Knowledge Discovery from Clinical and Administrative Data

Pedro Pereira Rodrigues

CINTESIS & LIAAD – INESC TEC

Faculty of Medicine – University of Porto, Portugal

@ECMLPKDD – Nancy, France

Who am I?

A privileged one, who being educated in machine learning, gets to teach medical students on research methodology and data science ;-)

- MSc (2005) and PhD (2010) on clustering data streams and stream sources.
- Last 6 years involved in medical informatics, clinical research and medical education.

Coordinator of the **BioData - Biostatistics and Intelligent Data Analysis** group of **CINTESIS - Centre for Health Technologies and Services Research** (100+ PhD research unit to start officially in 2015) and collaborator in **LIAAD – INESC TEC** (original research unit since 2003).

Agenda II

- Resistance to KDD from health data
- Contextual anomalies in health data
- Admission Discharge Transfer (ADT) data
- Uncertainty in recorded ADT and clinical data
- Impact of uncertainty in health services research
- Toy and real-world examples of misconceptions
- Lessons learned

Common KDD practitioner scenario...

Common KDD practitioner scenario...

«Hurray, we've got access to a medical database!»

Common KDD practitioner scenario...

«Hurray, we've got access to a medical database!»

Apply KDD process, including state-of-the-art machine learning methods

Common KDD practitioner scenario...

«Hurray, we've got access to a medical database!»

Apply KDD process, including state-of-the-art machine learning methods

Validate models using established validation procedures (e.g. X-validation)

Common KDD practitioner scenario...

«Hurray, we've got access to a medical database!»

Apply KDD process, including state-of-the-art machine learning methods

Validate models using established validation procedures (e.g. X-validation)

Present promissing results to principal investigator owning the data (i.e. MD)

Common KDD practitioner scenario...

«Hurray, we've got access to a medical database!»

Apply KDD process, including state-of-the-art machine learning methods

Validate models using established validation procedures (e.g. X-validation)

Present promissing results to principal investigator owning the data (i.e. MD)

…

Common KDD practitioner scenario...

«Hurray, we've got access to a medical database!»

Apply KDD process, including state-of-the-art machine learning methods

Validate models using established validation procedures (e.g. X-validation)

Present promissing results to principal investigator owning the data (i.e. MD)

…

«Nice, but I shall not use it...»

:-(

Common KDD practitioner scenario...

Why?

There are mainly four arguments why physicians hesitate to use our models (i.e. outside traditional biostatistics):

- *«I cannot interpret your model in order to assess its validity.»*

«OK, I'll lose the neural networks and build decision trees or Bayesian nets.»

Common KDD practitioner scenario...

Why?

There are mainly four arguments why physicians hesitate to use our models (i.e. outside traditional biostatistics):

- *«I cannot interpret your model in order to assess its validity.»*
- *«There's no clear statistical support in your machine learning models.»*

«But I can show you that the Gini's impurity coefficient is known to be closely related to both, the AU-ROC and the Mann-Whitney-U test.»

D. Hand and R. Till, “A simple generalisation of the area under the ROC curve for multiple class classification problems,” Mach. Learn., vol. 45, pp. 171–186, 2001.

Common KDD practitioner scenario...

Why?

There are mainly four arguments why physicians hesitate to use our models (i.e. outside traditional biostatistics):

- *«I cannot interpret your model in order to assess its validity.»*
- *«There's no clear statistical support in your machine learning models.»*
- *«The data you have used was not collected for that purpose.»*

«But I have the protocol that generated data collection; I can understand it.»

Common KDD practitioner scenario...

Why?

There are mainly four arguments why physicians hesitate to use our models (i.e. outside traditional biostatistics):

- *«I cannot interpret your model in order to assess its validity.»*
- *«There's no clear statistical support in your machine learning models.»*
- *«The data you have used was not collected for that purpose.»*
- *«The data is, simply, wrong.»*

Common KDD practitioner scenario...

Why?

There are mainly four arguments why physicians hesitate to use our models (i.e. outside traditional biostatistics):

- *«I cannot interpret your model in order to assess its validity.»*
- *«There's no clear statistical support in your machine learning models.»*
- *«The data you have used was not collected for that purpose.»*
- *«The data is, simply, wrong.»*

Err...

Problems in health data

"If I had only one hour to save the world, I would spend fifty-five minutes defining the problem, and only five minutes finding the solution."

Albert Einstein

Anomalies in health data?

There's an entire community devoted to data quality issues in health data...

…one should take into account, among others, the data:

accuracy / completion / relevance

timeliness / detail / representation

… and **context!**

J. C. Wyatt and J. L. Y. Liu, “Basic concepts in medical informatics.,” J. Epidemiol. Community Health, vol. 56, no. 11, pp. 808–12, Nov. 2002.

“Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house.”

Henri Poincarré (1952) Science and Hypothesis

(also borrowed from ECML/PKDD t-shirts, Pisa 2004)

Anomalies in health data depend on the context

from errors...

to outliers...

to hidden concepts...

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Anomalies in health (ADT) data

Some real examples

Admission Discharge Transfer (ADT) data

Admission-discharge-transfer (ADT) systems are a fundamental pillar regarding patient information on health care institutions.

They are used to maintain the master patient index, and the official list of patient encounters with the institution.

While it can include some clinical data, it mainly focus on scheduling and reporting patients encounters.

E. H. Shortliffe and J. J. Cimino, Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, 2006, p. 1064.

Anomalies in health data depend on the context

Nationwide admissions between 1993 and 2009, resulting in 160,853 admissions of patients with vascular disease, including information for 63 different variables.

Outcome: vascular disease was or was not the main diagnosis associated with each admission

Gritbot generated 491 rules were obtained that identify different types of anomalies

Note: Diagnosis-related group (DRG) is a system to classify hospital cases into homogeneous diagnosis group of each admission, from which are, for example, defined the payments made to the hospital. These codes can be generally clustered into medical or surgical type, thus variable GDHTIPO encodes the corresponding type of DRG (M: medical & C: surgical).

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Quinlan, R. (2007). GritBot:An Informal Tutorial, from http://www.rulequest.com/gritbot-unix.html

Contextual anomalies in medical data

Anomalies in health data depend on the context

case 37937: (label -259) [0.001]

CLTOTDIAS = -257 (156084 cases, mean 9, 99.99% >= -20)

This is considered an anomaly because it considers the total days of hospitalization as negative.

case 129252: (label 0) [0.000]

GDHTIPO = M (4807 cases, 99.98% `C')

SRG1 = 3522

Procedure 3522 with medical DRG (vs 99.98% of the subgroup sample).

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Contextual anomalies in medical data

Anomalies in health data depend on the context

case 58386: (label 0) [0.004]

GDHTIPO = M (1110 cases, 99.73% `C')

ADMTIP = 6

The rule means: admission for additional production of surgery encoded with medical DRG.

A possible explanation is that the patient did not actually had the surgery, for some reason, hence requiring coding with medical DRG.

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Contextual anomalies in medical data

Anomalies in health data depend on the context

case 85036: (label 0) [0.002]

DDXBin = no (1154 cases, 99.83% `yes')

CLIDADAN > 81 [85]

ADMTIP = 2

DRG = 135

Patient with more than 81 years, non-scheduled admission, coded with DRG 135, was not encoded with valvular heart disease as main diagnosis (vs 99.83% of the subgroup sample).

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Contextual anomalies in medical data

Anomalies in health data depend on the context

case 34461: (label 0) [0.011]

DDXBin = yes (1847 cases, 99.73% `no')

ADMTIP = 2

CLTOTDIAS <= 9 [6]

DRG = 122

Non-scheduled admission, inpatient less than 9 days and DRG 122 with a valvular disease as main diagnosis (vs 99.73% of the subgroup sample).

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Anomalies in health data depend on the context

case 13526: (label 0) [0.013]

DDXBin = no (2862 cases, 99.06% `yes')

CLIDADAN > 59 [74]

ADMTIP = 2

CLTOTDIAS > 5 [36]

DRG = 135

Non-scheduled admission, aged over 59 years, hospitalized for more than 5 days and DRG 135 does not have valvular disease as main diagnostis (vs 99.06% of the subgroup sample).

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

Take away message

M1: Anomalies in health data depend on the context.

S1: Better search for anomalies using a subgroup analysis.

Anomalies in health (clinical) data

The toy example

Uncertainty in clinical data records

Imagine you have access to a clinical record where there is a binary variable labeled “Penicillin”.

You ask the data curator (if they exist) or the MD responsible for that record what does it mean, and they say:

«Isn't it obvious? It records whether the patient is allergic to penicillin or not.»

So you happily use it in your knowledge discovery process as a well informed variable...

But does it really mean that?

Uncertainty in clinical data records

You ask to see the form used to gather that data and it reads:

[ ] Allergic to Penicillin

If the box is checked, then the patient is allergic to penicillin; but what if the box is left unchecked?

Uncertainty in clinical data records

You ask to see the form used to gather that data and it reads:

[ ] Allergic to Penicillin

If the box is checked, then the patient is allergic to penicillin; but what if the box is left unchecked?

New forms try to reduce the uncertainty in data registers by using

Is the patient allergic to penicillin?

( ) Yes (*) No

Uncertainty in clinical data records

You ask to see the form used to gather that data and it reads:

[ ] Allergic to Penicillin

If the box is checked, then the patient is allergic to penicillin; but what if the box is left unchecked?

New forms try to reduce the uncertainty in data registers by using

Is the patient allergic to penicillin?

( ) Yes (*) No

( ) Yes ( ) No (*) Unknown

Uncertainty in clinical data records

You ask to see the form used to gather that data and it reads:

[ ] Allergic to Penicillin

If the box is checked, then the patient is allergic to penicillin; but what if the box is left unchecked?

New forms try to reduce the uncertainty in data registers by using

Is the patient allergic to penicillin?

( ) Yes (*) No

( ) Yes ( ) No (*) Unknown

( ) Yes ( ) No ( ) Unknown (*) Not applicable

Uncertainty in clinical data records

You ask to see the form used to gather that data and it reads:

[ ] Allergic to Penicillin

If the box is checked, then the patient is allergic to penicillin; but what if the box is left unchecked?

New forms try to reduce the uncertainty in data registers by using

Is the patient allergic to penicillin?

( ) Yes (*) No

( ) Yes ( ) No (*) Unknown

( ) Yes ( ) No ( ) Unknown (*) Not applicable

( ) Yes ( ) No ( ) Unknown ( ) Not applicable (*) Not yet checked

Uncertainty in clinical data records

You ask to see the form used to gather that data and it reads:

[ ] Allergic to Penicillin

If the box is checked, then the patient is allergic to penicillin; but what if the box is left unchecked?

New forms try to reduce the uncertainty in data registers by using

Is the patient allergic to penicillin?

( ) Yes (*) No

( ) Yes ( ) No (*) Unknown

( ) Yes ( ) No ( ) Unknown (*) Not applicable

( ) Yes ( ) No ( ) Unknown ( ) Not applicable (*) Not yet checked

But clinical practice implies even **harder uncertainty**...

Uncertainty in clinical data records

But clinical practice implies even **harder uncertainty**...

Is the patient allergic to penicillin?

Uncertainty in clinical data records

But clinical practice implies even **harder uncertainty**...

Is the patient allergic to penicillin?

( ) Doctor knows “Yes” ( ) Doctor knows “No”

( ) Patient says “Yes” ( ) Patient says “No”

( ) Unknown ( ) Not applicable (*) Not yet checked

So, what's the meaning of our precious data variable “Penicillin” now?

Take away message

M1: Anomalies in health data depend on the context.

S1: Better search for anomalies using a subgroup analysis.

M2: Recorded (especially secondary) data is hard to interpret.

