Forensic Science International: Genetics 12 (2014) 77–85

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Error rates in forensic DNA analysis: Definition, numbers, impact and communication

Ate Kloosterman [a,b,c,*], Marjan Sjerps [b,d], Astrid Quak [a]

[a] Department of Human Biological Traces (HBS), Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands
[b] Department of Science, Interdisciplinary Research, Statistics and Knowledge Management (WISK), Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands
[c] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[d] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 29 April 2014
Accepted 30 April 2014

Keywords:
Lab error rate
Forensic DNA analysis
Contamination
Misidentification
Quality control

ABSTRACT

Forensic DNA casework is currently regarded as one of the most important types of forensic evidence, and important decisions in intelligence and justice are based on it. However, errors occasionally occur and may have very serious consequences. In other domains, error rates have been defined and published. The forensic domain is lagging behind concerning this transparency for various reasons.

In this paper we provide definitions and observed frequencies for different types of errors at the Human Biological Traces Department of the Netherlands Forensic Institute (NFI) over the years 2008–2012. Furthermore, we assess their actual and potential impact and describe how the NFI deals with the communication of these numbers to the legal justice system.

We conclude that the observed relative frequency of quality failures is comparable to studies from clinical laboratories and genetic testing centres. Furthermore, this frequency is constant over the five-year study period. The most common causes of failures related to the laboratory process were contamination and human error. Most human errors could be corrected, whereas gross contamination in crime samples often resulted in irreversible consequences. Hence this type of contamination is identified as the most significant source of error. Of the known contamination incidents, most were detected by the NFI quality control system before the report was issued to the authorities, and thus did not lead to flawed decisions like false convictions. However in a very limited number of cases crucial errors were detected after the report was issued, sometimes with severe consequences. Many of these errors were made in the post-analytical phase.

The error rates reported in this paper are useful for quality improvement and benchmarking, and contribute to an open research culture that promotes public trust. However, they are irrelevant in the context of a particular case. Here case-specific probabilities of undetected errors are needed. These should be reported, separately from the match probability, when requested by the court or when there are internal or external indications for error. It should also be made clear that there are various other issues to consider, like DNA transfer. Forensic statistical models, in particular Bayesian networks, may be useful to take the various uncertainties into account and demonstrate their effects on the evidential value of the forensic DNA results.

1. Introduction

Forensic DNA casework is currently regarded as one of the most important types of forensic evidence and has known many forensic success stories over the years. The basis of this is a combination of sound scientific principles, and a reliable and robust technological platform. An increasing number of judicial decisions currently rely on accurate DNA typing test results. However, as with all scientific investigations, analyses are subject to error. An error in forensic DNA analysis can lead to wrong decisions by investigative or legal authorities with far reaching consequences, such as conviction of innocent suspects, exoneration of guilty suspects, or failure to identify offenders. Furthermore, cases where a major miscarriage

[*] Corresponding author at: Department of Human Biological Traces (HBS), Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands. Tel.: +31 708886666.
E-mail address: a.kloosterman@nfi.minvenj.nl (A. Kloosterman).

http://dx.doi.org/10.1016/j.fsigen.2014.04.014
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

of justice was caused by an erroneous DNA result often generate a lot of media attention and damage the reputation of forensic laboratories [1].

The NFI had its share of such doomed forensic DNA cases. One case was the so called Avenger of Zuuk [2] where the DNA profile of an unknown woman in a series of arson and anonymous threats was identified from one of the postal items. The reference samples of more than 50 women who volunteered to provide their DNA in a mass screening in the Dutch rural community Zuuk had to be destroyed when it turned out that the supposedly incriminating DNA profile was the result of laboratory contamination and matched a NFI technician. Another case involved contamination of a crime sample with DNA from another sample resulting in a false match, that was noticed at a very late stage in the legal process. This case was extensively discussed by an influential journal for law representatives in the Netherlands [3]. The contamination in these cases was caused by gross contamination of the evidentiary sample which is a relatively rare event. The possibility of laboratory based background contamination which can give rise to drop-in alleles however is a far more common event and has to be considered on a routine base in the evaluation of complex DNA profiles from low template DNA samples [4]. Internationally the possibility of errors, although denied by some, is generally acknowledged by the field (see for critical views [5–7]).

