FACULTY OF MEDICINE AND HEALTH SCIENCES

DEPARTMENT OF PUBLIC HEALTH AND PRIMARY CARE
UNIT OF MEDICAL INFORMATICS AND STATISTICS

HOW CAN WE ACHIEVE HIGH QUALITY DATA RELATED TO MEDICAL TECHNOLOGIES?

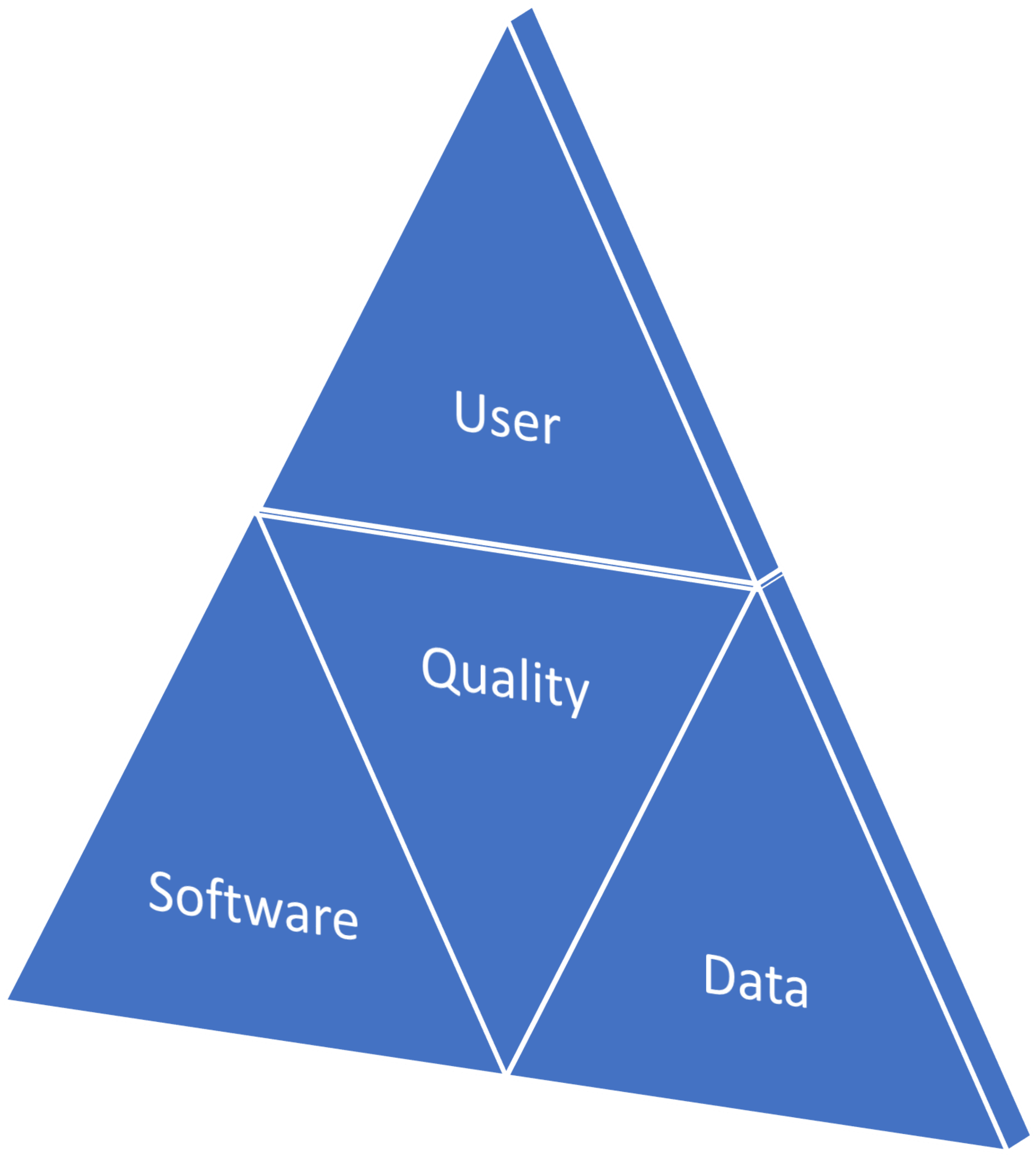
Prof. dr. Pascal Coorevits – beMedTech & i~HD Health Data Quality Seminar - May 17th 2022