S2: Better acknowledge the protocol used to collect the data.

Take away message

M1: Anomalies in health data depend on the context.

S1: Better search for anomalies using a subgroup analysis.

M2: Recorded (especially secondary) data is hard to interpret.

S2: Better acknowledge the protocol used to collect the data.

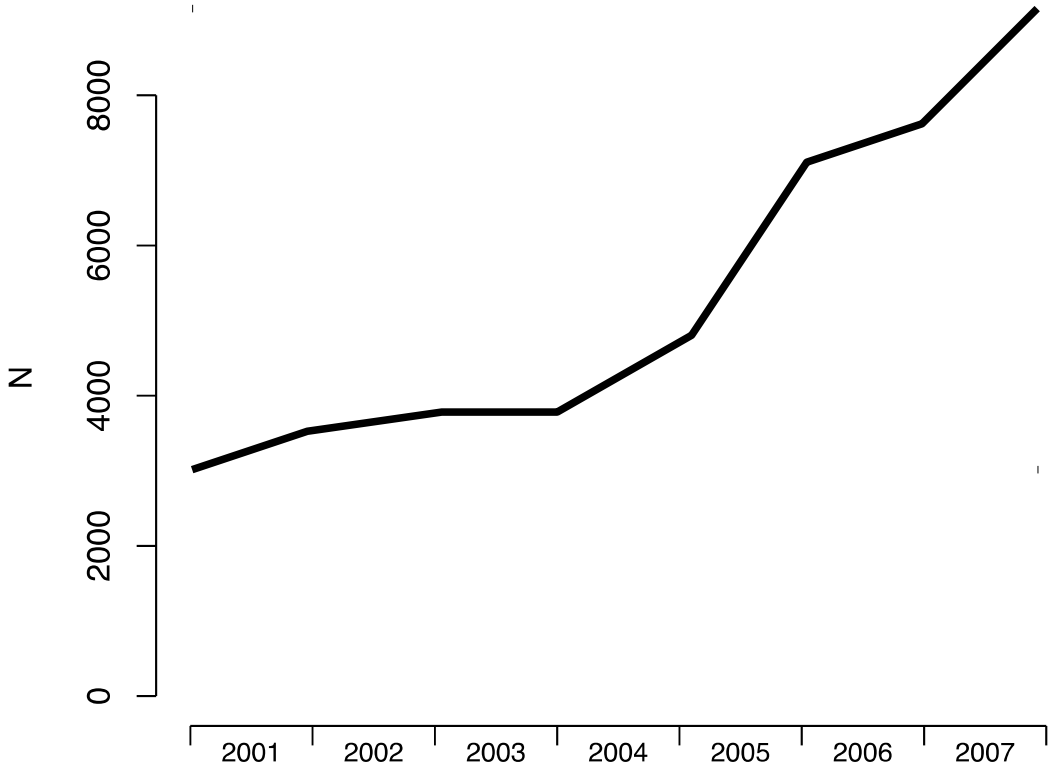
But has the protocol been correctly used?

Anomalies in health (clinical) data

More real examples

Anomalies in clinical records

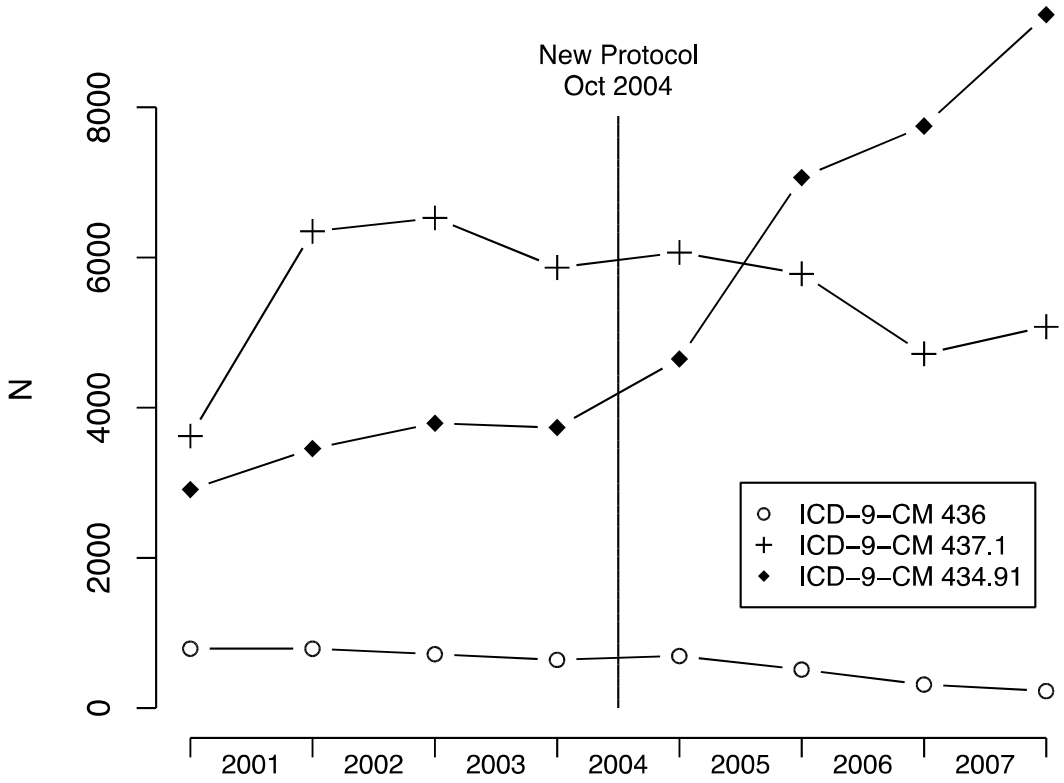
Using ICD-9-CM to code ischemic myocardial infarction (454.91)



R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

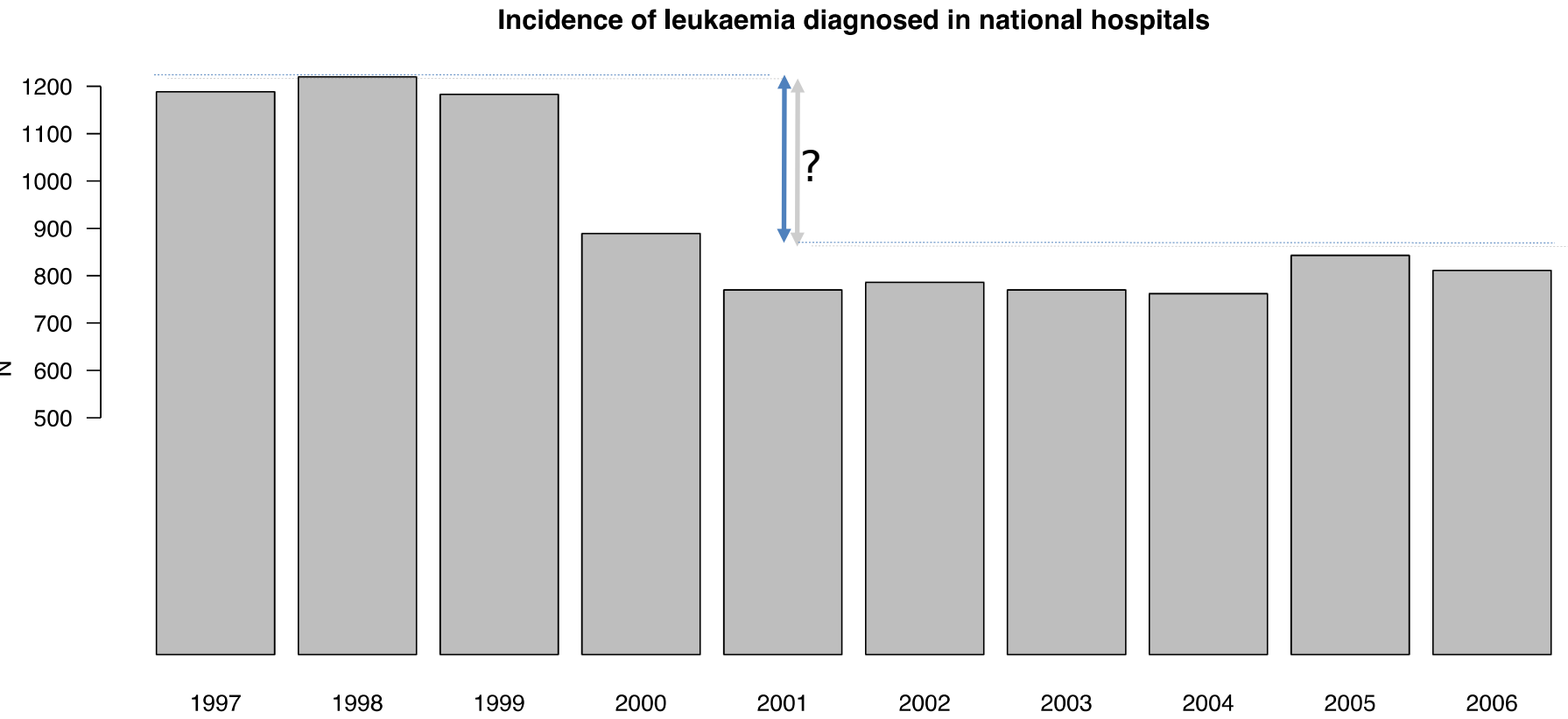
Anomalies in clinical records

Using ICD-9-CM to code ischemic myocardial infarction (454.91)



R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Anomalies in clinical records



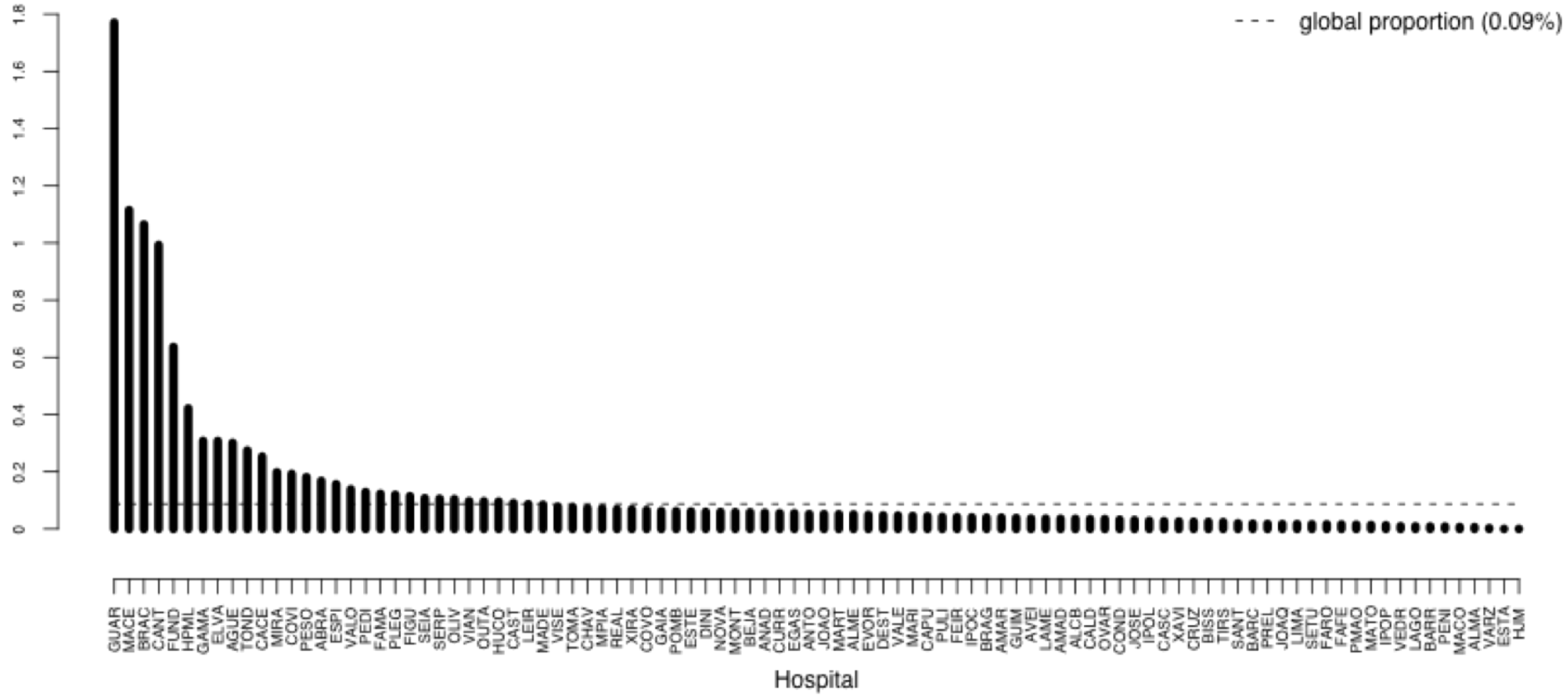
R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Anomalies in clinical data records

Proportion of admissions (Portugal, 2001-2007) with secondary diagnosis of flu

Anomalies in clinical data records

Proportion of admissions (Portugal, 2001-2007) with secondary diagnosis of flu



R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Anomalies in clinical data records

So now the problem is not that the data recording is anomalous...

