University of Pécs
Pécs

Now what?

After data extraction…

WHAT TO DO NEXT?

FORM GOUPS

BUT HOW?

UNDERSTAND YOUR RESULTS

WORK TOGETHER WITH A STATISTICIAN

FORMULATE YOUR HYPOTHESES

Process after data extraction

1. Form **groups** for the analysis (if necessary).
2. Take a look at the data set and data quality. Make tables and figures about **primary and secondary outcomes**.
3. Formulate your **hypotheses**.

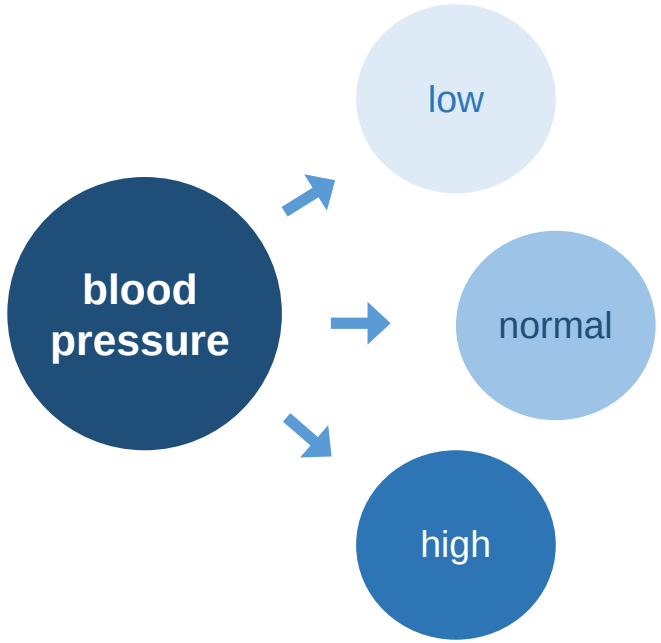
---

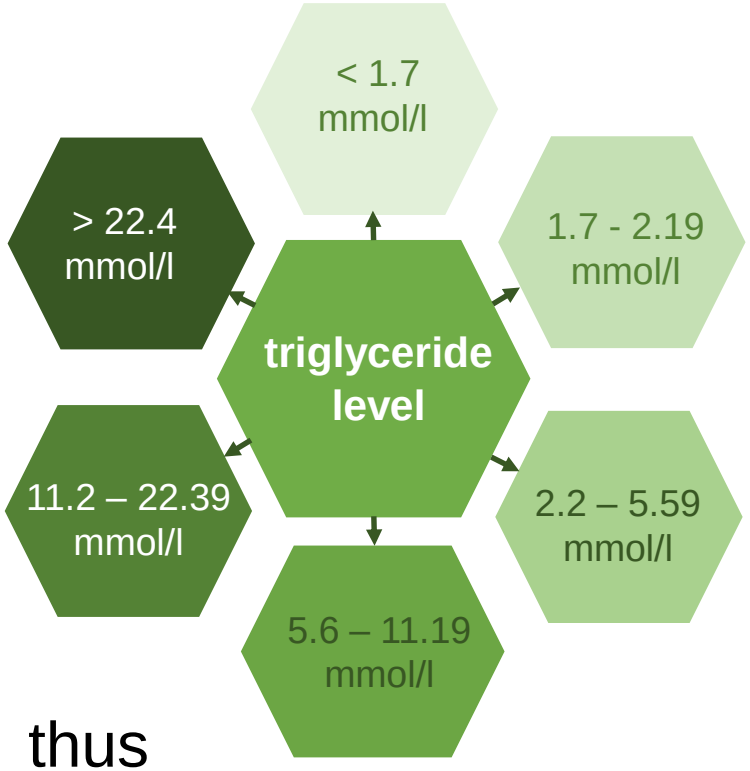
4. Perform **analysis**.
5. Understand your results. Make **further considerations** and statements. Form new hypotheses if necessary.
6. Conduct further statistical **analysis**.
7. Understand your results, start **writing your paper**.

This is the point from where a statistician SHOULD BE INVOLVED!

Opportunity to form groups

Forming groups - according to biological/medical considerations.





Grouped data has been 'classified' and thus some level of data analysis has taken place which means that the data is no longer raw → **always keep original values**!

Hypothesis

Hypothesis

an assumption about certain characteristics of a population

↓

Null Hypothesis

there is **no effect** or **no relationship** between phenomena or populations

Alternate Hypothesis

observations are influenced by a non-random factor

Developing a hypothesis

ASK A QUESTION

The question should be focused, specific and researchable.

TAKE A LOOK AT YOUR PARAMETERS

Look for parameters according to your question.

FORMULATE YOUR HYPOTHESIS

Write your initial answer to the question in a clear sentence.

REFINE YOUR HYPOTHESIS

The hypothesis should contain: the relevant variable(s), the specific group(s), the predicted outcome.

Examples of good hypothesis

- COPD patients have higher blood pressure than the recommended value of the average population.
- Hypertension is more frequent in patients with COPD than in those without COPD.
- Hypertension predicts 5-y mortality in COPD with high accuracy.
- Hypertension is an independent predictor of 5-y mortality in COPD.