The frequency with which errors occur has also received some attention. In the well-known case Daubert v. Merrell Dow Pharmaceuticals, Inc. [8], the Court listed the "known or potential rate of error of a particular scientific technique" as an important factor that might be considered by a trial judge. The importance of error rates has further been emphasized in numerous papers [9–12] and a thesis [13] and in three influential reports by the National Research Council [14–16]. For instance, recommendation 5 of the 1996 NRC report calls for more research to quantify and characterize the amount of error. It is even asserted that "The assessment of the accuracy of the conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science." (pp. 4–9)

In other scientific domains, it has since long been considered good practice to define, register and publish error rates [17]. A 'quality failure' in relation to a laboratory test can be defined [18] as any failure to meet the required output quality necessary for optimal patient care. This includes problems occurring in the pre-analytical, analytical or post-analytical phases. Table 1 shows examples of the frequency of such failures that were published in the medical literature.

In line with these studies the quality failure rate of a Genetic Test for Familial Hypercholesterolemia (FH) in the Netherlands has been assessed [19]. 1003 cases were included in this study. The overall accuracy of the genetic test for FH screening was 99.8%, with two false positive results identified.

Table 1
Published rates of quality failures in laboratories that analyze human DNA.

	Medical laboratory[a]	Medical laboratory[b]	Molecular genetic testing centre[c]	Medical laboratory[d]
Year of Publication	1997	1998	1999	2007
Data collection period	3 months	3 years	1 year	3 months
No. of tests	40 490	676 564	88 394	51 746
No. of errors	189	4135	293	160
Frequency (%)	0.47	0.61	0.33	0.31

[a] Plebani and Carraro [33].
[b] Stahl et al. [34].
[c] Hofgärtner and Tait [35].
[d] Carraro and Plebani [36].

Forensic practitioners, however, have struggled with the outcry for transparency on error rates by the academic community [20]. Even though forensic DNA analysis is often seen as the "gold standard" of forensic science, error rates in casework have, to the best of our knowledge, never been published at a detailed level. General rates, including "contamination" and "laboratorive failures (*sic*)" have been published by the Swedish forensic laboratory SKL [21]. We think that there are several reasons for the general lack of detailed figures. Besides the fear of possible reputation damage, there is still discussion and confusion about the definition and relevance of error rates, the estimation of them, and the communication to the legal justice system. Mnookin et al. [22] criticize this lack of transparency concerning mistakes and argue that a research culture is needed. Ansell [21] emphasizes that transparency is important both for quality improvement and for trust from the criminal justice system and the general public. We share these views, and consider this paper as our contribution to this culture change. The numbers provided may also be useful for benchmarking.

In this paper we provide definitions and observed frequencies for different types of errors at the Human Biological Traces Department (HBS) of the Netherlands Forensic Institute (NFI) over the years 2008–2012. Furthermore, we assess their impact and describe how the NFI deals with the communication of these numbers to the legal justice system. We start in Section 2 by describing the way that "quality issue notifications" (QINs) are registered at the NFI, and how their impact is assessed. In Section 3 we classify different types of QINs and report their frequencies over the years 2008–2012. Since contamination is an important type of failure, we zoom in on this type in Section 3.2. The consequences of the QINs are reported in Section 3.3.

Section 4 addresses the important question what to do with these numbers. Obviously, the relevant errors for the legal justice system are those that are not recognized as erroneous in the specific case. The observed frequency of recognized errors reported in this paper thus provides useful information to assess the relevant error rate, but they are not the same. So what should we do with this information in the context of a case? Should error rates be part of the forensic testimony? Should they be combined with the profile random match probability in a single number describing the evidential strength of the match? Should the court be guided in their interpretation of the numbers? We describe and motivate the pragmatic approach of the NFI to deal with these questions in Sections 4.1–4.3. Section 5 discusses the quality system and research culture needed to cope with errors in a professional way and the use of Bayesian networks for error detection, analysis and communication.

2. Data

2.1. Quality issue notifications (QINs): registration and handling

Notification of quality issues is part of the standard process of quality improvement. All staff members are authorized to report quality issues and have access to the electronic quality system that is installed on their computers (standard workplace). The quality issue notification concerns the following items: administrative details, description of the quality issue, cause, extent of the problem, action and measures taken to correct the failure and prevent future incidents and operational nature of the improvement. All notifications are scrutinized by the quality control manager of the department. The root cause of the quality issue is identified and appropriate corrective actions are taken. The head of each department has to approve the handling of all notifications. Raw data (in Dutch) for each department of the NFI are made public and can be downloaded at http://www.nederlandsforensischinstituut.nl/over_het_nfi/organisatie/kwaliteit/kwaliteitsrapportages.aspx. Data are available from 1997 onwards for the biology department (DNA only) and from 2008 onwards for all departments of the NFI. For privacy reasons, the personal information such as names of people and identifiers of the criminal case are removed and only the general information is published at the website. In this paper, we focus on the notifications of the NFI HBS department from 2008 to 2012.