… but the fact that the way humans follow protocol is uncertain!

So, let's take it to health services research...

Virtual Patient Record

Clinical Department HIS

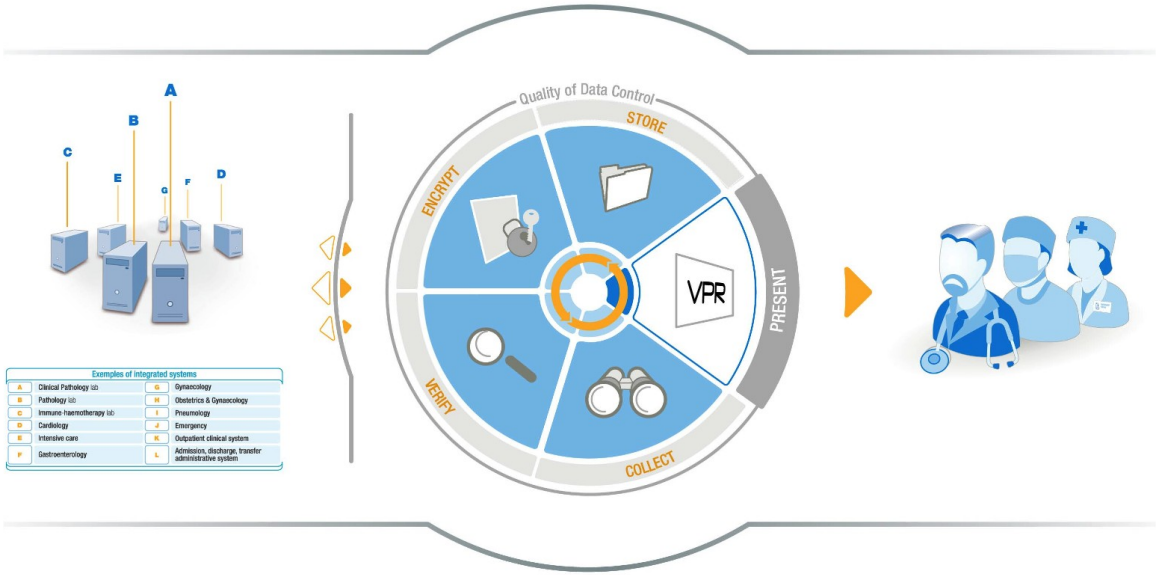
Obstetrics
Gynecology
Pneumology
Cardiology
Hematology
Breast Pathology
Psychiatry
Anesthesiology

Central HIS

EHR
Emergency
Management

LIS/RIS/PACS

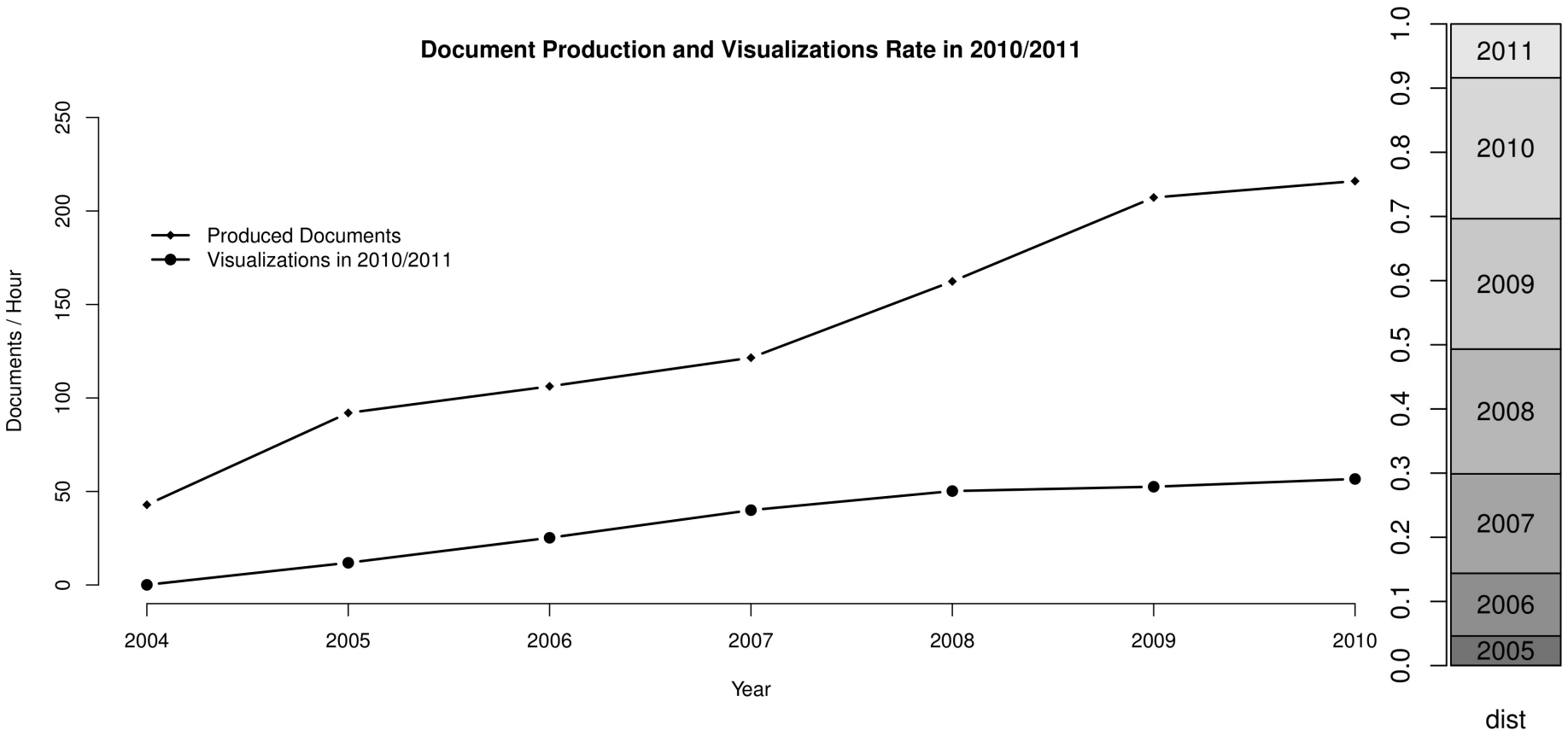
Clinical Pathology
Imunohemotherapy
Pathologic Anatomy
Radiology



Virtual Patient Record

Data de Entrada	Tipo	Serviço	Autor	Data de Emissão
04-04-2011 16:01	Exame - Hematologia e BioquImica	Patologia Clínica | HSJ e HSM	Medico Teste (HSJ)	22-10-2009 13:54
21-06-2011 16:01	Ecocardiografia - Problema Cardiaco	Obstetrícia | HSJ e HSM	Joana Obstetra (HSM)	21-06-2011 16:01
21-06-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	21-06-2011 16:01
21-06-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	21-06-2011 16:01
21-06-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	21-06-2011 16:01
21-06-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	21-06-2011 16:01
21-06-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	21-06-2011 16:01
21-06-2011 16:01	Ecocardiografia - Problema Cardiaco	Cardiologia - Cir. Torácica | HSJ e HSM	Cardiologia (HSJ)	21-06-2011 16:01
21-06-2011 16:01	Ecocardiografia - Problema Cardiaco	Cardiologia - Cir. Torácica | HSJ e HSM	Cardiologia (HSM)	21-06-2011 16:01
21-06-2011 16:01	Relatório cirúrgico de cesariana - Cesariana	Obstetrícia | HSJ e HSM	Lucinda Calejo	21-06-2011 16:01
21-06-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	21-06-2011 16:01
21-06-2011 16:01	Ecocardiografia - Problema Cardiaco	Obstetrícia | HSJ e HSM	Joana Obstetra (HSM)	21-06-2011 16:01
21-06-2011 16:01	Nota de Alta - Pré-eclampsia	Obstetrícia | HSJ e HSM	Teresa Rodrigues (HSJ)	21-06-2011 16:01
16-04-2004 12:17	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	08-04-2004 13:33
16-04-2004 12:17	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	08-04-2004 13:32
16-04-2004 12:17	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	08-04-2004 13:21
16-04-2004 12:17	Template nº1 -	Anatomia Patológica | HSJ e HSM	Autor desconhecido	08-04-2004 13:19
16-04-2004 12:17	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	08-04-2004 11:07
04-04-2011 16:01	Template nº1 -	Cuidados Intensivos | HSJ	Autor desconhecido	08-04-2004 11:00

Virtual Patient Record



P. P. Rodrigues and R. C. Correia, “Streaming Virtual Patient Records,” in Real-World Challenges for Data Stream Mining, 2013, pp. 34–37.

Virtual Patient Record

From 2010 to the first quarter of 2011

The hospital had **+530K records**:

+210K (39.33%) from immunohemotherapy

+146K (27.34%) anatomo-pathology

+127K (23.83%) clinical pathology

+17K (3.24%) cardiothoracic surgery

+10K (1.94%) gastroenterology

+8K (1.65%) obstetrics

+4.8K (0.91%) pneumology

+3.9K (0.75%) clinical hematology

+2.1K (0.41%) intensive care

+1.1K (0.22%) breast pathology

+1.1K (0.21%) from the gynaecology endoscopy unit

P. P. Rodrigues, C. C. Dias, D. Rocha, I. Boldt, A. Teixeira-Pinto, and R. Cruz-Correia, “Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 461–464.