Statistical analysis – practical aspects

Types of statistical tests

Comparison of central tendencies (e.g. means)

looks for the difference between the means of variables

Correlational

looks for an association between variables

Regression

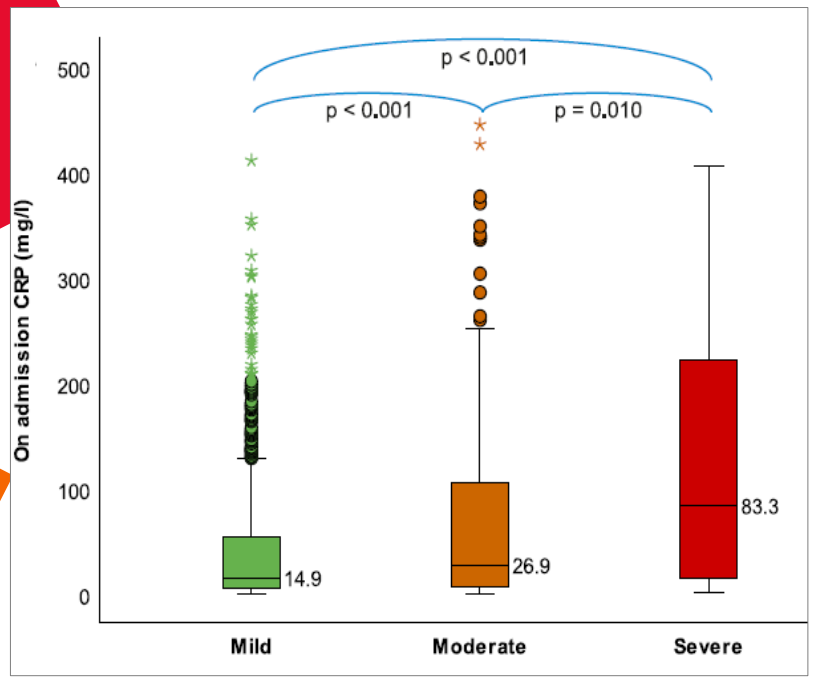
assess if change in one variable predicts change in another variable

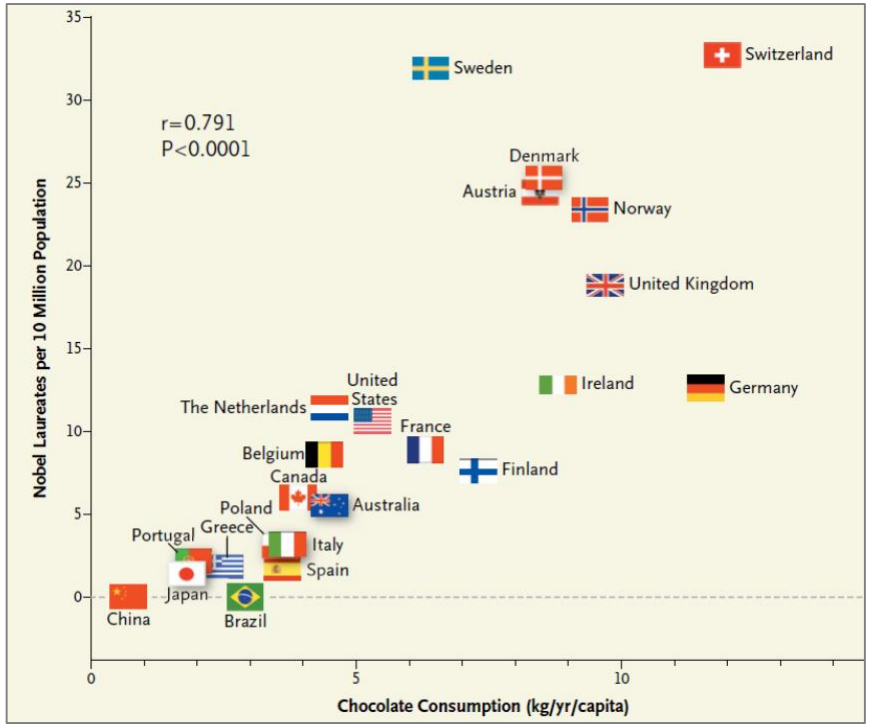
Statistical analysis – practical aspects