2.2. Quality issue notifications (QINs): assessing impact

The QINs are very diverse and range from harmless typographical errors to crucial errors. Before any conclusions about consequences can be drawn, it is necessary to assess the impact of each QIN and add this assessment to the data. There are several ways to assess the impact of a QIN; this section describes our choices.

When assessing the impact of a QIN we first have to define what we actually try to achieve with this concept. Thus we need to answer the question: impact on what? Various choices are possible here, ranging from impact on just the involved samples to impact on the final decision by the court. Our choice is based on the idea that the impact should somehow represent meaningful effects on the case, for either the prosecution, defence, or the court. However, a practical restriction is that we have only limited information about the case, hence an assessment of the consequences of each QIN for the court would be very speculative. Based on these considerations, we have chosen to define the impact as the consequences of the QIN for the conclusions of the NFI report.

We also need to answer the question: what kind of impact? As detailed below, there is an important distinction between potential and actual impact. It may seem appealing to focus only on actual impact. However, both types are very informative: the potential consequence of a QIN is relevant for improvement of the procedures, and the actual consequence is useful information to evaluate the importance of errors in casework. Therefore, we chose to assess both potential and actual impact of each QIN.

Next we have to define a way of measuring actual and potential impact. This can also be done in numerous ways. Preferably, the measurement is a number produced by an objective tool. However, due to the diversity of QINs and our definition of impact this is simply infeasible. A practical solution is to subjectively grade each QIN using a limited number of categories. This subjective grading is done by the quality control manager in cooperation with a senior DNA expert, and is finally approved by the head of the department. We consider this further in Section 5.

For the assessment of the *potential impact* of QINs, it is of course relevant whether or not the QIN affects the conclusions in the NFI report. Furthermore, it is important whether or not the consequence is irreversible. The potential impact is therefore scored in three categories:

0. *no consequences*: the QIN has no consequences for the conclusions of the NFI report
1. *repairable consequences*: the QIN potentially has consequences for the conclusions of the NFI report, but they can be corrected
2. *irreversible consequences*: the QIN potentially has consequences for the conclusions of the NFI report, that cannot be corrected

For instance, when a reference sample is destroyed, there are no potential consequences for the conclusion of the NFI report, because one of the back-up samples can be used or even a new sample can be taken from the person. When samples are switched, this potentially has serious consequences, however they can easily be corrected when the switch is recognized at an early stage in the process. When a crime sample is contaminated with relatively large amounts of DNA, this potentially has serious consequences for the conclusions of the NFI report that cannot be corrected.

The *actual impact* can differ drastically from the potential impact. For example, when an analyst cuts himself on paper, this can potentially have major consequences because the crime sample may have been contaminated. However, when the DNA profile of the crime sample does not indicate contamination with the analyst's DNA profile, the actual consequence for the conclusions of the NFI report are nil.

For the assessment of the actual impact, again a distinction is made between affecting the conclusions or not, and whether correction is possible or not. In addition, it is important whether the QIN was made before or after the report was issued. This is because a wrong conclusion in a report may have serious consequences for the way a case proceeds externally, even if that conclusion is corrected later on. For instance, if it is erroneously concluded that the suspect did not contribute DNA to the crime sample, the suspect may be released and commit a second crime. The actual impact is scored in four categories:

0. *no consequences*: corrections were made (when necessary) before the report was issued and the QIN had no consequences for the conclusions of the NFI report
1. *consequences corrected in revised report*: corrections were made after the report was issued and a revised report was issued
2. *irreversible consequences stated in report*: the QIN was made before the report was issued but corrections were not possible; consequences were stated in the NFI report
3. *irreversible consequences stated in revised report*: the error was noticed and the QIN registered after the report was issued. Corrections were not possible and the consequences were stated in a revised report

3. Results

3.1. Types and rates of notifications

The first relevant question concerning quality issues is how often they occur. In Table 2 the absolute number of QINs is presented for the years 2008–2012, as well as their frequency relative to the total number of DNA analyses. Although a single DNA analysis may concern several different cases the number of criminal cases in which DNA-analysis was performed is far less than the number of analyzed samples, since most cases often involve several analyses.