Visualization depends on some factors

Setting:

- Consult reports (OR=0.098)
- Inpatient stays reports (OR=4.007)
- Emergency encounters (OR=5.641)

Department:

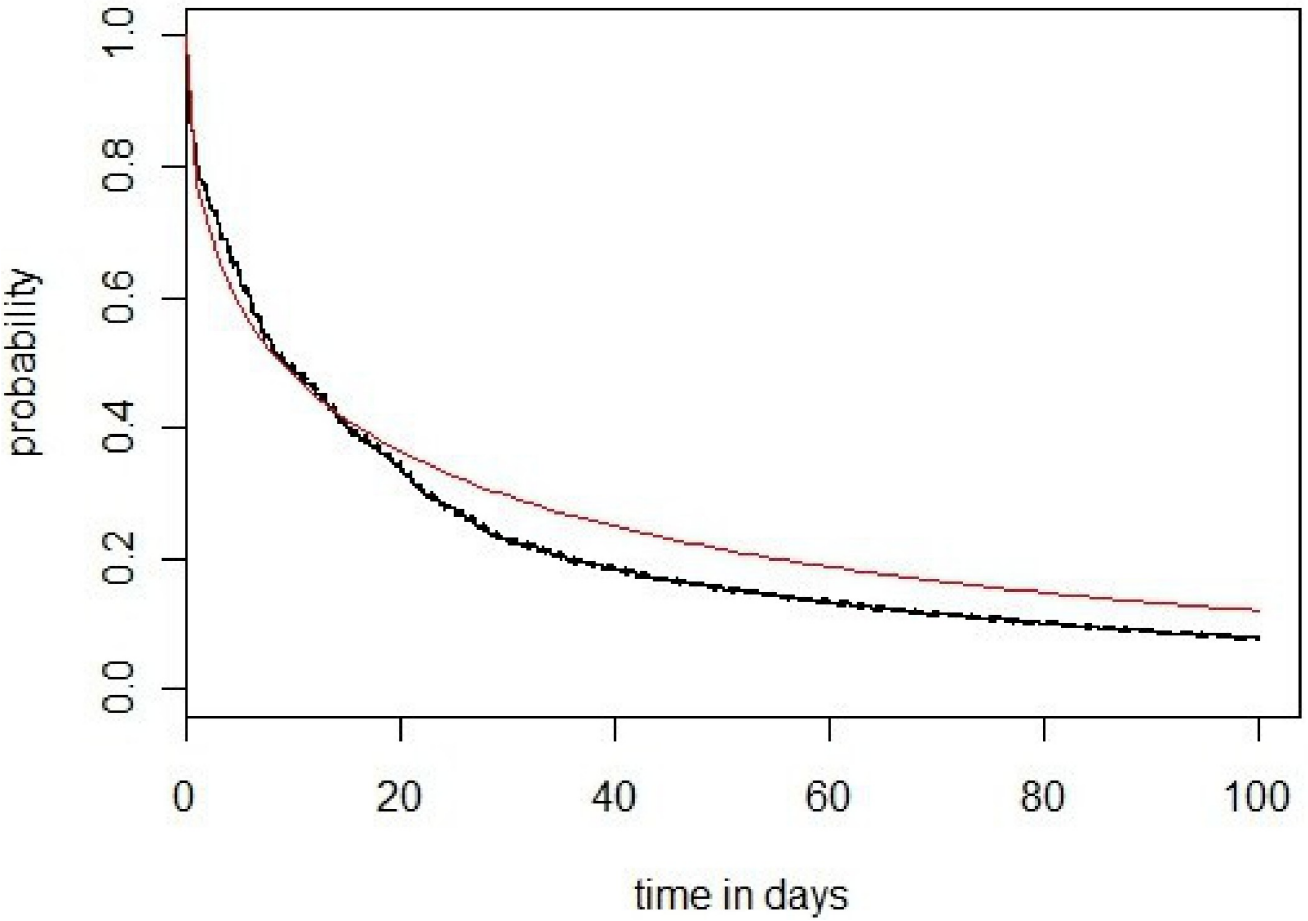
- immunohemotherapy (OR=2.418)
- gynecologic endoscopy unit ones (OR=0.106)

Type of report:

- gastroenterology reports are only slightly more likely to be visualized (OR=1.018) unless they are of type 11 case when they are much more likely to be visualized (OR=6.753)
- cardiotoraccic surgery report are less likely to be visualized (OR=0.205) unless they are of type 27 case when they are more likely to be visualized (OR=2.762).

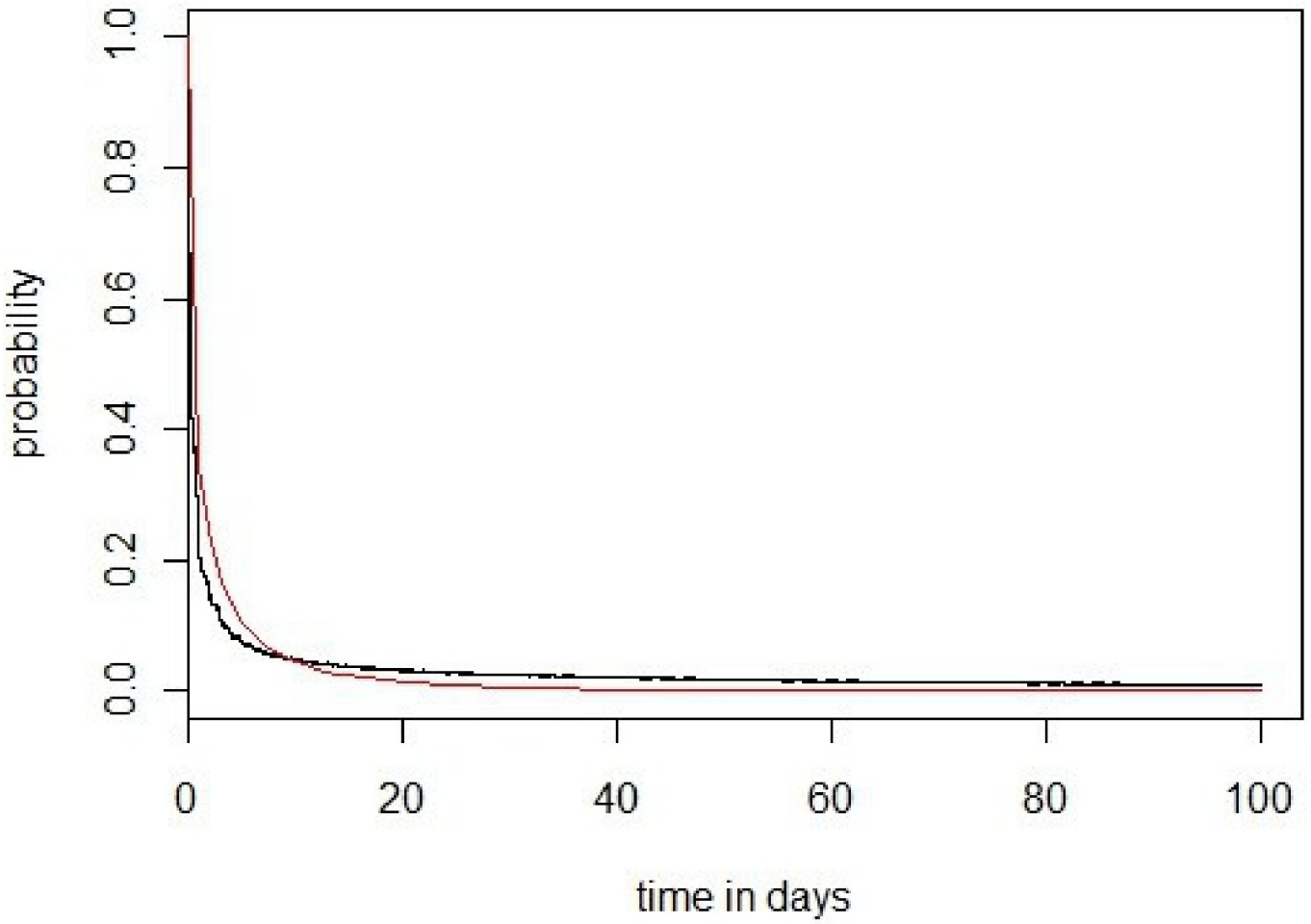
P. P. Rodrigues, C. C. Dias, D. Rocha, I. Boldt, A. Teixeira-Pinto, and R. Cruz-Correia, “Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 461–464.

Survival models for single type of report



P. P. Rodrigues, C. C. Dias, D. Rocha, I. Boldt, A. Teixeira-Pinto, and R. Cruz-Correia, “Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 461–464.

Survival models for single type of report



P. P. Rodrigues, C. C. Dias, D. Rocha, I. Boldt, A. Teixeira-Pinto, and R. Cruz-Correia, “Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 461–464.

Prediction error when using survival models

The median error of using those models compared to the curves of actual data was:

- 6% (min:1%, max: 52%, for outpatient consults),
- 17.5% (min=1%, max=50%, for inpatient stays), and
- 21% (min=3%, max=28%, for emergency encounters).

P. P. Rodrigues, C. C. Dias, D. Rocha, I. Boldt, A. Teixeira-Pinto, and R. Cruz-Correia, “Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 461–464.

Probability of visualization

Probability of visualization of radiology reports (X-ray, CT, MRI)

Setting:

- Consult reports (brown)
- Inpatient stays reports (green)
- Emergency encounters (blue)

Kaplan-Meier curve resulted in astonishing results...

Probability of visualization

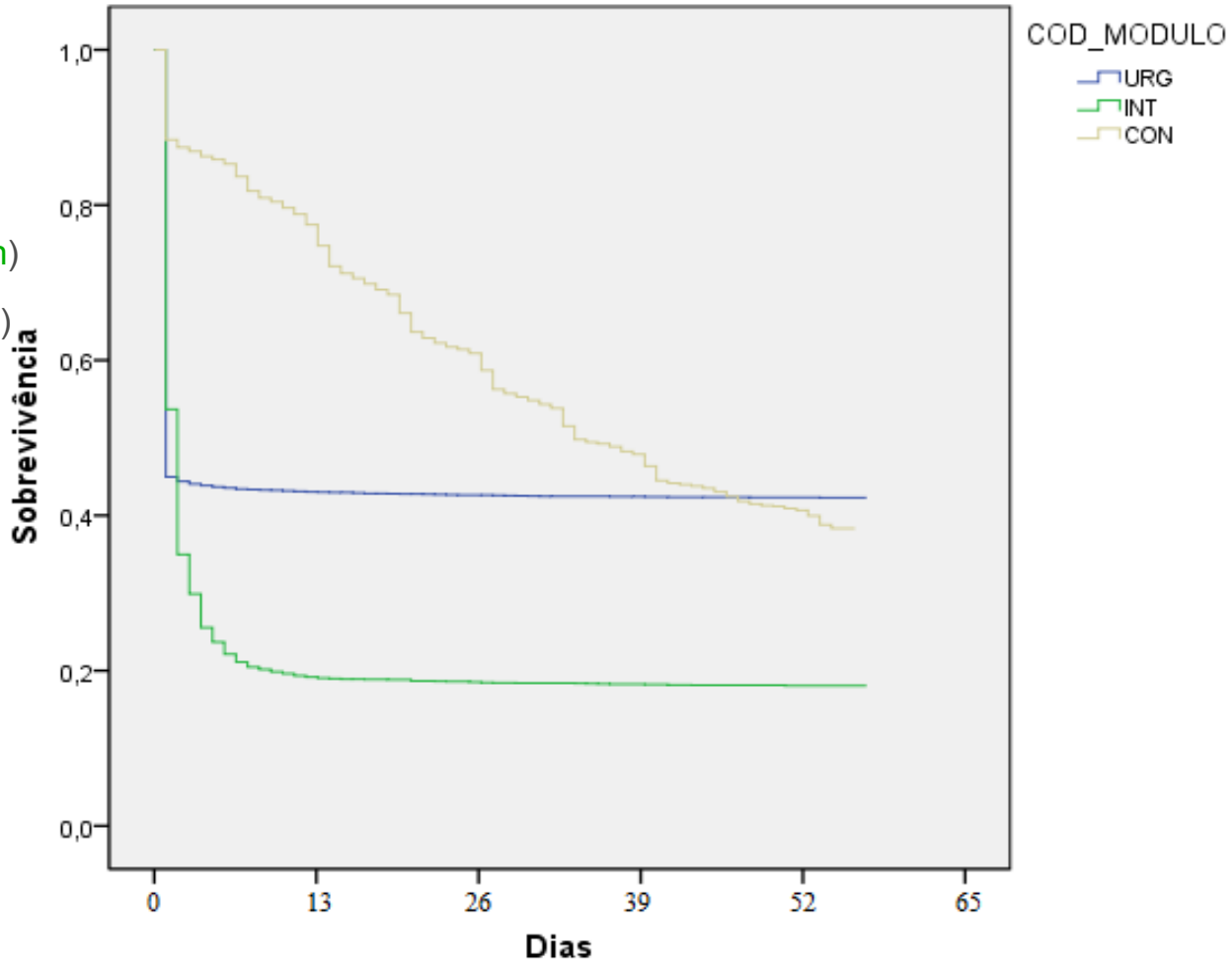
Probability of visualization of radiology reports (X-ray, CT, MRI)