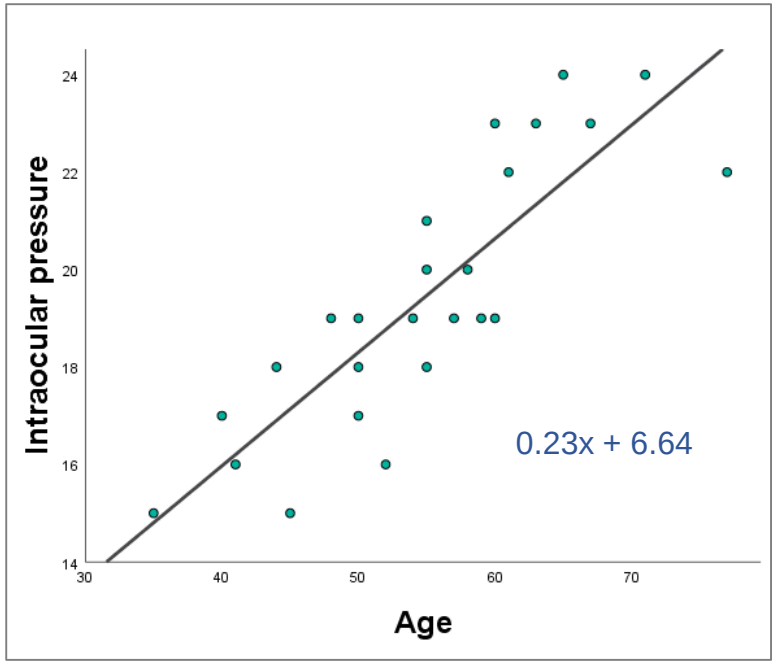
Comparison of means

Correlational

Regression







TAKE HOME MESSAGE

- A **good hypothesis** should contain the followings: the relevant variable(s), the specific group(s) and the predicted outcome!
- Researcher should keep in mind the **types of statistical tests** when formulating hypotheses!
- If any question arises regarding the data set, hypotheses or analysis always **consult with a statistician!**

Thank you for your attention!

Lilla Hanák

biostatistics@tm-centre.org

Publication strategy

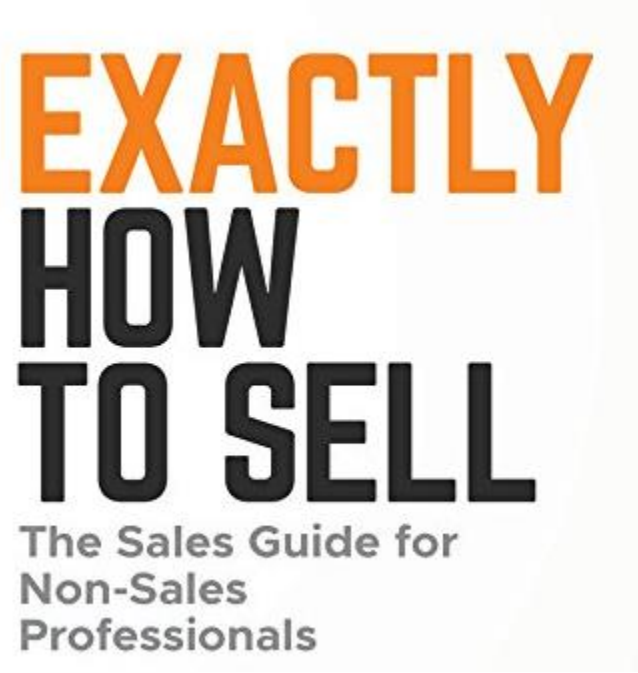
Péter Hegyi

Centre for Translational Medicine

p.hegyi@tm-centre.org

2nd October, 2019

University of Pécs
Pécs


EXACTLY
HOW
TO SELL
The Sales Guide for
Non-Sales
Professionals

Q1 WHAT ARE THE ELEMENTS OF A PUBLICATION

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULT
DISCUSSION
CONCLUSION

Q2 WHICH ORDER SHOULD I START?

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

CONCLUSIONS

- the **most usable** ones in practice
- no more than **two or three** points
- highlight the **importance**
- Point the the **future**

THIS IS THE FINAL CLAIM!