We see an increase in the absolute number of notifications. This may be explained by a management decision in 2009 to include all clerical errors as a QIN, and by the increase in the number of analyses. The relative frequency of QINs is stable over the years.

The numbers presented in Table 2 are only to a certain extent comparable to the frequencies of quality failures presented in Table 1, because the QINs involve not only failures but also other types of notifications. We distinguished between external origin of the quality issue (categories a and b below), i.e., the origin of the error lies outside the NFI, and internal origin (categories c–g). For

Table 2
(Relative) frequency of quality issue notifications (QINs) at the NFI in the years 2008–2012.

	2008	2009	2010	2011	2012
No. of DNA-analyses	66 391	82 896	89 977	100 407	132 456
No. of notifications	328	329	435	526	572
Frequency (%)	0.5	0.4	0.5	0.5	0.4

Table 3
Types of quality issue notifications (QINs) at the NFI in the years 2008–2012. In 2011 it was decided to no longer incorporate the type c QIN: opportunities for improvement ($n = 2$ in 2011 and $n = 10$ in 2012) in the yearly totals of this overview.

	2008	2009	2010	2011	2012
a. External origin	23	10	23	54	100
b. External contamination	3	0	5	24	22
c. Room for improvement	11	6	3	(2)	(10)
d. Positive response	19	9	11	6	17
e. Clerical (no adverse outcome)	29	25	92	77	82
f. Not related to case work	13	9	20	10	5
g. Other (NFI related)	230	270	281	355	346
Total	328	329	435	526	572

the internal origins, a further distinction is made between issues clearly not related to quality failures in NFI casework (category c, d and f), minor clerical errors (category e) that have adverse outcome and issues that may result in such a quality failure (category g).

We define the following categories of notifications:

a. The quality issue does not concern contamination and the origin is external (outside the NFI)
b. The quality issue concerns contamination and the origin is external
c. Remarks for improving the process, for instance, its efficiency. From 2011 onwards, these kinds of remarks were processed in a different way and are no longer part of the notification data set
d. Positive remarks such as compliments
e. Shortcomings that have no adverse outcome for the case, e.g., clerical errors in the administration of the case
f. Quality issues that are not related to casework, e.g., DNA positive surface wipe test from laboratory bench or failed acceptance test for reagents
g. All other NFI related internal notifications

The frequencies of the different types are presented in Table 3.

We see that most QINs fall into category g. This category contains the QINs where the origin lies within the NFI and may result in a quality failure related to NFI casework; we will focus on this important category. Their frequency relative to the number of DNA analyses ranges between 0.3% and 0.4%, i.e., is in line with quality failure rates published in the medical domain (Table 1). The different causes of the type g-QINs are categorized in Table 4.

We see that the most common causes are contamination and human error. Contamination events were detected in analyzed crime samples, reference samples, positive controls and reagent blanks. Human errors mostly concerned clerical errors that could usually be corrected and rarely had serious consequences for the case. Examples of other human errors are positive and negative controls that were overlooked to analyze and test results that were not administrated correctly. Technical problems concerned leaking of DNA extract from damaged mini tubes or 96 well plates; damaged cover foil on 96 well plates, problems with the extraction robot and liquid handling workstation.

Table 4
Causes of quality issue notifications (QINs) of type g: other (NFI related) at the NFI in the years 2008–2012.

	2008	2009	2010	2011	2012
Contamination	49	56	57	130	135
Human error	105	124	135	139	114
Technical problem	17	28	37	21	19
Deviation quality document	0	0	3	5	2
Capacity/planning	1	1	0	1	0
Deviation from competence matrix	0	1	0	0	0
Sample mix-up	24	32	25	30	34
Other	34	28	24	29	40
Ongoing					2
Total	230	270	281	355	346

Incidents of sample mix-up were identified in the phase of collecting the secured samples for DNA extraction, mix-up of DNA extracts in setting up the PCR, and during preparation of the amplified products for capillary electrophoresis. Most incidents in the sample mix-up category concerned crime samples that were mixed up. But also incidents with mix-up of reference samples have been detected.

A planning issue notification was made when DNA extracts from case samples were stored in the freezer without registration of the correct location.

The NFI competence matrix is used to determine competence requirements for each step in the analysis process against which staff members assigned to specific task(s) are assessed. Incidents where staff members perform tasks that they are not assigned to are registered.