Setting:

Consult reports (brown)

Inpatient stays reports (green)

Emergency encounters (blue)



Probability of visualization

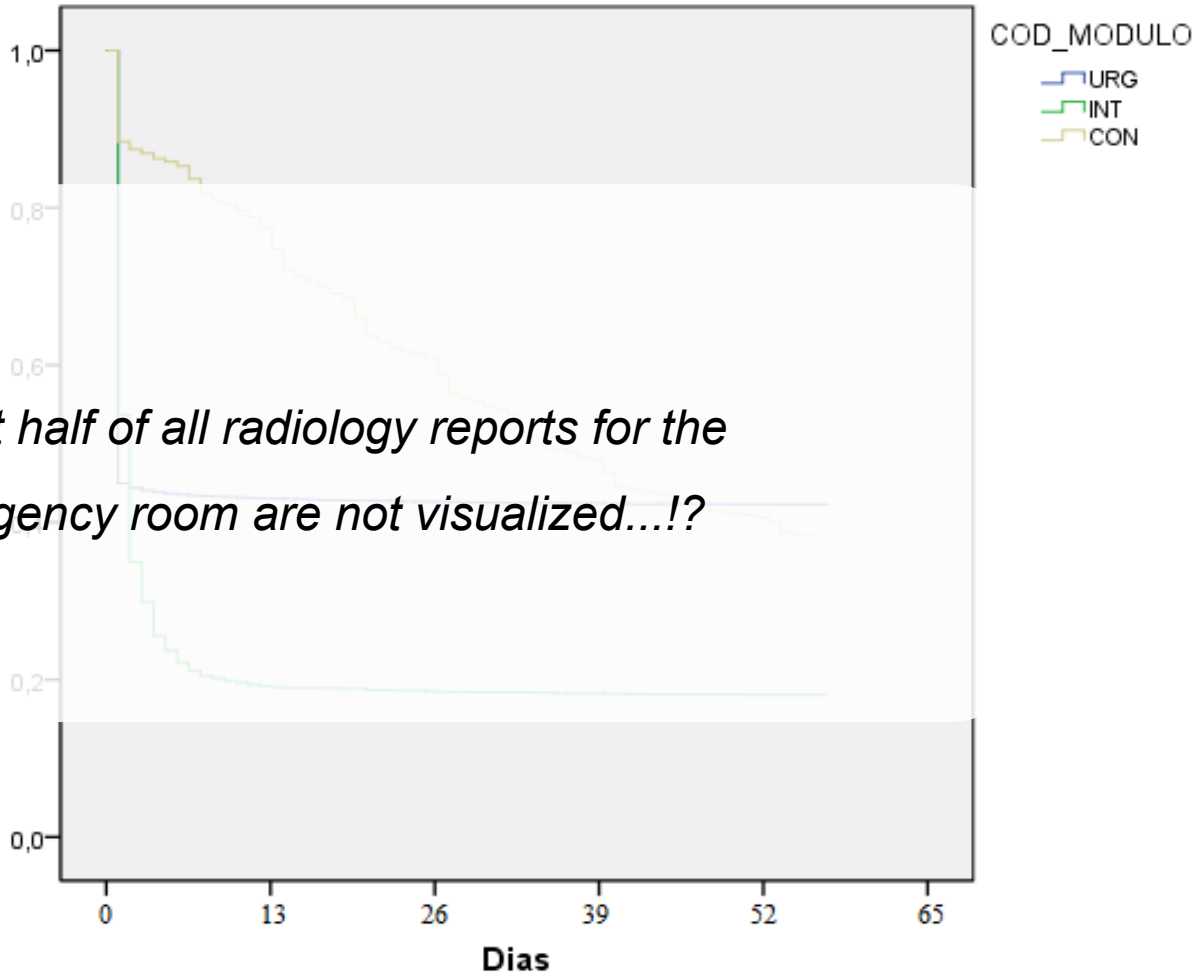
Probability of visualization of radiology reports (X-ray, CT, MRI)

Setting:

Consult reports (brown)

Inpatient stays reports (green)

Emergency encounters (blue)



Almost half of all radiology reports for the emergency room are not visualized...!?

Probability of visualization

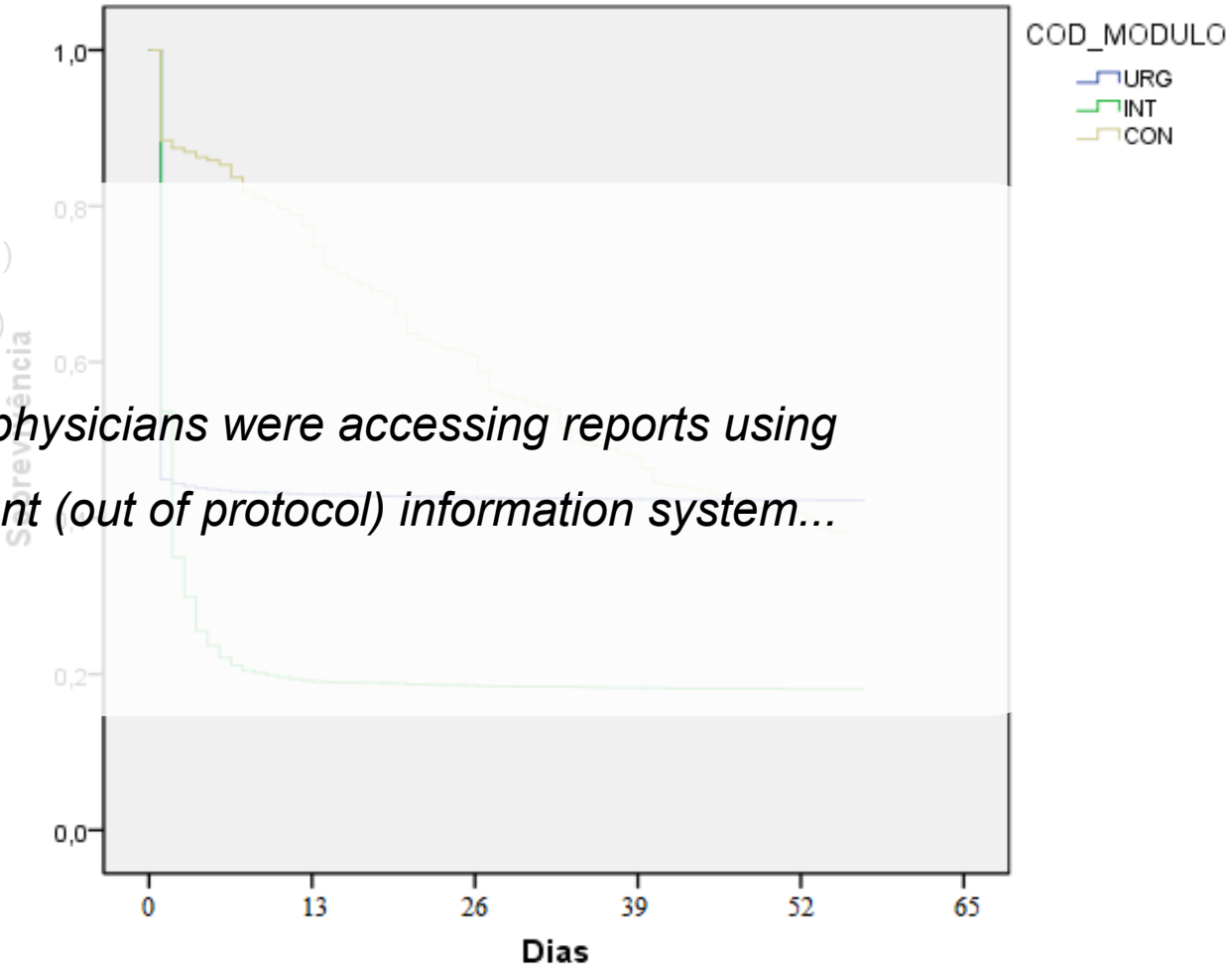
Probability of visualization of radiology reports (X-ray, CT, MRI)

Setting:

Consult reports (brown)

Inpatient stays reports (green)

Emergency encounters (blue)



In fact, physicians were accessing reports using a different (out of protocol) information system...

Take away messages

M1: Anomalies in health data depend on the context.

S1: Better search for anomalies using a subgroup analysis.

M2: Recorded (especially secondary) data is hard to interpret.

S2: Better acknowledge the protocol used to collect the data.

M3: Humans tend to override the protocol... quite often.

S3: Better expect several bias in data entry points.

Take away messages

M1: Anomalies in health data depend on the context.

S1: Better search for anomalies using a subgroup analysis.

M2: Recorded (especially secondary) data is hard to interpret.

S2: Better acknowledge the protocol used to collect the data.

M3: Humans tend to override the protocol... quite often.

S3: Better expect several bias in data entry points.

Can simpler data be as unreliable?

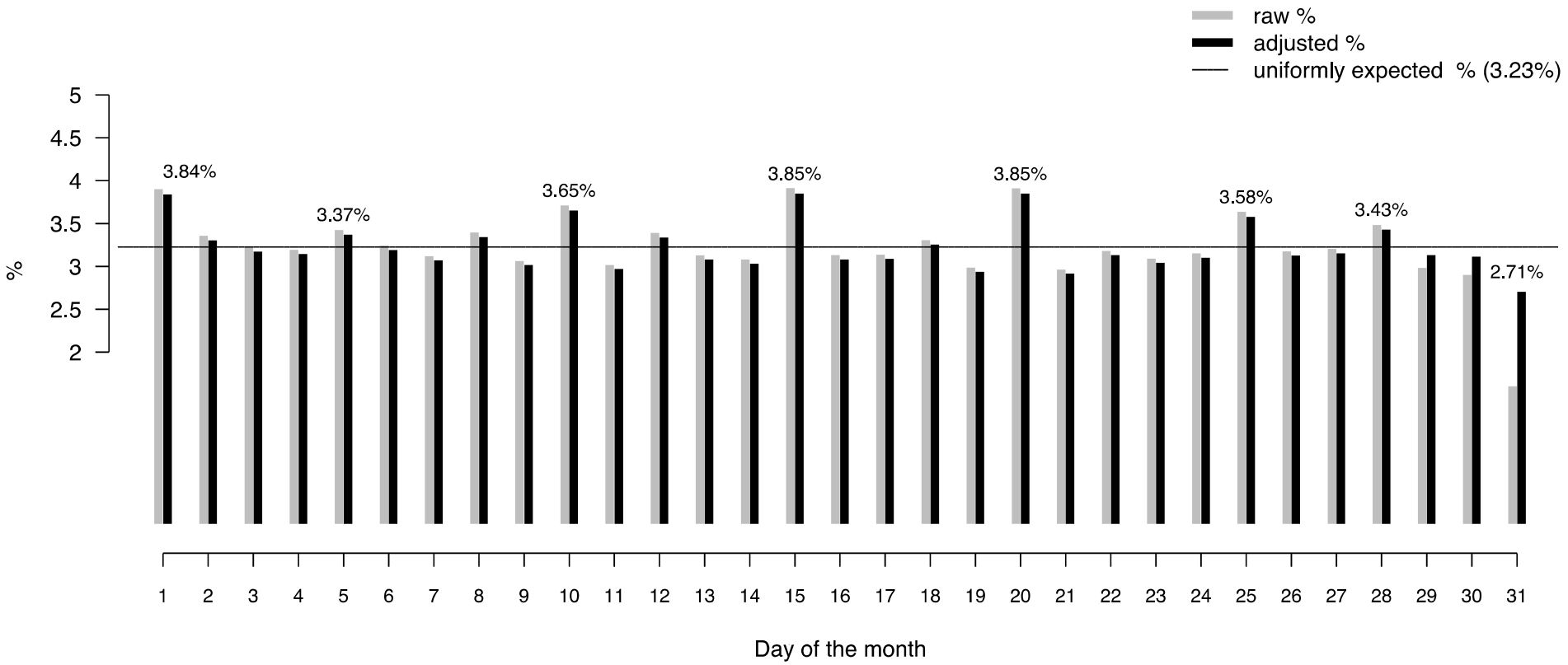
Uncertainty in ADT data records

Frequency of patients' birthday by day of the month (hospital admissions 2000-2007)

Uniformly distributed?