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

METHODS

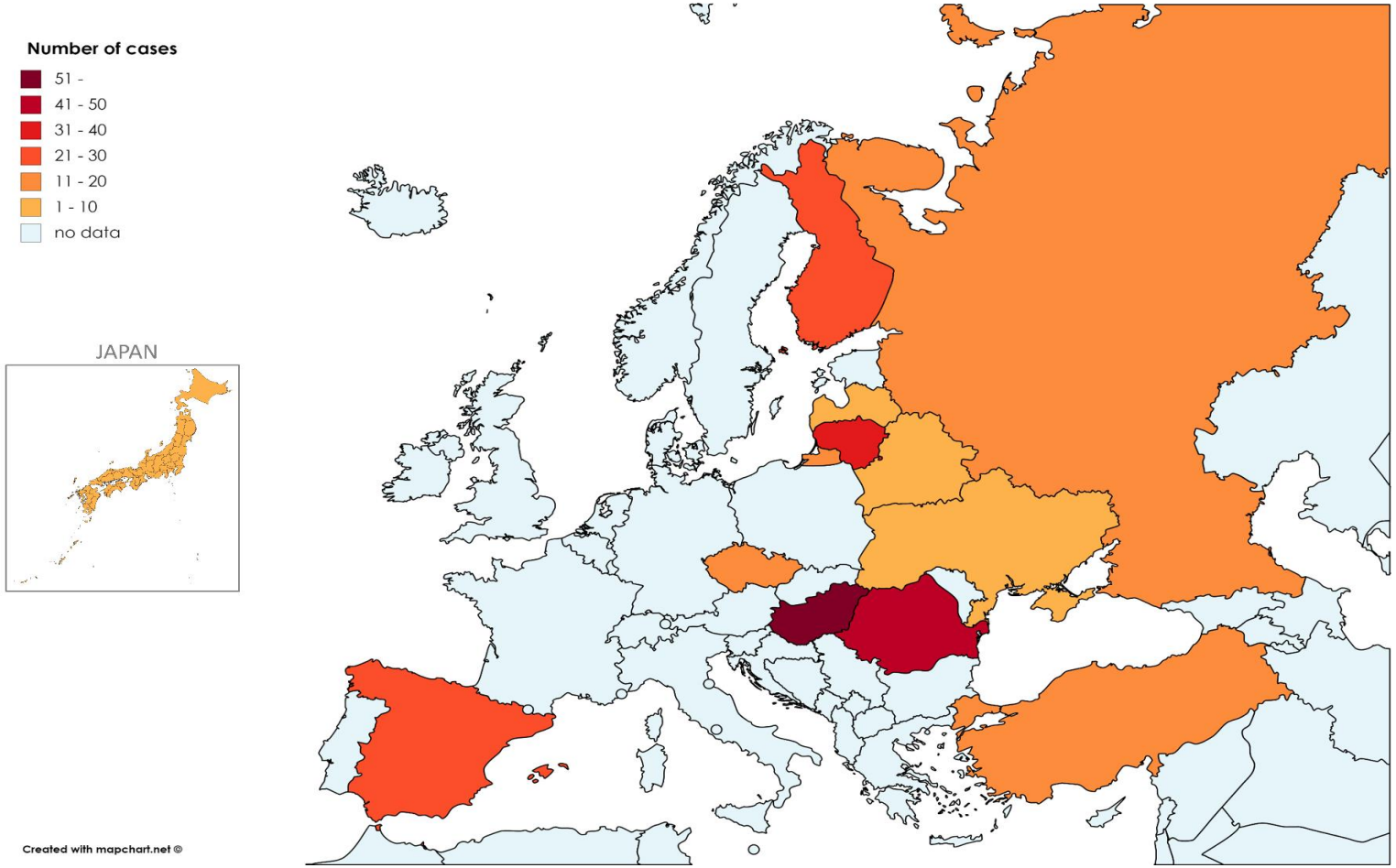
- Only a **summary** of the method
- All details can go to the **supplementary materials**

RESULTS

- **point by point** (like in the guidelines)
- put them in a logical order (**make a story**)
- put only the **undisposable** ones into the main text (**must have**)
- put into the section which **justify your conclusion**
- put every other figures to the supplementary part (**nice to have**)
- **connect** them
- highlight the **new discoveries**, **make a table**
- you can **change the order at any time**

WHERE WERE YOUR DATA COLLECTED?

DISTRIBUTION OF CASES



WHAT IS THE QUALITY OF OUR THE DATA?