GHENT UNIVERSITY

QUALITY & SAFETY GHENT
Expert centre for Quality and Safety in Primary Care

iHD The European Institute for Innovation through Health Data

be MedTech Medical Technologies Belgium


User
Quality
Software
Data

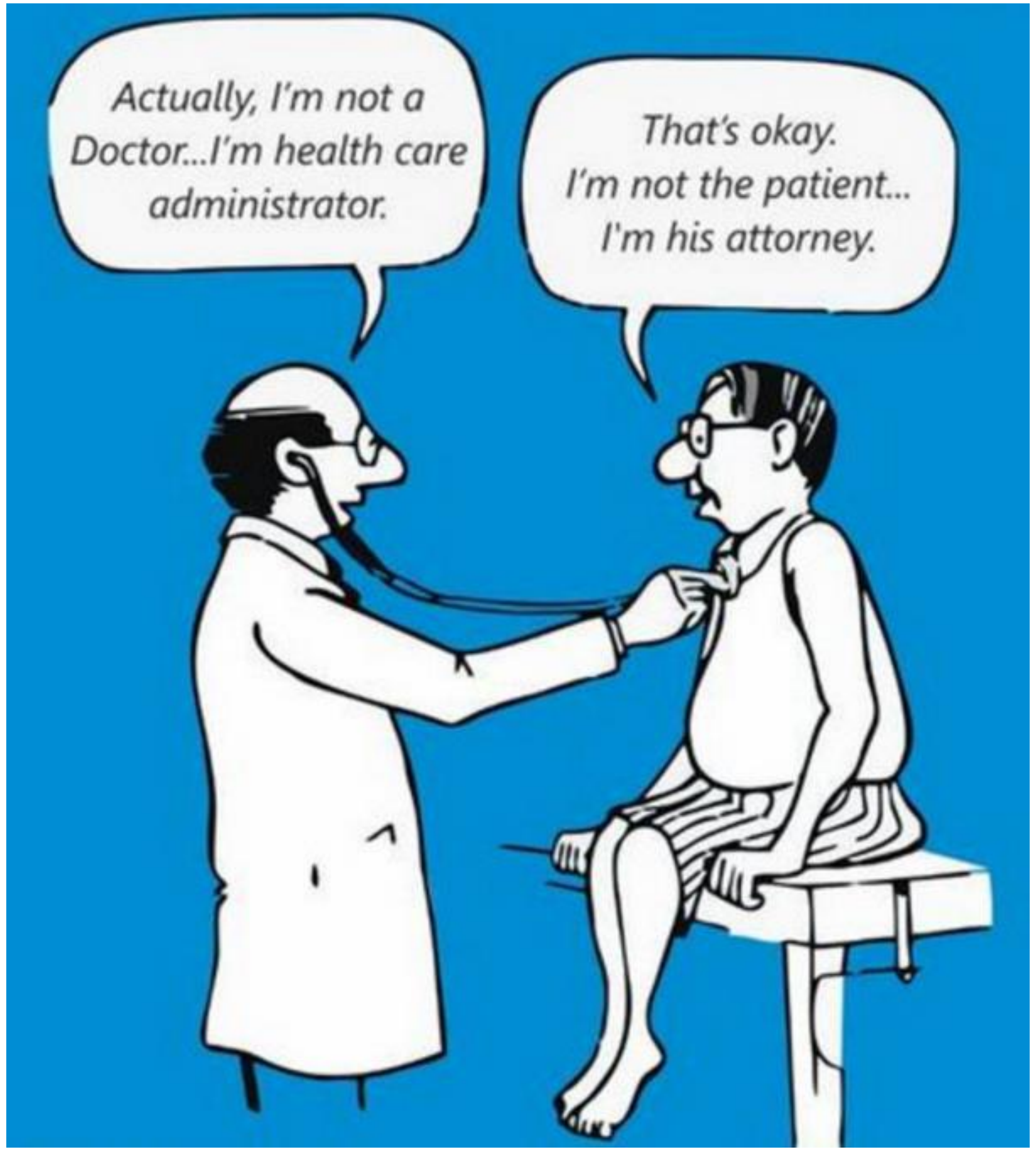
Quality of systems and data quality

- Depending on the use-case (e.g. providing daily routine care for patients, clinical research, big data analytics, …) a number of functionalities are needed (e.g. security, confidentiality, trustworthiness, …)

- Functions are required to ensure e.g. data correctness, completeness, accuracy, …

- Quality assurance is essential

- Quality labelling & certification are needed




QUALITY

What is data quality?

- A thing you think about briefly while you start data cleaning procedures after data has been collected ✗

- A complex scientific discipline ✓
  → Fit for use
  → Multi-dimensional
  → No consensus on exact definition, meaning and assessment methodology

Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research

Nicole Gray Weiskopf, Chunhua Weng

A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research

Michael G. Kahn, MD, PhD,*† Marsha A. Raebel, PharmD,‡§ Jason M. Glanz, PhD, MS,‡|| Karen Riedlinger, MPH, MT (ASCP),¶ and John F. Steiner, MD, MPH‡

A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data

Michael G. Kahn

Secondary Use of EHR: Data Quality Issues and Informatics Opportunities

Taxiarchis Botsis[a,b], Gunnar Hartvigsen[a,c], Fei Chen[b], Chunhua Weng[b]

A Data Quality Assessment Guideline for Electronic Health Record Data Reuse

Nicole G. Weiskopf, PhD;[i] Suzanne Bakken, RN, PhD;[ii,iii] George Hripcsak, MD, MS;[ii] Chunhua Weng, PhD[ii]

A practical framework for data management processes and their evaluation in population-based medical registries