Two notifications from 2012 are still being evaluated for the exact cause of the incident. One incident concerns a number blood samples from two cases that failed to produce a DNA profile for yet unknown reasons. The other notification concerns a possible mix-up of samples from a collaborative exercise.

The category "other" contains a miscellanea of QINs that are different to the other categories and which absolute numbers are too small for a separate evaluation. We will present the consequences of the QINs in the g-category, including those which were caused by human error, in Section 3.3.

Contamination of a crime sample can have serious consequences for the case. We will therefore now investigate contamination as an important cause of QINs in more detail.

3.2. Contamination

Notifications concerning contamination fall into different categories a–g defined earlier. Table 5 lists the number of all contamination events over the years 2008–2012, not just the g-type QINs as in Table 4 but also external contamination (category b in Table 3) and contaminations that are not related to case work, i.e., contamination in a wipe sample from the surface of a laboratory bench (category f in Table 3). In order to categorize contamination events we have made a distinction between internal contamination of the evidence/reference sample to be analyzed with DNA from a staff member, contamination with DNA from another crime/reference sample and contamination with DNA from an external source (e.g., forensic specialists from the police, manufacturers of reagents or consumables).

We see an increase in the total number of contaminations over the years. This may be explained in several ways: the number of analyses has increased; in 2011 a new more sensitive DNA analytical system was introduced (Next Generation Multiplex, NGM); the 'DNA elimination database' containing reference profiles of people (internal and external) involved in the handling

Table 5
Quality issue notifications (QINs) concerning contamination in the years 2008–2012: contamination sources.

	2008	2009	2010	2011	2012
Internal contamination with DNA from a staff member	21	18	17	26	53
Contamination other (sample)	29	40	50	108	84
External DNA contamination	3	0	5	24	22
Total	53	58	72	158	159

of DNA samples grew, so contaminations were recognized as such more often; the use of enhancement techniques, i.e., low template DNA analysis by hyperamplification and enhanced electrophoresis settings [23] has increased; and finally, we have noticed an increase in the number of requests for analyzing contact traces with minimal amounts of DNA. On the other hand the registration of spurious contaminations in blank samples was discontinued in 2012.

An important question concerns the source of the contamination: "who or what contaminated the sample?". Of course, this is impossible to establish with certainty, but an 'identification' can be made based on the DNA profile, and knowledge about persons and samples or objects that may have been in (in)direct contact with the samples. We distinguish three different categories:

1. contaminations where the source was identified as an NFI employee
2. all other contaminations, including contaminations where the source was identified as another sample, and contaminations where the source could not be identified
3. contaminations where the source was identified as external (outside NFI)

We note that most contaminations fall into the "other" category, i.e., they cannot be identified as an external source or as an NFI employee. Furthermore we see an increase in the number of contaminations in all categories. It is unknown whether the three explanations mentioned above (increasing number of analyses, introduction of NGM, and growing 'DNA elimination database'), fully account for this.

Another important question concerns the object that was contaminated. Here, we distinguished four categories:

1. Control samples: these samples are not related to a criminal case but are processed simultaneously with the crime samples for quality assurance purposes. Reagent blanks (samples containing no DNA), negative controls (i.e., female control DNA in Y chromosome STR analysis), and samples with known genetic composition are used as controls.
2. Reference samples: DNA samples that are taken directly from a person. The standard procedure is to take four different buccal swabs from the person involved. One swab for routine DNA analysis, one swab to allow for reanalysis (mainly QC purposes) one swab to allow a counter expertise and one swab for back up and additional analysis (i.e., analysis of extra autosomal loci, Y-chromosomal loci and/or mitochondrial DNA typing).
3. Crime samples: these are the biological traces that are collected for a forensic investigation in a criminal case.
4. Wipe samples: these samples are not related to a criminal case but are processed apart from the crime samples for quality assurance purposes. They are taken from tables and tools such as scissors and tweezers that are used in collecting and securing the crime samples.

Table 6
Quality issue notifications (QINs) concerning contamination in the years 2008–2012: contaminated objects.

	2008	2009	2010	2011	2012
Contamination in control (blank, negative and positive control)	23	28	39	102	47
Contamination in a reference sample	9	5	6	8	39
Contamination in a crime sample	20	23	18	46	71
Contamination in wipe sample (bench monitoring)	1	2	9	2	2
Total	53	58	72	158	159

Table 7
Quality issue notifications (QINs) concerning contamination in control samples in the years 2008–2012: extent of contamination.