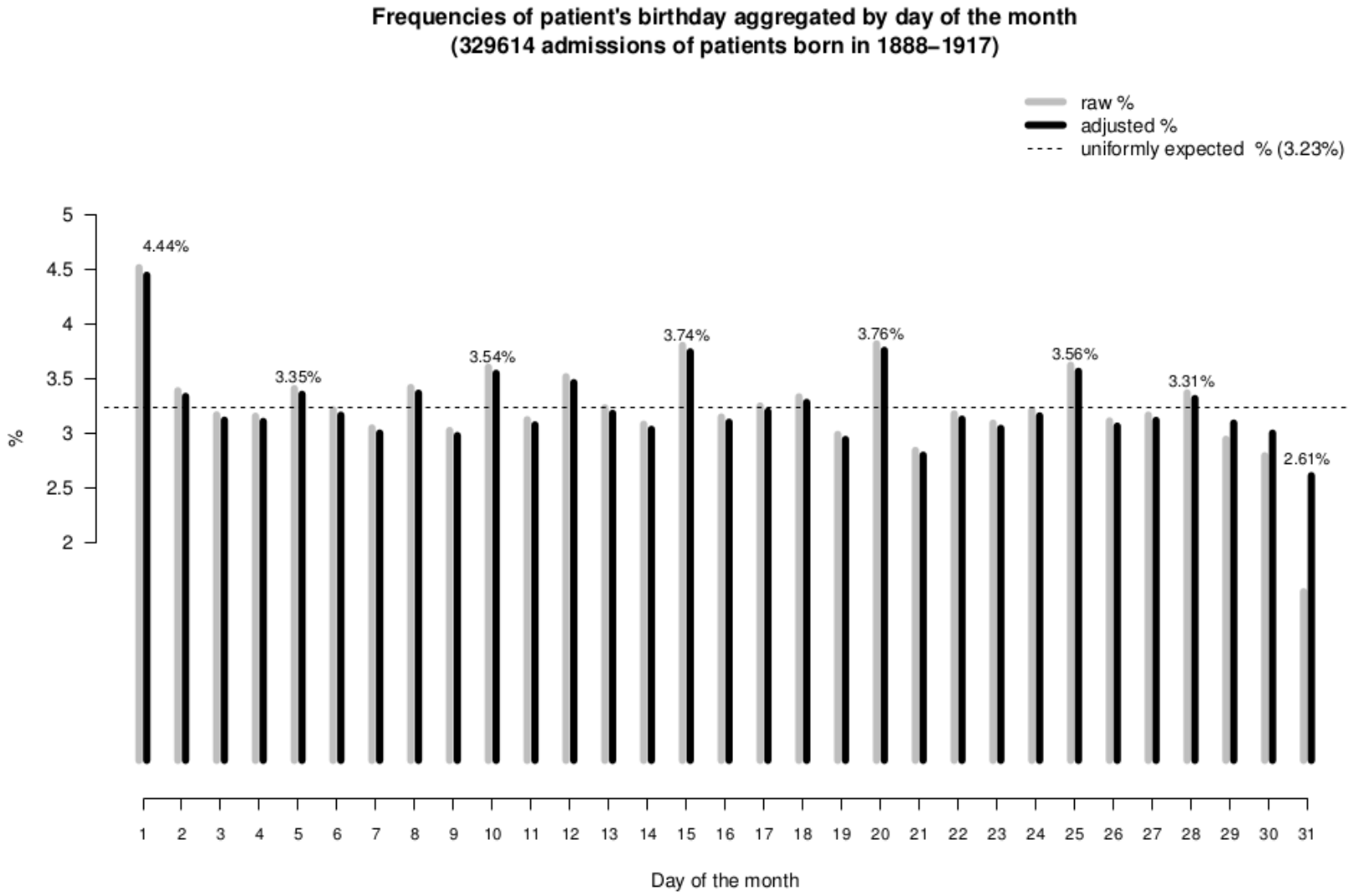
Uncertainty in ADT data records

Frequency of patients' birthday by day of the month (hospital admissions 2000-2007)