M. SARIYAR[1], A. BORG[1], O. HEIDINGER[2] & K. POMMERENING[1]

RECEIVED 30 August 2015
REVISED 21 December 2015
ACCEPTED 17 January 2016

Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories

Carlos Sáez[1,2], Oscar Zurriaga[3,4,5], Jordi Pérez-Panadés[3], Inma Melchor[3], Montserrat Robles[1] and Juan M García-Gómez[1,6]

i~HD Data Quality Assessment

9 data quality dimensions

Designed for patient care, organisational learning and research

For health data providers, users and supporters

Developed by Data Quality Task Force

Based on scientific literature

Approved by 70 hospital representatives

i~HD
Hospital Network Workshop
February 9th, 2017
Brussels, Belgium
www.i-hd.eu
The European Institute For Innovation Thr~ugh Health Data
i~HD enriching knowledge and enhancing care through health data
i~HD Session
(Workshop 2)
"The reuse of EHRs for Learning Health Systems"
When: Friday 13.09.2019
Timing: 4 PM – 5:30 PM
Location: Meet & Greet Center in the Ghelamco Arena, Gent
Uncover the insights hidden in your EHRs
i~HD
PARIS, 5 - 6 MARCH 2018
A two-day workshop run by the i~HD Hospital Network of Excellence and Data Quality Task Force, in collaboration with
Gaining the Benefits of Improved Health Data Quality
ICHOM
EPF European Patients Forum
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
WIND - DSI AP-HP
i~HD Hospital Network of Excellence Data Quality Workshop
Towards better data quality in hospitals
Tuesday 23rd May 2017 - Wednesday 24th May 2017
Realising the value of health data~
showcasing practical examples from research and healthcare
REPORT
i~HD Annual Conference
Gothenburg, Sweden 19 - 20 November 2018
Realising the Value from Health Data ~
Improving Care and Research
JOINT EVENT
September 21-22, 2017
MADRID, SPAIN
i+12
Instituto de Investigación Hospital 12 de Octubre
The European Institute For Innovation Thr~ugh Health Data
EMIF
KEEPING IN TOUCH
i~HD Annual Conference 2017
www.i-HD.eu/AC2017
i~HD @ VITALIS
21 - 23 May 2019
Gothenburg
Joining the Dots Conference - 27 & 28 November - Brussels
Trillium II
Digital Health Europe
i~HD
EHR2EDC
C3CLOUD
EHDEN
Exchanging use cases, insights, methods, successes, challenges
Quality and interoperability of high-value data sets
Promoting and sustaining federated learning health ecosystems
Patient empowerment in access and use of health data
Advancing solutions for multi morbidity
Calls to action for health ministries
Joint meeting with the eHealth Network on 28th November
Save the date - registration opens soon
i~HD@
LIFE SCIENCE LIVE
MAY 15 - 16, 2019
TURKU, FINLAND
i~HD @
CONGRESS '19
EAHM
BELGIUM - GHENT
11-14 September, 2019
Ghent, Belgium

Data Quality dimensions

Completeness
Data values are present

Consistency
Data satisfy constraints

Correctness
Values are true and unbiased

Uniqueness
Patient records are not duplicated

Stability
Data are comparable among sources and over time

Timeliness
Data is promptly processed and up-to-date

Contextualisation
Data are annotated with acquisition context

Trustworthiness
Data can be trusted based on owner’s reputation

Representative-ness
Data are representative of population

PatientID	Date-of-admission	Date-of-birth	Sex-at-birth	Weight-kg	Height-cm	Medication-Insulin	Glucose-blood-mg/dL	Glucose-blood-fasting	Prior MI
2310	05/03/2020	17/06/1976	F			0			0
10003	12/05/2020	24/11/1965	M	77	181	1	86	Yes	1
10003	23/08/2020	24/11/1965	M	76	181	1	91	Yes	1
10003	02/10/2020	24/11/1965	M	78	181	1	121	No	0
811	06/11/2020	01/26/1990	C	"55kg"	1.68	0	76	Yes	
345	23/01/2020	02/08/1939	F	53	162	0	134		
6786	11/08/2020	11/09/1946	M	68	177	0			1
6786	11/08/2020	11/09/1946	M	68	177	0			1
6786	11/08/2020	11/09/1946	M	83	177	0	95	Yes	1
10009	16/05/2020	24/02/1953	F	93	165	1	221	No	

Incomplete Inconsistent Duplicate Incorrect

CORRECTNESS – “VALUES ARE TRUE AND UNBIASED”

Osteoporos Int
DOI 10.1007/s00198-016-3635-2

ORIGINAL ARTICLE

Clinical height measurements are unreliable: a call for improvement

A. L. Mikula[1] · S. J. Hetzel[2] · N. Binkley[3] · P. A. Anderson[4]

“Fifty percent of clinic staff reported they on occassion enter patient reported height into the EHR rather than performing a measurement”