	2008	2009	2010	2011	2012
Sporadic contamination	9	8	17	74	19
Gross contamination (origin identified)	13	12	18	24	21
Gross contamination (origin not identified)	1	8	4	4	7
Total number of contaminations in controls	23	28	39	102	47

Table 6 lists the observed frequencies for each type of object over the years 2008–2012.

We see that most QINs concern a contamination in a control sample. The relevant question for these contaminations is whether the crime sample was also contaminated. When a control or wipe sample is contaminated, all samples that were simultaneously analyzed are screened for contamination. Also, the cause of the contamination is investigated and additional analyses may be conducted. A contamination in a reference sample can easily be resolved by analysing the back-up samples.

When a crime sample is contaminated, the extent of the contamination is utterly important. When a relatively large amount of contaminating DNA was transferred, the DNA profile of the crime sample is mixed with the contaminant profile and may even hide the actual DNA-profile from the crime stain. This is called "gross contamination". The origin of the contamination can sometimes be identified, which eases the interpretation. Comparisons with other profiles are hampered but are often still possible; however the evidential value will suffer. New crime samples may also be taken when possible, but nevertheless a crucial piece of evidence may have been destroyed.

When a tiny amount of DNA was transferred, the DNA profile of the crime sample is mixed with a few allelic peaks of low intensity of the contaminant profile(s) from one or more individuals. This is called sporadic contamination. It is impossible to reliably identify the source(s) of such contaminations, since only a few alleles are visible. Unless the crime sample itself contained only tiny amounts of DNA, the DNA profile of the crime sample is usually clearly distinguished.

For the control samples, it is possible to classify the QINs according to the amount of contamination and identification of the origin. The numbers in the three different categories are presented in Table 7.

3.3. Consequences

For each QIN in the g-category above, the potential and actual consequences for the conclusions of the NFI report were assessed as described in Section 2.2. Table 8 presents the assessed potential

Table 8
Potential consequences of quality issue notifications (QINs) of type g: other (NFI related) in the years 2008–2012.

	2008	2009	2010	2011	2012
0. No adverse outcome	39	22	78	158	125
1. Repairable	144	197	138	155	137
2. Irreversible	47	51	65	42	81
Ongoing	0	0	0	0	3
Total	230	270	281	355	346

Table 9
Actual consequences of quality issue notifications (QINs) of type g: other (NFI related) for which the potential consequences were irreversible, at the NFI in the years 2008–2012.

	2008	2009	2010	2011	2012
0. Without adverse outcome	22	16	36	20	17
1. Adverse outcome; failure corrected; revised forensic report	0	0	3	0	1
2. Adverse outcome; irreversible; stated in the forensic report	21	32	23	22	60
3. Adverse outcome; irreversible; revised forensic report	4	3	3	0	0
4. Actual impact unknown	0	0	0	0	3
Total irreversible	47	51	65	42	81

consequences for the years 2008–2012.We see a strong increase in the number of QINs with no potential adverse outcome in 2011. This is mostly caused by an increase in the number of sporadic contaminations in control samples (see Section 3.2). We see that about 20% of QINs potentially has irreversible consequences. The sharp increase in the number of irreversible consequences in 2012 (from 42 in 2011 to 81 in 2012) is mainly explained by the increase in the number of contaminations in trace samples.Table 9 presents the assessed actual consequences for these QINs.

A considerable number of QINs with potential irreversible consequences actually had no adverse outcome. Many of these QINs concerned technical difficulties. Also many QINs did have consequences for the conclusions of the NFI report. This was relatively often caused by a contamination by an NFI employee in a crime sample (see Section 3.2). The sharp rise in the number of registered contaminations (Table 6) increased the number of notifications with adverse outcome. Many of such QINs were discovered before the report was issued.