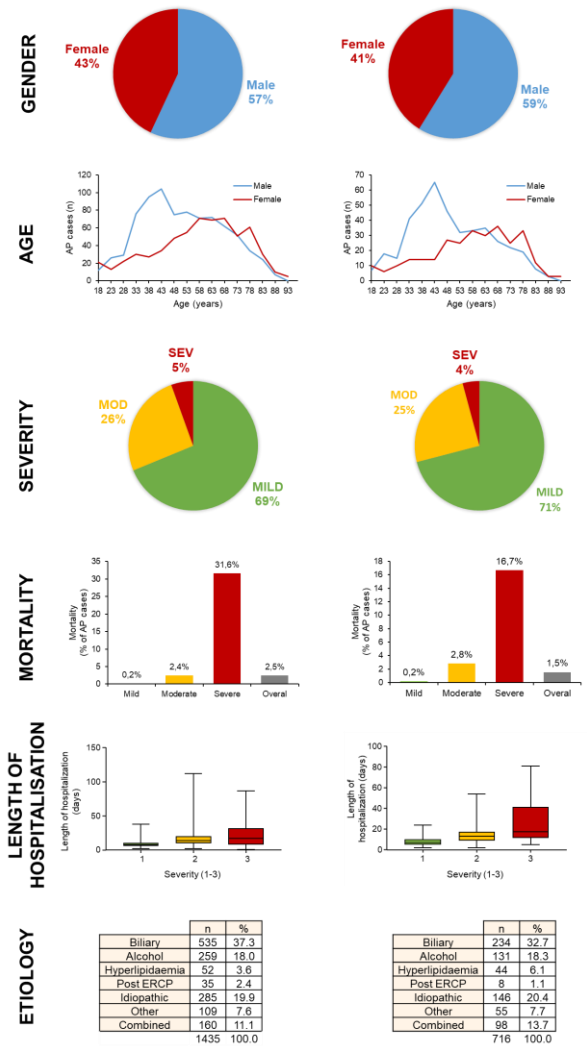
TRANSLATIONAL MEDICINE

taking discoveries for patients benefits

TM TRANSLATIONAL MEDICINE

DATA QUALITY OF INVESTIGATED PARAMETERS

parameter	overall	uploaded data	%
Age at the time of admission	1435	1435	100.0%
Gender	1435	1435	100.0%
Severity	1435	1435	100.0%
Mortality	1435	1435	100.0%
LOH	1435	1435	100.0%
Abdominal pain	1435	1432	99.8%
Abdominal pain length before admission	1435	1202	83.8%
Ad Antibiotic therapy	1435	1291	90.0%
Ad White blood cell (WBC) count (G/l)	1435	1288	89.8%
D1 White blood cell (WBC) count (G/l)	1435	865	60.3%
D2 White blood cell (WBC) count (G/l)	1435	746	52.0%
D3 White blood cell (WBC) count (G/l)	1435	657	45.8%
D4 White blood cell (WBC) count (G/l)	1435	518	36.1%
D5 White blood cell (WBC) count (G/l)	1435	429	29.9%
D6 White blood cell (WBC) count (G/l)	1435	374	26.1%
D7 White blood cell (WBC) count (G/l)	1435	338	23.6%
Ad C-reactive protein (mg/l)	1435	1177	82.0%
D1 C-reactive protein (mg/l)	1435	775	54.0%
D2 C-reactive protein (mg/l)	1435	674	47.0%
D3 C-reactive protein (mg/l)	1435	640	44.6%
D4 C-reactive protein (mg/l)	1435	520	36.2%
D5 C-reactive protein (mg/l)	1435	422	29.4%
D6 C-reactive protein (mg/l)	1435	365	25.4%
D7 C-reactive protein (mg/l)	1435	316	22.0%
TOTAL	34440	21204	61.6%

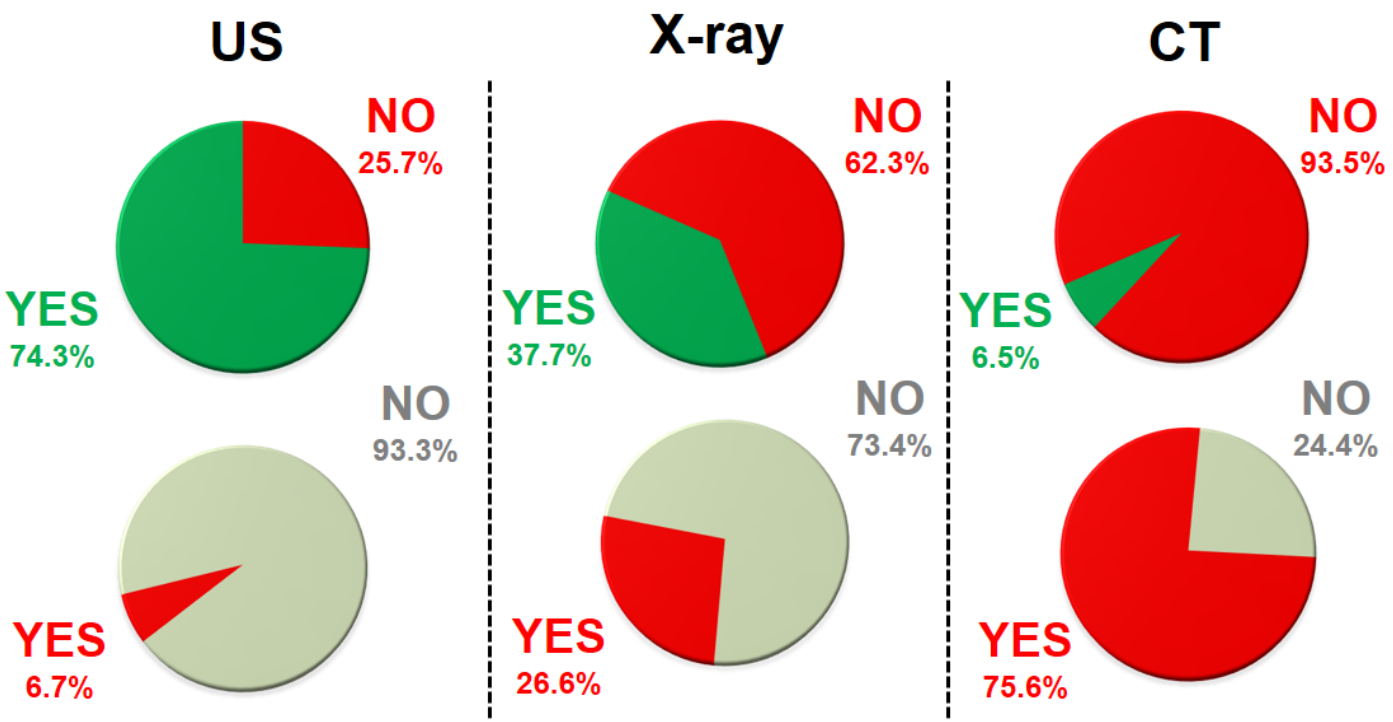


IT MUST BE DETERMINED **WHAT** YOUR STUDY POPULATION REPRESENTS

DATA **INTERPRETATION** STRONGLY **DEPENDS** ON YOUR POPULATION

WHAT CONCLUSION CAN WE MAKE?