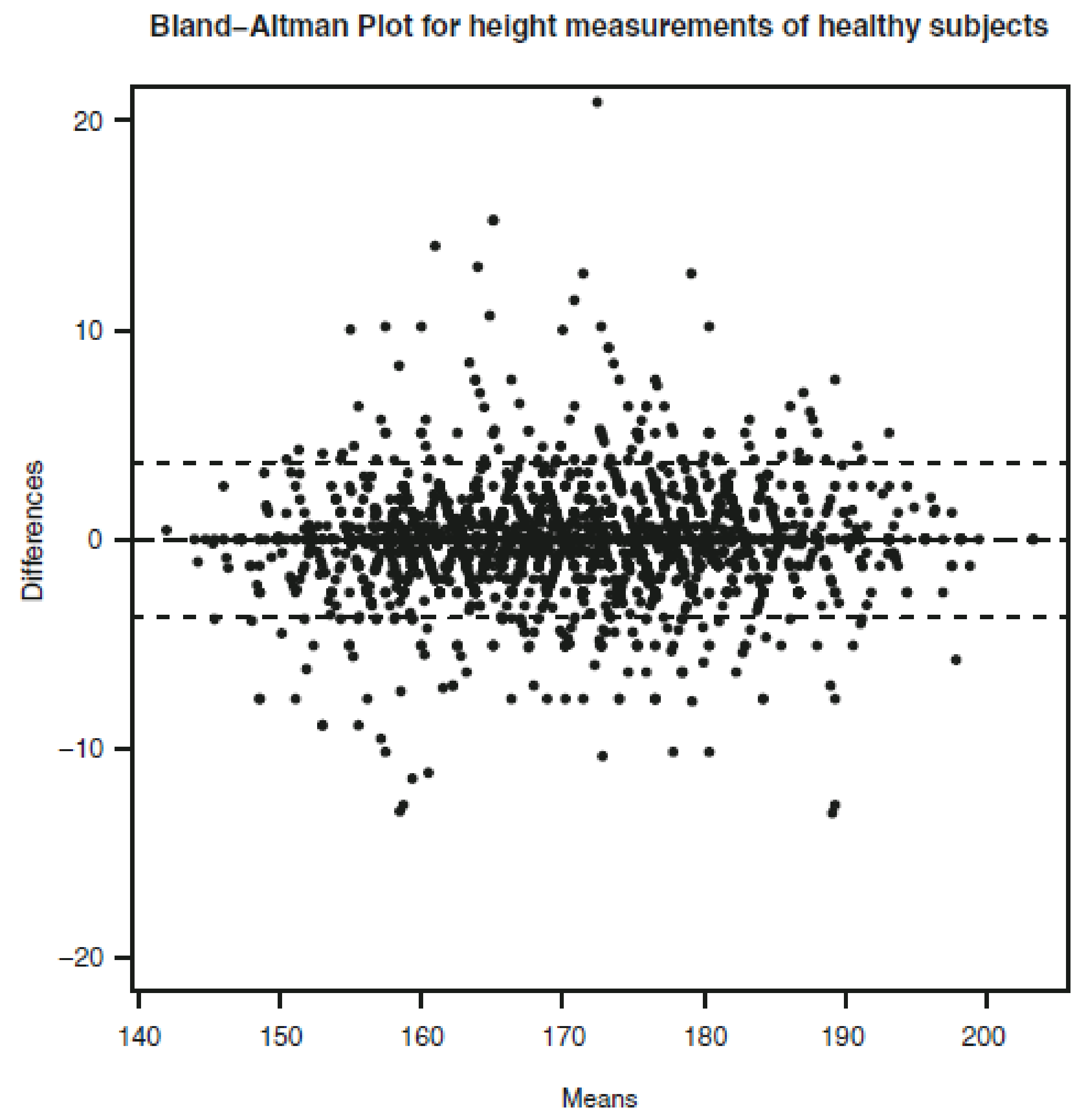


Fig. 4 Bland-Altman plot for height measurements of subjects. Each data point represents a single patient. *X axis* represent mean patient height in centimeters. *Y axis* represents difference between the first and the last height measurement for the individual patients in centimeters. The *dotted lines* represent 95 % CI