Although the destruction of trace material or a decrease of evidential value of a trace is a serious failure, they do not result in misleading evidence. False matches (declaring a match when in fact the person did not contribute DNA to the sample: type 1 error) and false non-matches (declaring an exclusion when in fact the person did contribute a measurable amount of DNA to the sample: type 2 error) are misleading evidence, and therefore the most serious type of error that can be made. Quality procedures are targeted specifically to prevent such errors, but sometimes all defences fail and a false (non) match is reported. Table 10 shows the number of times this happened in the years 2008–2012. So far the NFI has not systematically registered near match/non-match errors. A near match/non-match error is defined as an event that has the potential to lead to the reporting of a wrongful match/non-match but does not because of chance prevention at the final state of the forensic DNA analysis procedure: the preparation of the forensic testimony. Consideration should therefore be given to establish a system that encourages reporting of near misses and actual errors.We see that despite all efforts to avoid crucial errors, they still occur at rare occasions. Misidentifications from the past that are the result of wrong registration of the DNA profiles are often discovered years later but registered in the year of discovery. In this respect it is relevant to understand that the root cause of a number of registered misidentifications in 2012 was in earlier years.Errors can occur at various phases in the DNA typing process. We distinguish between:

Table 10
Error rates by the NFI in the years 2008–2012: type 1 error (false DNA match) and type 2 error (false DNA non-match) that were reported and where the error was notified by internal control or by external (non-NFI) persons.

	2008	2009	2010	2011	2012	Total
Type 1 error	2	1	0	0	0	3
Type 2 error	4	3	1	2	4	14
Type 1 and type 2 errors	0	2	0	0	2	4
Total	6	6	1	2	6	21

Table 11
Process phase that caused type 1 and type 2 errors made by the NFI in the years 2008–2012.

	2008	2009	2010	2011	2012	Total
Pre-analytical	1	0	0	0	1	2
Analytical	0	3	0	1	2	6
Post-analytical	5	3	1	1	3	13
Total	6	6	1	2	6	21

- Pre-analytical (incidents that occurred before the actual investigation commenced at the department of HBS, i.e., the exhibit was incorrectly packaged or labelled)
- Analytical, occurring within the laboratory
- Post-analytical, whereby a correct DNA analysis result is generated but is incorrectly evaluated or reported, i.e., wrongful statistical calculation of the evidential value of a DNA match. Another example is clerical errors in DNA-profiles that were added to the CODIS DNA database to be searched. This type of error in the post-analytical phase has caused the majority of wrongful reported non-matches in Table 10

The phases in the analytical process where the crucial failures of Table 10 were observed are shown in Table 11.

Most errors are made in the post-analytical phase.

The four combined type 1 and type 2 errors (Table 10) that have been registered over the years 2008–2012 were all due to sample mix-up. In three instances there was a mix-up of crime samples, in one instance there had been mix-up of reference samples.

The fourteen type 2 errors were due to one occasion of reference sample mix-up, seven occasions concerned incorrect reporting of the results by the forensic DNA-expert and in one occasion a clerical error was made during import of the DNA-profile in the CODIS DNA database. Five type 2 errors were due to wrongful interpretation of the DNA-typing results by the CODIS DNA database staff.

The three type 1 errors were caused by one occasion of reference sample mix-up, one occasion of crime sample mix-up and one occasion of incorrect reporting to the Criminal Justice System.

This makes clear that despite automation of the analytical process (DNA extraction, amplification and analysis) the post analytical process is still prone to errors mainly because many non-automated steps take place in this part of the analysis process.

4. Dealing with error rates in the context of a legal case

Many authors have noted that in a specific legal case, the possibility of errors affects the evidential value of the DNA profile comparison [9–12,14,15]. Although all agree that it is an important issue, opinions still diverge on how to deal with this in practice. More specifically, if a forensic scientist wants to address the issue in a legal case he or she needs to decide on (1) whether or not to report error rates, and if so, what type of error rates (2) whether or not such rate(s) should be combined with the profile frequency, and if so, how (3) whether to guide the court in their interpretation, and if so, how. These issues have been addressed in various papers and discussed in various platforms [24]. We will not reiterate the arguments here, but instead describe and motivate a pragmatic approach.

4.1. Should error rates be reported?

Let us first consider two hypothetical murder cases. In case 1, there are numerous DNA samples taken from clothes, skin, fingernails and both semen and blood stains. They all match with the suspect, and all matches are confirmed by a counter-laboratory. In such a case, it would be a waste of time to consider the possibility of errors such as contamination or analytical error, and the investigation should focus on the probability of transfer of the suspect's DNA in the guilty and innocent scenarios. In case 2, we have a single sample containing a tiny amount of DNA that matches the suspect's profile. Now the possibility of various types of error is a relevant issue, and the criminal justice system should be informed about this.