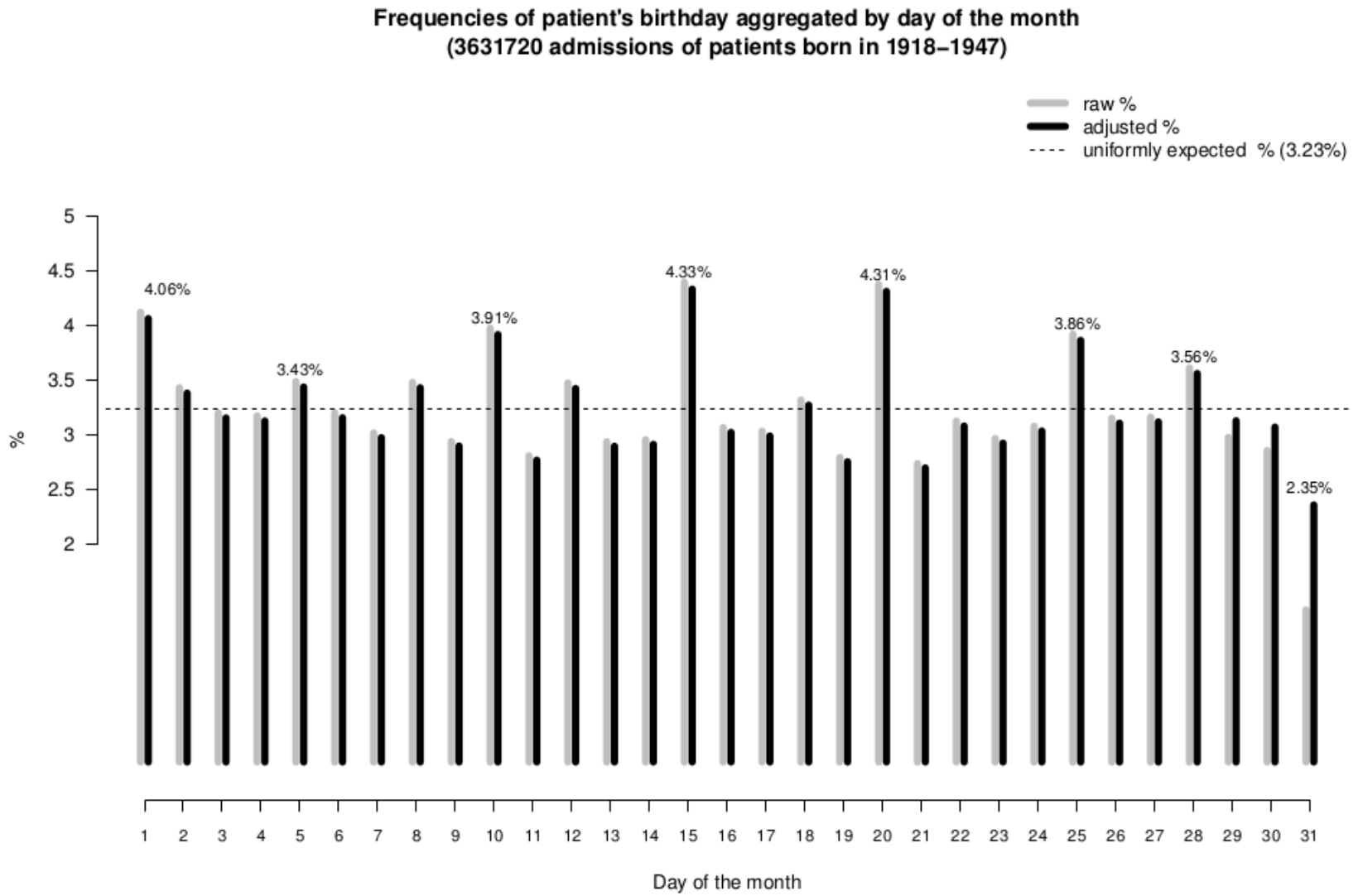


R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

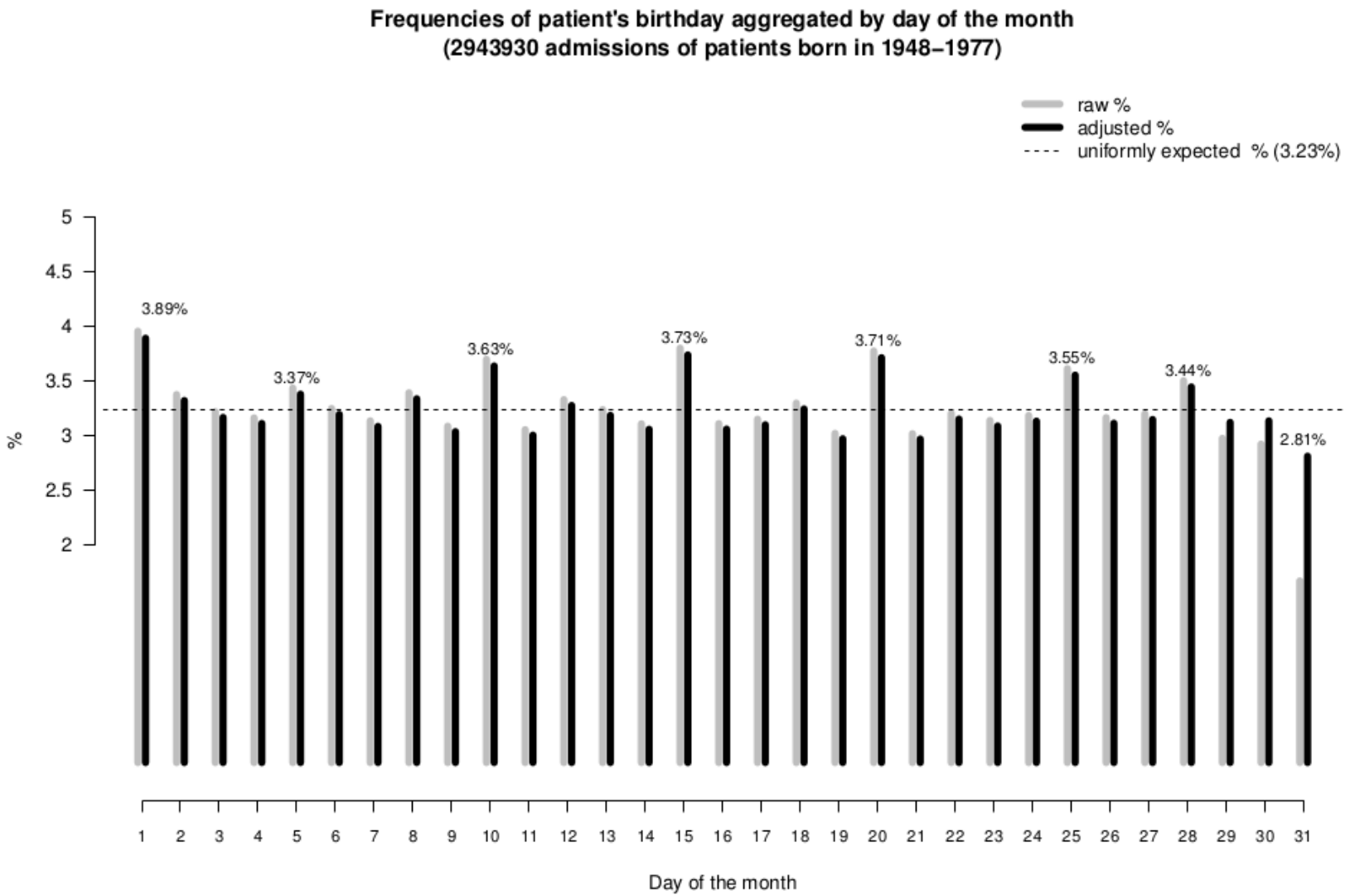
Uncertainty in ADT data records



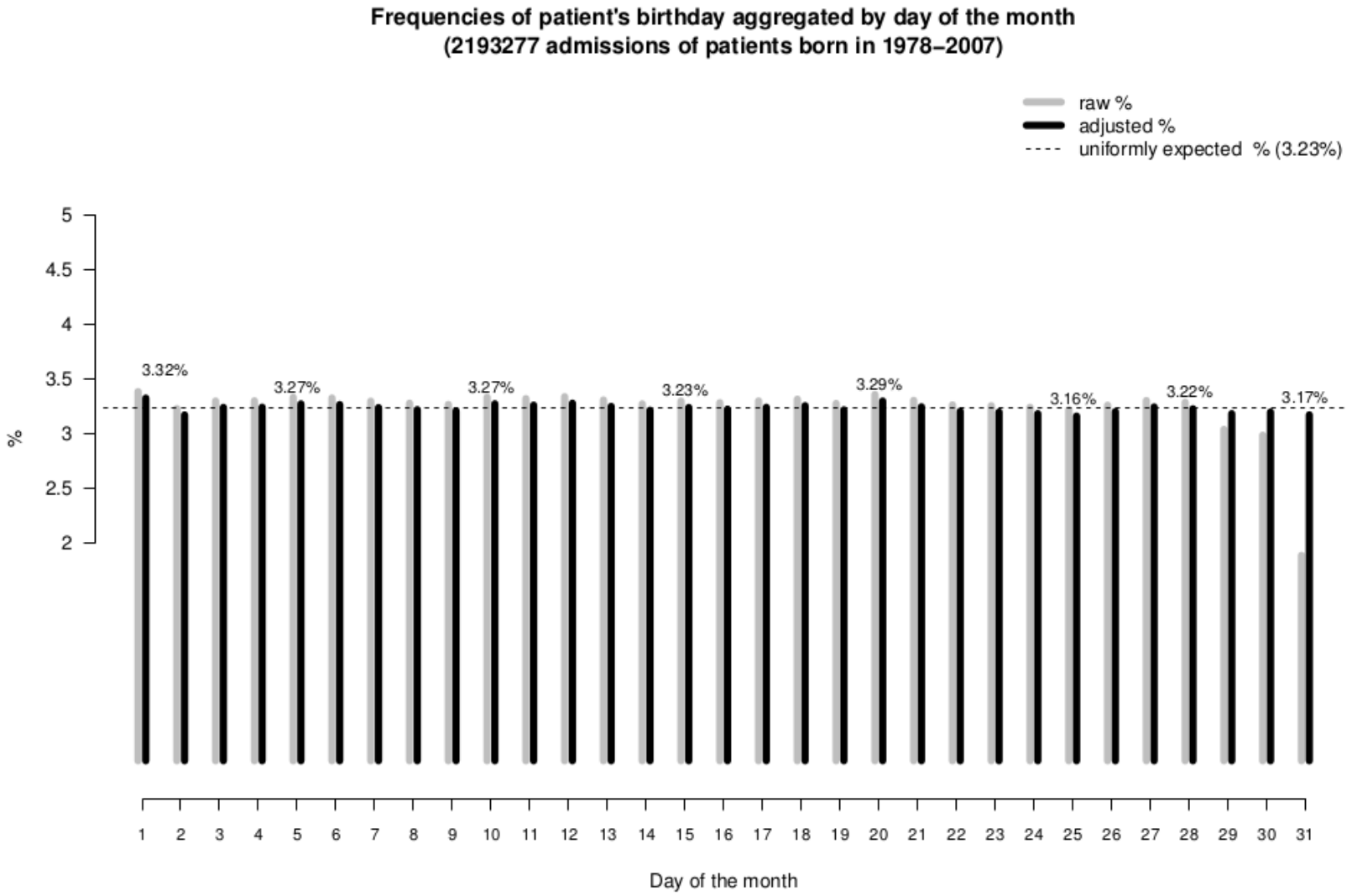
Uncertainty in ADT data records



Uncertainty in ADT data records



Uncertainty in ADT data records

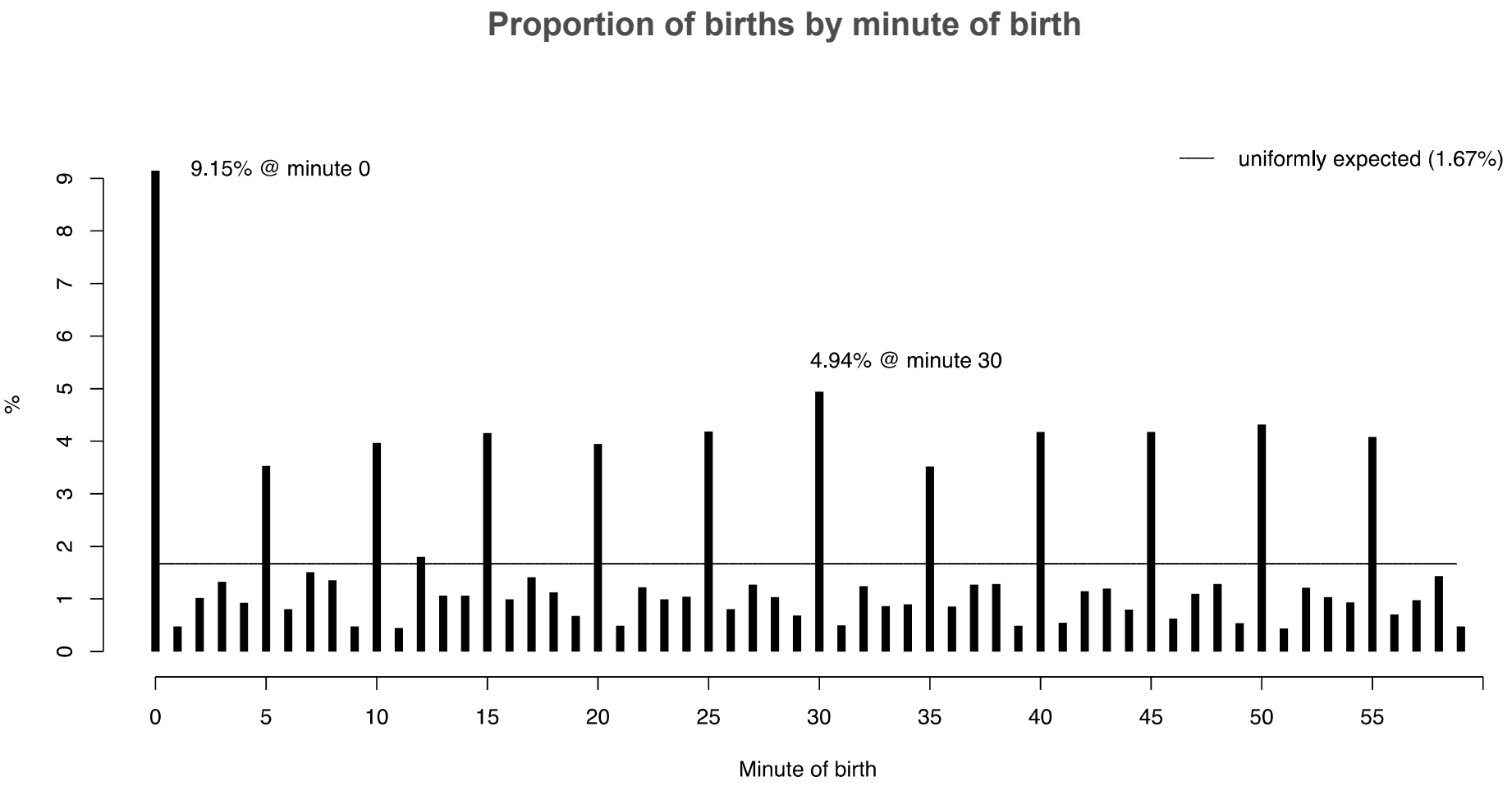


Uncertainty in ADT/clinical data records

Proportion of births by minute of birth

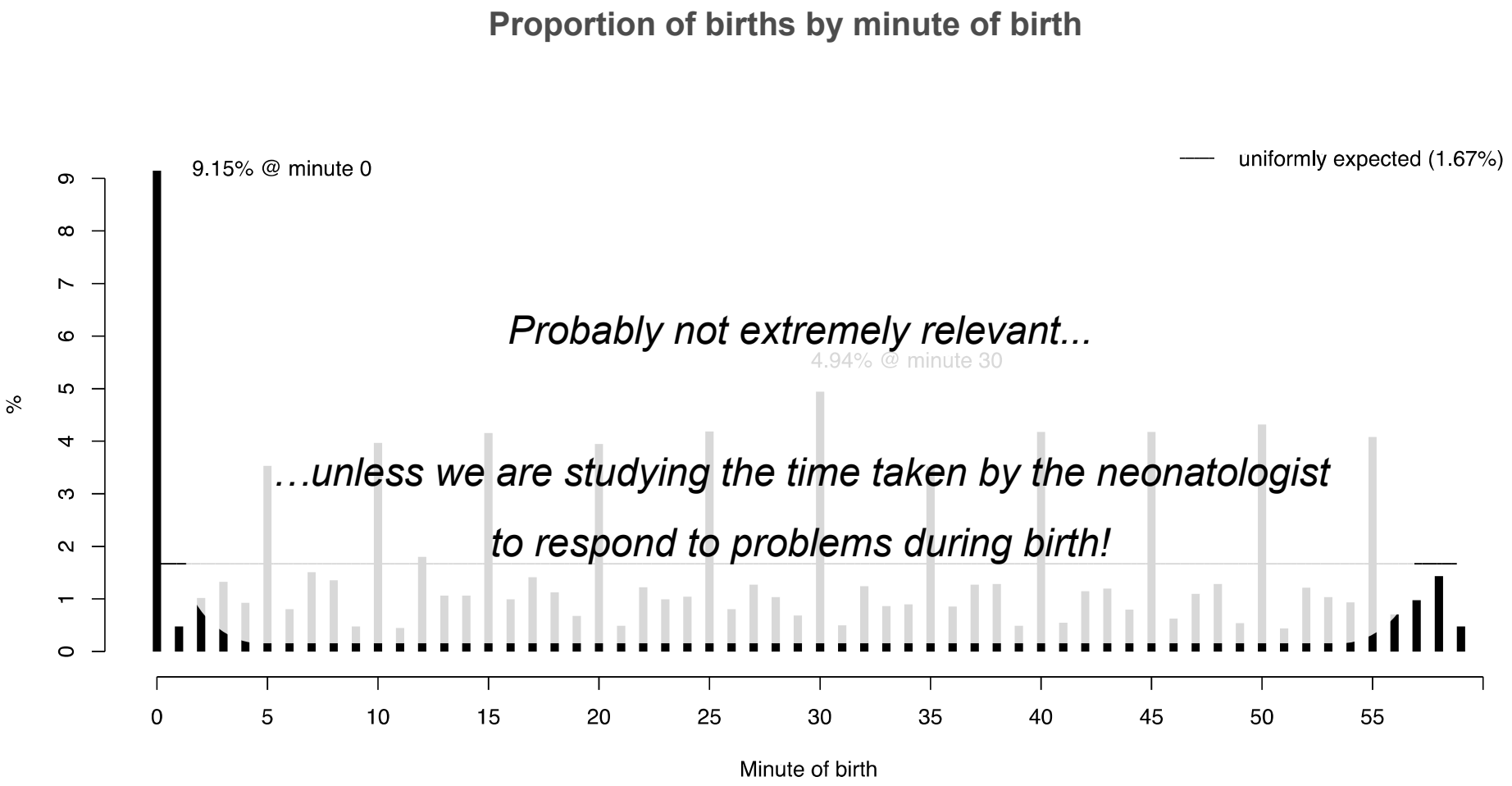
Uniformly distributed?