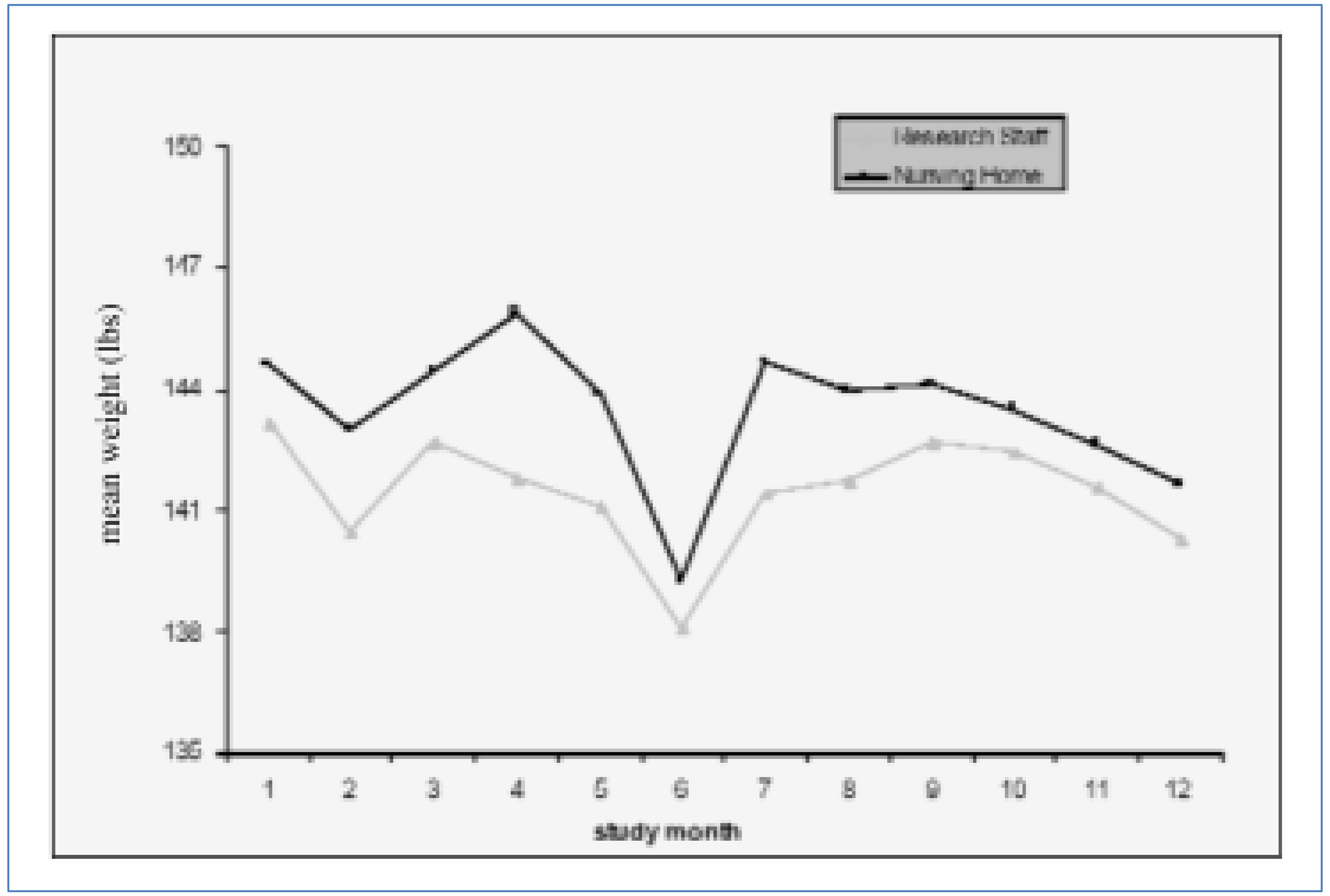
Bron: Carlos Saez and Juan M Garcia Gomez, UPV

CORRECTNESS – "VALUES ARE TRUE AND UNBIASED"

Published in final edited form as:
J Nutr Health Aging. 2009 March ; 13(3): 284–288.

THE ACCURACY OF MONTHLY WEIGHT ASSESSMENTS IN NURSING HOMES: IMPLICATIONS FOR THE IDENTIFICATION OF WEIGHT LOSS

S.F. SIMMONS[1,2], **E.N. PETERSON**[1], and **C. YOU**[1]
[1]Vanderbilt University, School of Medicine, Division of General Internal Medicine and Public Health, Center for Quality Aging, Nashville, TN
[2]Veterans Administration Tennessee Valley Healthcare System; VA Geriatric Research, Education, and Clinical Center, Nashville, TN



Comparison of weight measurements between:
- Trained research staff -> Gold standard
- Nursing homes staff -> Biased

Bron: Carlos Saez and Juan M Garcia Gomez, UPV

COMPLETENESS – "DATA VALUES ARE PRESENT"

- Incompleteness in clinical trials (EHR4CR standardized data inventory)