THE INCIDENCE RATE OF **PLEURAL FLUID** IN ACUTE PANCREATITIS



Severity and mortality with (yes) or without (no) pleural complications

	MILD	MOD	SEV	MORT
YES	39.1%	47.8%	13.0%	33.0%
NO	63.0%	28.9%	8.1%	0

	MILD	MOD	SEV	MORT
YES	28.6%	41.1%	30.4%	58.8%
NO	64.3%	27.9%	7.8%	0

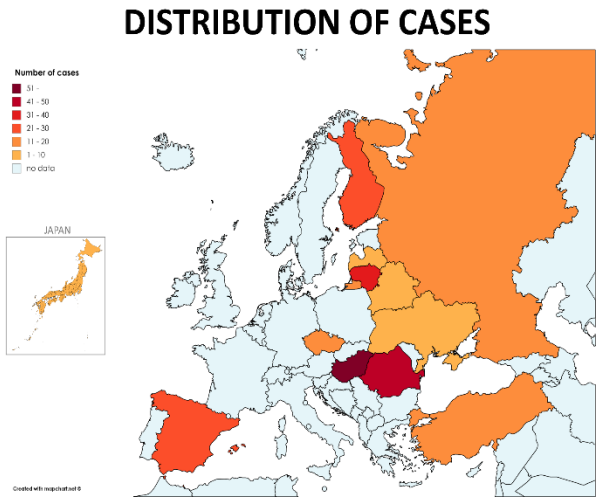
	MILD	MOD	SEV	MORT
YES	14.3%	61.7%	25.0%	43.0%
NO	33.3%	55.6%	11.1%	0

SAME COHORT
DIFFERENT METHODS
DIFFERENT RESULTS

BECAUSE OF THE DIFFERENCES BETWEEN THE STUDY POPULATION!

SUPPLEMENTARY FIGURES

CENTRES



SFig1

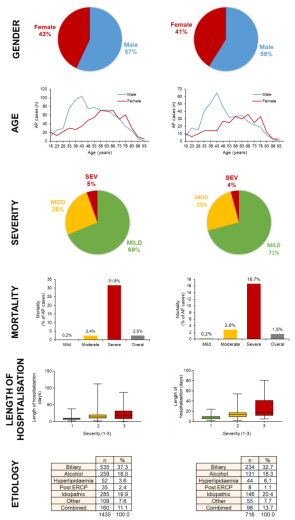
QUALITY

DATA QUALITY OF INVESTIGATED PARAMETERS

parameter	overall	uploaded data	%
Age at the time of admission	1435	1435	100.0%
Gender	1435	1435	100.0%
Severity	1435	1435	100.0%
Mortality	1435	1435	100.0%
LOH	1435	1435	100.0%
Abdominal pain	1435	1432	99.8%
Abdominal pain length before admission	1435	1202	83.8%
Ad Antibiotic therapy	1435	1291	90.0%
Ad White blood cell (WBC) count (G/l)	1435	1288	89.8%
D1 White blood cell (WBC) count (G/l)	1435	865	60.3%
D2 White blood cell (WBC) count (G/l)	1435	746	52.0%
D3 White blood cell (WBC) count (G/l)	1435	657	45.8%
D4 White blood cell (WBC) count (G/l)	1435	518	36.1%
D5 White blood cell (WBC) count (G/l)	1435	429	29.9%
D6 White blood cell (WBC) count (G/l)	1435	374	26.1%
D7 White blood cell (WBC) count (G/l)	1435	338	23.6%
Ad C-reactive protein (mg/l)	1435	1177	82.0%
D1 C-reactive protein (mg/l)	1435	775	54.0%
D2 C-reactive protein (mg/l)	1435	674	47.0%
D3 C-reactive protein (mg/l)	1435	640	44.6%
D4 C-reactive protein (mg/l)	1435	520	36.2%
D5 C-reactive protein (mg/l)	1435	422	29.4%
D6 C-reactive protein (mg/l)	1435	365	25.4%
D7 C-reactive protein (mg/l)	1435	316	22.0%
TOTAL	34440	21204	61.6%