Uncertainty in ADT/clinical data records



R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Uncertainty in ADT/clinical data records

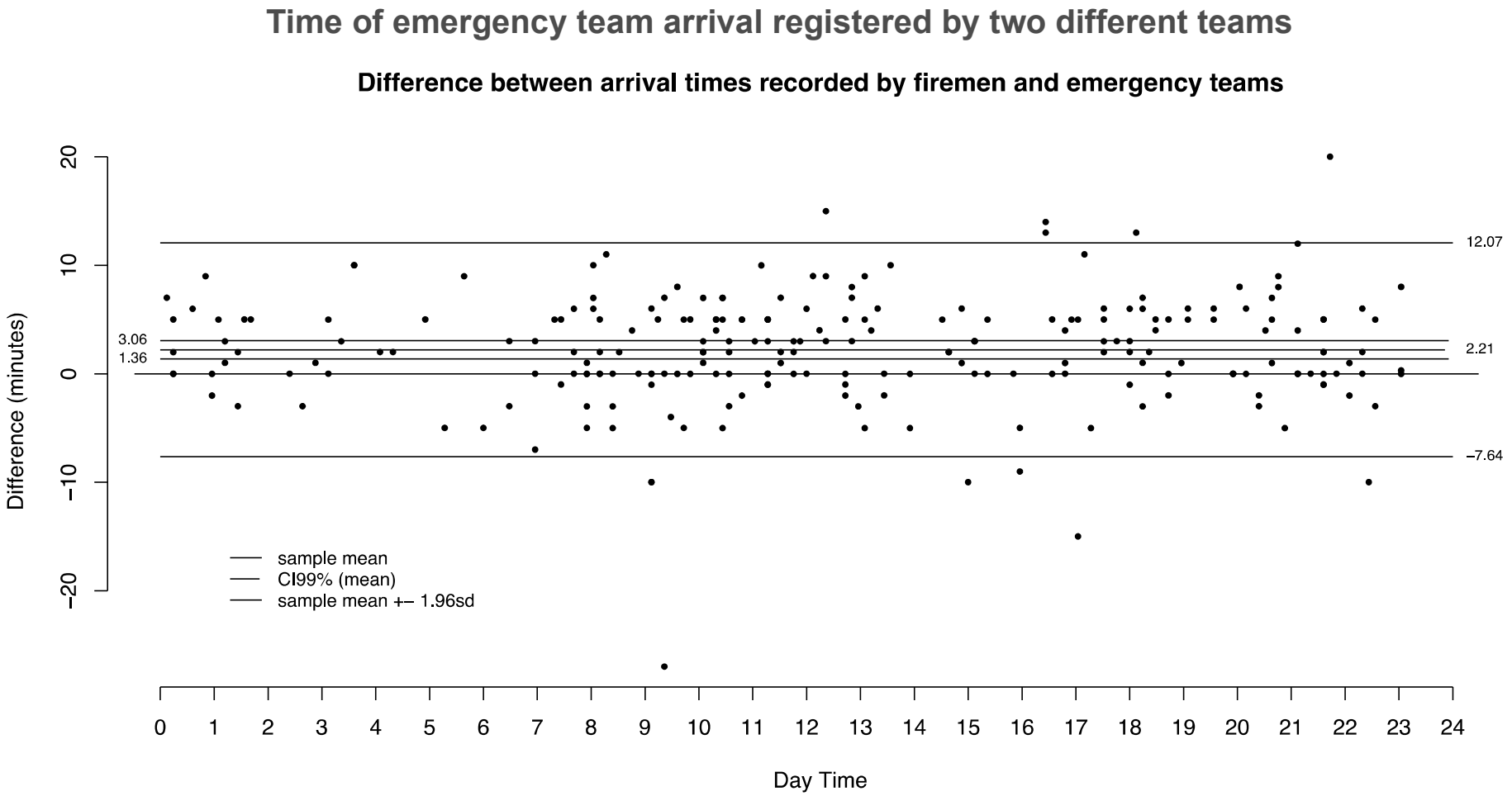


R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Uncertainty in ADT/clinical data records

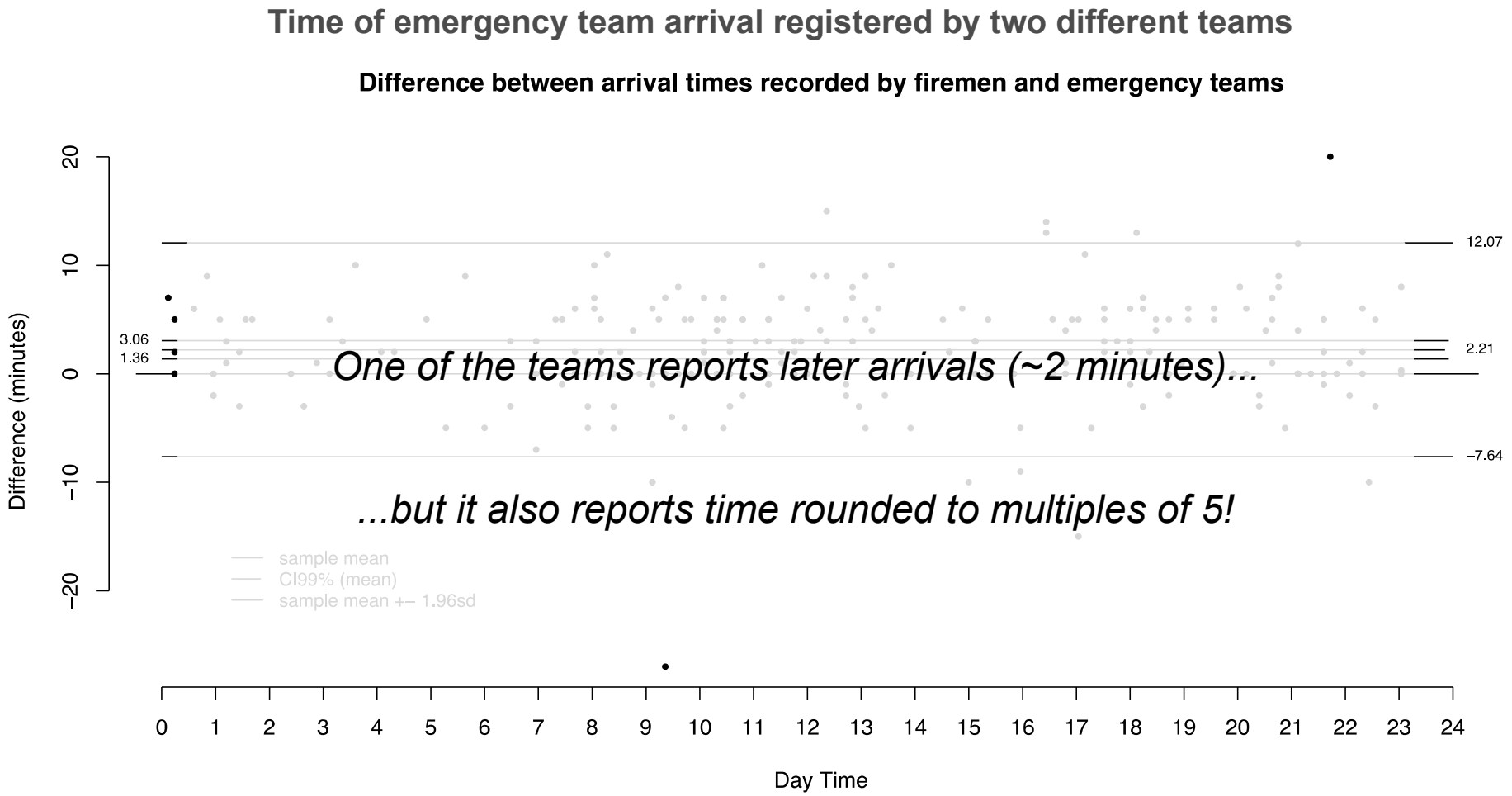
Time of emergency team arrival registered by two different teams

Uncertainty in ADT/clinical data records



R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Uncertainty in ADT/clinical data records



R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

Take away messages

M1: Anomalies in health data depend on the context.

S1: Better search for anomalies using a subgroup analysis.

M2: Recorded (especially secondary) data is hard to interpret.

S2: Better acknowledge the protocol used to collect the data.

M3: Humans tend to override the protocol... quite often.

S3: Better expect several bias in data entry points.

M4: Recorded (especially secondary) data is never what it seems at first.

S4: Better suspect positive results and proceed with caution...

Health data problems

"There are a lot of small data problems that occur in big data. They do not disappear because you have got lots of the stuff. They get worse."

David Spielgelhalter (2014)

Acknowledgements

Ricardo Cruz-Correia

Alberto Freitas

Altamiro Costa-Pereira

Daniela Vasco

João Gama

Leila Pereira

Cláudia Camila Dias

Armando Teixeira-Pinto

Diana Rocha

Isabel Boldt

References

R. Cruz-Correia, P. P. Rodrigues, A. Freitas, F. Almeida, R. Chen, and A. Costa-Pereira, “Data Quality and Integration Issues in Electronic Health Records,” in Information Discovery on Electronic Health Records, V. Hristidis, Ed. CRC Press, 2009, pp. 55–95.

D. Vasco, P. P. Rodrigues, and J. Gama, “Contextual anomalies in medical data,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 544–545.

P. P. Rodrigues, C. C. Dias, D. Rocha, I. Boldt, A. Teixeira-Pinto, and R. Cruz-Correia, “Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record,” in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 461–464.

J. C. Wyatt and J. L. Y. Liu, “Basic concepts in medical informatics.,” J. Epidemiol. Community Health, vol. 56, no. 11, pp. 808–12, Nov. 2002.

P. P. Rodrigues and R. C. Correia, “Streaming Virtual Patient Records,” in Real-World Challenges for Data Stream Mining, 2013, pp. 34–37.

Quinlan, R. (2007). GritBot: An Informal Tutorial, from http://www.rulequest.com/gritbot-unix.html

D. Hand and R. Till, “A simple generalisation of the area under the ROC curve for multiple class classification problems,” Mach. Learn., vol. 45, pp. 171–186, 2001.

E. H. Shortliffe and J. J. Cimino, Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, 2006, p. 1064.