Data Group	Data Item	Avg. usage	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 9	Site 9
Demographics	**Gender**	100%	100,00%	100,00%	100,00%	100,00%	100,00%	100%	100%	100,0%	100,00%
Demographics	**Case Status**	96%	99,87%	100,00%	60,00%	100,00%	100,00%	100,00%	100%	100,0%	100,00%
Demographics	**Date of Birth**	89%	100,00%	100,00%	99,00%	NA	100,00%	100%	100%	100,0%	100,00%
Demographics	**Admission date**	84%	100,00%	100,00%	100,00%	NA	100,00%	99,53%	58%	100,0%	100,00%
Diagnosis	**Diagnosis Text**	81%	50,46%	84,02%	100,00%	100,00%	98,05%	100,00%	14%	100,0%	80,98%
Diagnosis	**Diagnosis Code**	81%	50,46%	84,02%	100,00%	100,00%	98,05%	100,00%	14%	100,0%	80,98%
Demographics	**Discharge date**	75%	100,00%	100,00%	100,00%	NA	100,00%	100,00%	58%	100,0%	14,18%
Diagnosis	**Diagnosis Date**	70%	50,46%	84,02%	100,00%	100,00%	100,00%	NA	13%	100,0%	80,98%
Medication	**Dosage**	25%	20,36%	0,00%	NA	NA	94,43%	95%	NA	NA	12,21%
Findings	**Weight**	25%	29,56%	18,24%	NA	NA	89,17%	27,20%	36%	7,5%	13,82%
Laboratory Findings	**Platelets Blood**	48%	52,78%	33,14%	63,73%	NA	100,00%	100%	45%	NA	33,88%
Laboratory Findings	**SGPT (ALT) in serum**	47%	33,61%	22,29%	100,00%	NA	100,00%	100%	47%	NA	21,86%
Laboratory Findings	**Total Protein in serum**	46%	52,37%	14,96%	86,53%	NA	100,00%	100%	47%	NA	16,34%
Laboratory Findings	**Total Bilirubin in serum**	46%	33,03%	16,99%	100,00%	NA	100,00%	100%	47%	NA	19,58%

Variables such as Weight are quite frequently not present

Doods *et al. Trials* 2014, 15:18
http://www.trialsjournal.com/content/15/1/18

TRIALS

RESEARCH **Open Access**

A European inventory of common electronic health record data elements for clinical trial feasibility

Justin Doods[1], Florence Botteri[2], Martin Dugas[1], Fleur Fritz[1*] and on behalf of EHR4CR WP7

Bron: Carlos Saez and Juan M Garcia Gomez, UPV

- 20% of scientific papers with supplementary data in Excel contain errors
  - Cause: automatic conversion of gene symbols in dates and numbers by Excel

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT **Open Access**

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Oryza sativa* and *Saccharomyces cerevisiae* [2]. The regex search used was similar to that described previously by Zeeberg and colleagues [1], with the added screen for dates in other formats (e.g. DD/MM/YY and MM-DD-YY). To expedite analysis of supplementary files from multi-disciplinary journals, we limited the articles screened to those that have the keyword 'genome' in the title or abstract (*Science*,

CONSISTENCY – "DATA SATISFIES CONSTRAINTS"

Bron: Carlos Saez and Juan M Garcia Gomez, UPV

- Data quality issues found in a survival analysis of pancreatic cancer patients (Columbia University Medical Center, New York)

➢ Information inconsistency between different EHR data sources:
  - In a few cases, pancreatitis was **diagnosed as being chronic in the pathology reports** but it was **reported as being only acute in the clinical notes**

➢ Information inconsistency within the same data sources :
  - Some patients received **simultaneously two different ICD-9-CM codes** for their diagnoses of diabetes, both **250.01 and 250.02 for type-1 and type-2 respectively**

CONSISTENCY – "DATA SATISFIES CONSTRAINTS"

Summit on Translat Bioinforma. 2010; 2010: 1–5. PMCID: PMC3041534
Published online 2010 Mar 1.

Secondary Use of EHR: Data Quality Issues and Informatics Opportunities

Taxiarchis Botsis,[a,b] Gunnar Hartvigsen,[a,c] Fei Chen,[b] and Chunhua Weng[b]

Bron: Carlos Saez and Juan M Garcia Gomez, UPV

› JMIR Med Inform. 2021 Aug 4;9(8):e27842. doi: 10.2196/27842.

Quality of Hospital Electronic Health Record (EHR) Data Based on the International Consortium for Health Outcomes Measurement (ICHOM) in Heart Failure: Pilot Data Quality Assessment Study

Hannelore Aerts [1, 2], Dipak Kalra [1, 2], Carlos Sáez [3], Juan Manuel Ramírez-Anguita [4], Miguel-Angel Mayer [4], Juan M Garcia-Gomez [3], Marta Durà-Hernández [3], Geert Thienpont [2, 5], Pascal Coorevits [1]

Affiliations − collapse

Affiliations

1 Medical Informatics and Statistics Unit, Department of Public Health and Primary Care, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium.
2 The European Institute for Innovation through Health Data (i~HD), Ghent, Belgium.
3 Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Valencia, Spain.
4 Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona, Spain.
5 Research in Advanced Medical Informatics and Telematics (RAMIT), Ghent, Belgium.