SFig2

POPULATION



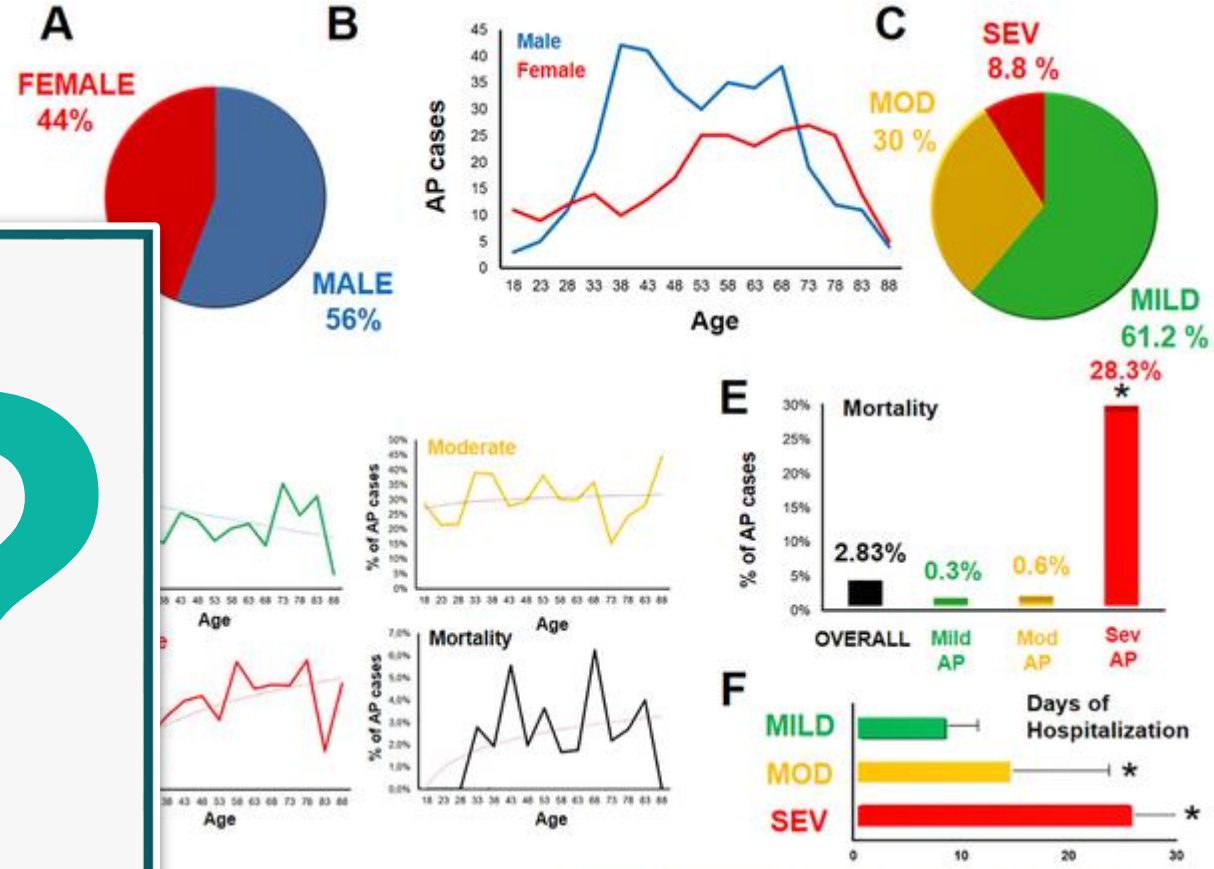
SFig3

THE STYLE OF PUBLICATION

MAJOR FIGURES

Factors affecting the LTBI treatment recommendation by a physician.

	Treatment offered (N = 302)	Visiting the clinic (N = 609) Bivariate analysis	Multiva
Age (year), median, range	42 (21–62)	0.94 (0.92–0.95)	0.975 (
Gender, female No (%)	210 (69.5)	1.69 (1.21–2.37)	1.00 (
BMI (kg/m^2), median (IQR)	22 (20.8–24.7)	0.92 (0.84–0.97)	0.98 (
Never smoker[a]	228 (84.1)	2.05 (1.36–3.09)	1.12 (
HTN	18 (6.0)	0.52 (0.28–0.95)	0.92 (
DM	12 (4.0)	1.37 (0.56–3.30)	
Profession			
Administrative	38 (12.6)	reference	re
Technician[b]	23 (7.6)	1.79 (0.93–3.47)	1.27 (
Health aid[c]	97 (32.1)	2.28 (1.42–3.64)	1.70 (
Physician	25 (8.3)	2.38 (1.19–4.53)	1.86 (0.90–3.85)
Nurse	119 (39.4)	5.23 (3.19–8.59)	3.43 (1.88–6.28)
Working duration, month, median (IQR)	237.4 (103.5–296.8)	0.99 (0.99–1.00)	
IFN-γ (TB Ag-Nil) concentration (IU/mL; median, IQR)	2.385 (0.878–5.865)	0.99 (0.97–1.02)	



	number	%	MALE	FEMALE	MILD	MOD	SEV
Biliary	263	43.83%	31.94%	**58.87%**[a]	64.26%	28.52%	7.22%
Alcohol	102	17.00%	**27.16%**[b]	4.15%	61.76%	30.39%	7.84%
Alcohol + High fat	57	9.50%	**11.94%**[c]	6.42%	57.89%	29.82%	12.28%
Idiopathic	98	16.33%	16.12%	16.600%	62.24%	28.57%	9.18%
Hyperlipidaemia	37	6.17%	**8.06%**[d]	3.77%	**32.43%**[f]	**48.65%**[g]	**18.92%**[h]
Post ERCP	22	3.67%	1.79%	**6.04%**[e]	68.18%	22.73%	9.09%
Other	21	3.50%	2.99%	4.15%	66.67%	28.57%	4.76%
	600	100,00%					

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

TITLE

- **Avoid**: „chatacterization…., effects of…, investigation of…
- The strongest **short** statement

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

DISCUSSION

- Discuss **all the relevant articles** which support or are against your results
- **AVOID:** repeating the result session
- **Do not describe** important knowledge which is **not relevant** to understand the study
- describe the **limitations**
- Highlight the **usefulness** of the result

TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

INTRODUCTION

- Two or three relevant points which **introduce the necessity** of the **work**
- **Do not describe** important knowledge which is **not relevant** to understand the study