PMID: 34346902 PMCID: PMC8374665 DOI: 10.2196/27842

FULL TEXT LINKS

JMIR Publications OPEN ACCESS

PMC FREE Full text

ACTIONS

Cite

Favorites

SHARE

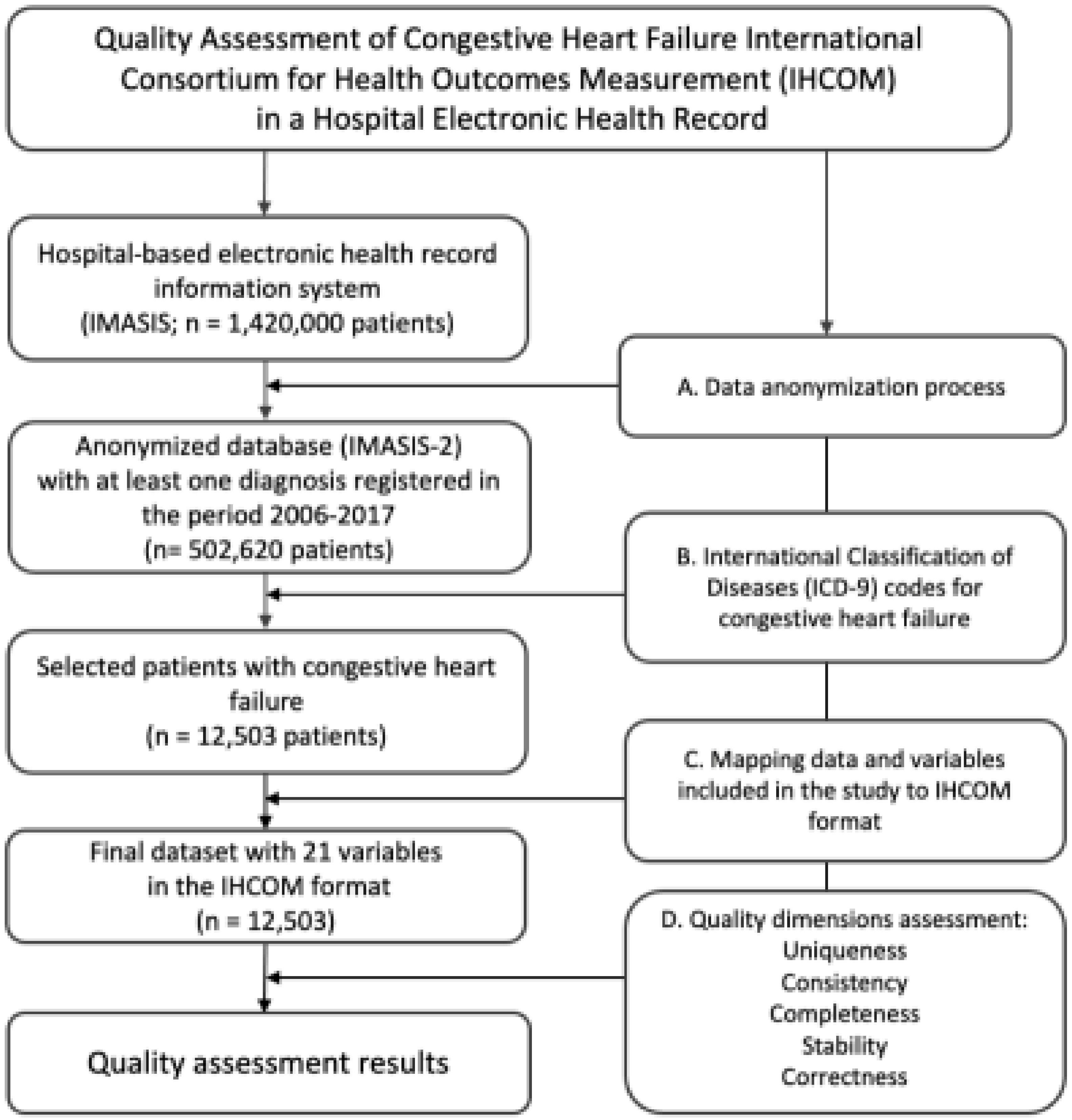
PAGE NAVIGATION

Title & authors

Abstract

Conflict of interest statement

Figures


Quality Assessment of Congestive Heart Failure International Consortium for Health Outcomes Measurement (IHCOM) in a Hospital Electronic Health Record
Hospital-based electronic health record information system (IMASIS; n = 1,420,000 patients)
A. Data anonymization process
Anonymized database (IMASIS-2) with at least one diagnosis registered in the period 2006-2017 (n= 502,620 patients)
B. International Classification of Diseases (ICD-9) codes for congestive heart failure
Selected patients with congestive heart failure (n = 12,503 patients)
C. Mapping data and variables included in the study to IHCOM format
Final dataset with 21 variables in the IHCOM format (n = 12,503)
D. Quality dimensions assessment: Uniqueness Consistency Completeness Stability Correctness
Quality assessment results

- It is essential to have **considerable knowledge of the EHR** (types of data available, how the data were collected or who collected it)

- The **assessment** of the data is the very first step to improve the quality of your data

- Once you know about the quality of your data, it is important to **monitor it regularly**

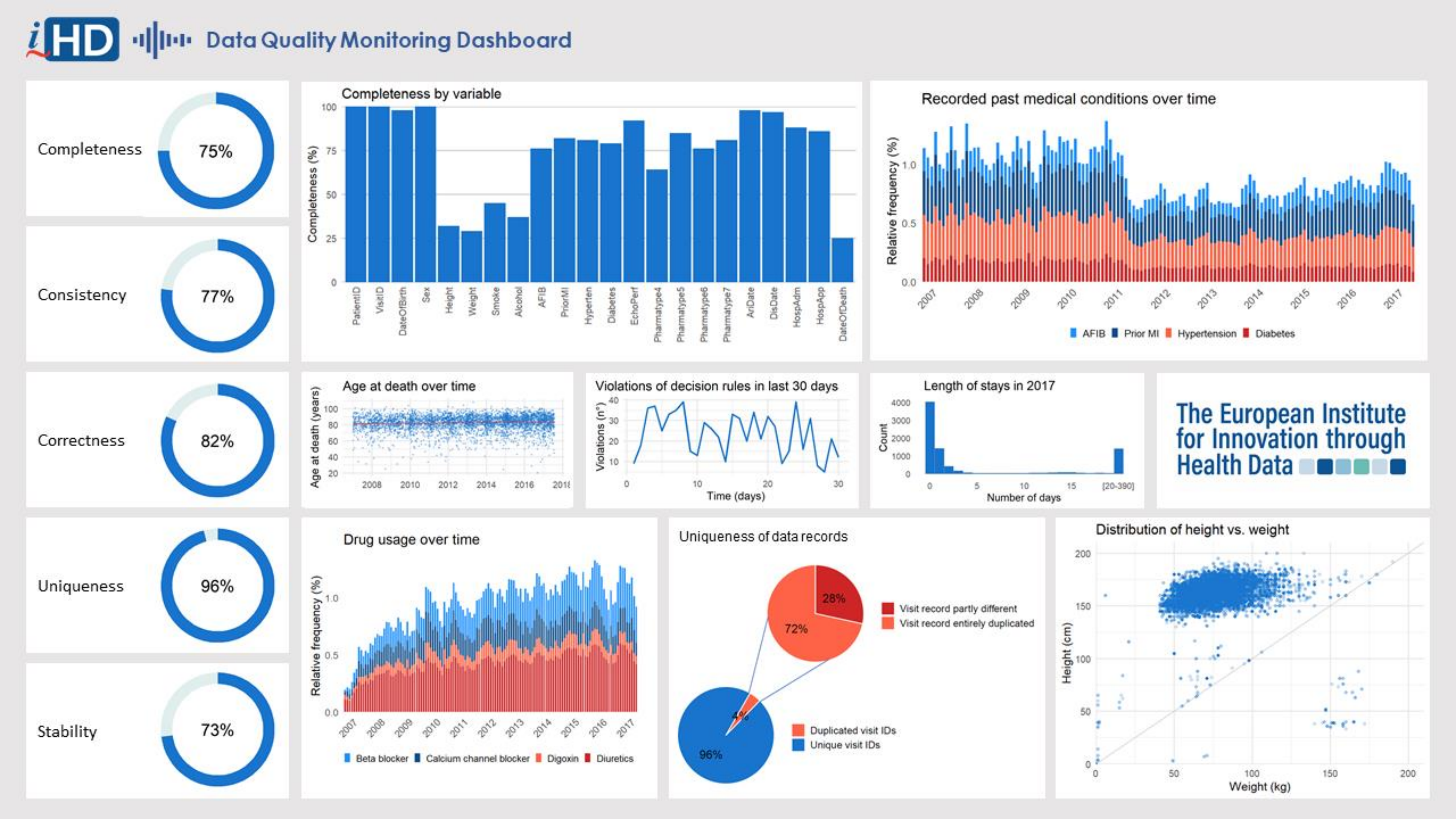
- It is of value that an **external assessment** of the data quality is performed by an independent organization

- High-quality data **enhance the validity and reliability** of study findings

- It is critical to ensure that **the metrics** are feasible, valid, and meaningful for a specific EHR and purpose and its quality improvement

- **Multidisciplinary** approach is highly recommended

- Thinking of using EHR for different purposes such as research, EHR data models would need to be **expanded and redesigned** and data quality assessment can assist in doing these tasks


iHD
Data Quality Monitoring Dashboard
Completeness 75%
Consistency 77%
Correctness 82%
Uniqueness 96%
Stability 73%
Completeness by variable
Completeness (%)
PatientID
VisitID
DateOfBirth
Sex
Height
Weight
Smoke
Alcohol
AFIB
PriorMI
Hyperten
Diabetes
EchoPerf
Pharmatype4
Pharmatype5
Pharmatype6
Pharmatype7
AriDate
DisDate
HospAdm
HospApp
DateOfDeath
Recorded past medical conditions over time
Relative frequency (%)
AFIB
Prior MI
Hypertension
Diabetes
Age at death over time
Age at death (years)
Violations of decision rules in last 30 days
Violations (n°)
Time (days)
Length of stays in 2017
Count
Number of days
[20-390]
The European Institute for Innovation through Health Data
Drug usage over time
Relative frequency (%)
Beta blocker
Calcium channel blocker
Digoxin
Diuretics
Uniqueness of data records
28%
72%
4%
96%
Visit record partly different
Visit record entirely duplicated
Duplicated visit IDs
Unique visit IDs
Distribution of height vs. weight
Height (cm)
Weight (kg)

Thank you!

pascal.coorevits@ugent.be

www.linkedin.com/pascalcoorevits