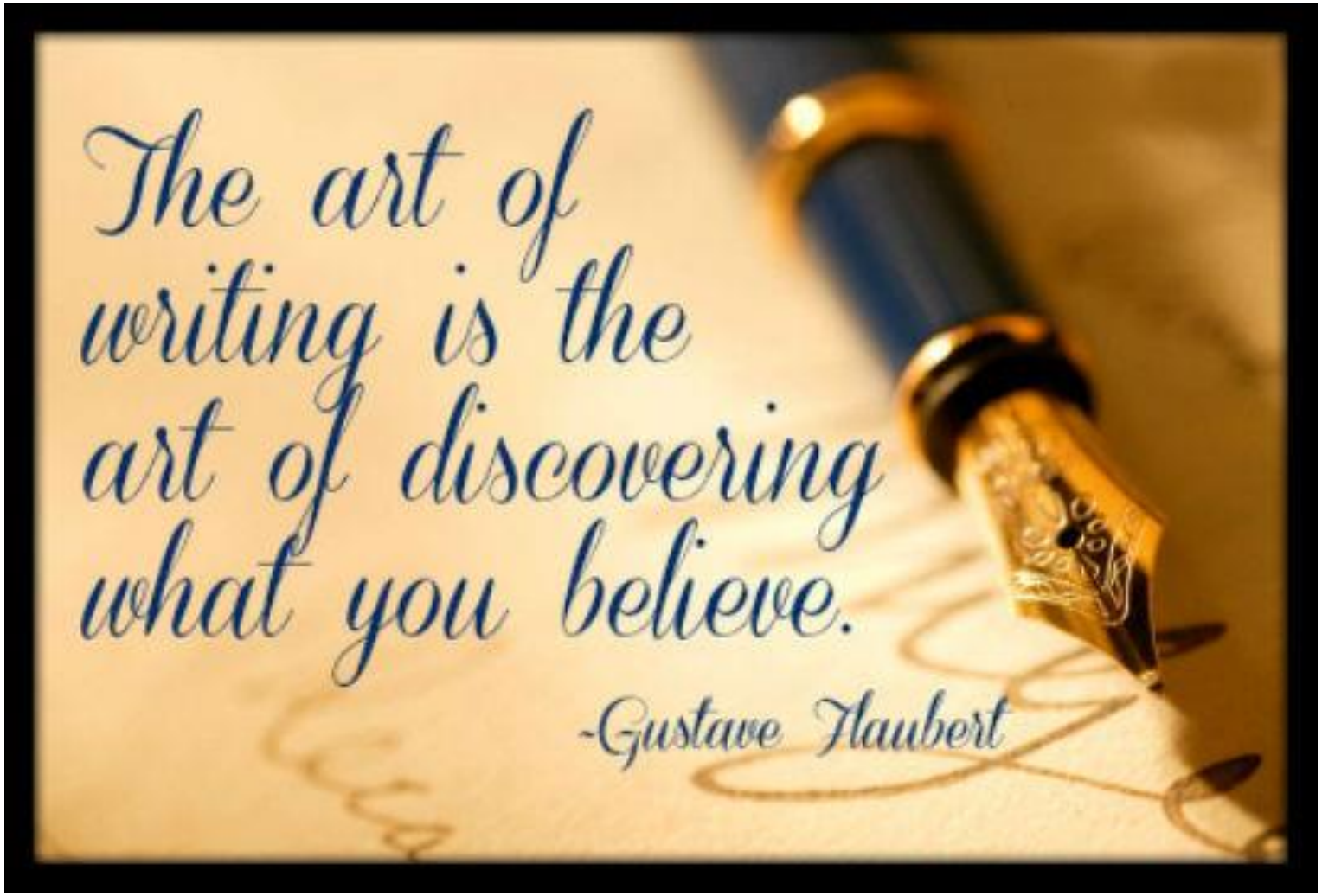
TITLE
ABSTRACT
INTRODUCTION
METHODS
RESULTS
DISCUSSION
CONCLUSIONS

ABSTRACT

- **SHORT**
- INFORMATIVE
- VERY MUCH DEPENDS ON THE JOURNAL STYLE

TRANSLATIONAL MEDICINE

taking discoveries for patients benefits



Thank you for your attention!

Péter Hegyi

p.hegyi@tm-centre.org

www.tm-centre.org

Establishing and Operating Registries Summary

Dalma Erdősi

2 October 2019, Pécs

Establishing a registry
Main points

1. Determining the purpose of the registry
2. International research
3. CRF
- Create
- Overview
- Final approval
4. Ethical approval

Main points

5. eCRF development
- Test version
- Live version
- User guide
6. Educate data managers
7. Organize patient involvement
8. Continuous involvement of national and international centers

Operating a registry
Main points

1. Data collection and upload
2. Quality control
3. Data extraction
   - Consultation : determining the data group
   - Extraction
   - Quality control
4. Statistical analysis
5. Publication

tm-centre.org

Thank you for yor attention!

TM TRANSLATIONAL MEDICINE Foundation

PRACTICE:
Interpretation of statistical analyses in publications from patient registries

Zsolt Szakács
Pécs, Hungary

3 Questions

Feedback presentation from 3 groups

1 Question each

1. Which prognostic factors did the study identify? Are they dependent or independent factors? Interpret the survival curves.

2. What does Fig 4 say?

3. What limitations does the study have? To which population are the findings representative?

Thank you for your participation!