The point is that in both cases it would be essentially meaningless and potentially misleading to report general error rates from proficiency testing, as suggested by, e.g., Koehler [9], or from the lab's casework, as in Section 3. In the first case, error is an academic discussion, and in the second, the general numbers are not representative for the specific circumstances of the case. The second NRC report [15] lists a number of reasons why quoting such numbers in a case report is problematic and states (pp. 85–86): "The question to be decided is not the general error rate for a laboratory or laboratories over time but rather whether the laboratory doing DNA testing in this particular case made a critical error." Balding [25] further specifies: "What matters are not probabilities of *any* profiling or handling errors, but only the probabilities of errors which could have led to the observed DNA profile match." Furthermore, we should take into account that the match has been reported to the court, and has not been identified as erroneous. To summarize: only case-specific rates of undetected errors are relevant, and only in certain cases.

In theory, it is possible for an expert to list such case specific undetected error rates and assess at least an order of magnitude. For example, the expert could assess the probability of contamination between two samples by checking whether they have ever been in (in)direct contact, taking the quality system of the chain of custody into account. Samples can also be retested to rule out certain sources of error. In practice however, investigating the processing of the samples in detail is an enormous lot of work that simply is only feasible and relevant in a very limited number of cases. Hence, it is important to select these cases carefully. Obviously, a detailed investigation concerning the possibility of error should be performed when it is deemed relevant by the court. Furthermore, there may be other external or internal signals that something may be wrong. These signals should be taken seriously, and also lead to a detailed investigation.

However, this is a circular problem, since such signals may not be given if people are not alerted to the possibility of errors somehow. Therefore, it is also important to provide general information, provided that it is clear that the particular case may deviate from the general picture. The NFI has provided professionals in the criminal justice system with general information on errors in forensic DNA analysis in the NFI sourcebook on forensic DNA typing [26]. The error issue is further discussed in papers [27,28] and in courses given to lawyers and police. Figures on error rates can be downloaded from the NFI website (http://www.nederlandsforensischinstituut.nl/over_het_nfi/organisatie/kwaliteit/kwaliteitsrapportages.aspx) and are provided in yearly reports of the national DNA database. Through these communication channels, all professionals that are involved in the criminal case are encouraged to contact the DNA expert if they have any reason to believe that an error was made.

We disagree with the NRC report [15] and Budowle et al. [20] who claims that "Errors of consequence due to mistakes or bias, which are a serious concern for all, are identified and addressed best through peer review by retesting, reanalysis, and/or blind verification." We agree with Balding [25] that such peer review does not "eliminate the need to assess possible handling errors before the samples are divided." Furthermore, such peer review is very rare. Our pragmatic solution is to inform the criminal justice system on general error rates through non-case related communication channels, and to provide case specific error assessments only when requested by the court or when there are external or internal signals that something may be wrong.

4.2. Should error rates be incorporated in an overall evidential value?

An interesting question is whether the expert should not only provide the necessary information about error rates, but also combine the information with the profiles rarity, or provide guidance on how the court could combine it. Thompson [10], Thompson et al. [11], and Balding [29] show how the various probabilities can logically be combined in a single number, the likelihood ratio that can be interpreted as a measure of evidential value. Taroni et al. [30, p. 169] and Fenton et al. [31] present a Bayesian network approach, that is very flexible and can be extended to include various sources of error. Thus, provided that we have identified the possible sources of error that are relevant for the case at hand that we have reliable probability assessments for them, and that we know their correlation structure, it is very well possible to incorporate all this into a single numerical likelihood ratio.

The question is to what extent the conditions are met, and whether this is desirable anyway. The NRC report [15] mentions legal obstacles, apart from the problems of obtaining suitable error rates mentioned above. The conclusion is (p. 87): "For all those reasons, we believe that a calculation that combines error rates with match probabilities is inappropriate. The risk of error is properly considered case by case, taking into account the record of the laboratory performing the tests, the extent of redundancy, and the overall quality of the results." Balding [29] notes that the likelihood ratio for the DNA evidence also depends on, for instance, the probability that the evidence was planted. He concludes (p. 99) "there seems to be no role for a forensic scientist to predict whether a juror might pursue such a line of reasoning, and hence the only option is to supply the juror with a match probability but also try to convey an understanding of the circumstances under which the match probability would be effectively irrelevant." We agree with these arguments, and report information on errors separately.

4.3. How to inform the court?

Balding [25] criticizes the NRC report [15] for not giving "guidance on how experts should convey to jurors a fair assessment of the evidence in view of the possibilities of both "chance match" and "handling or laboratory error". Fenton et al. [31] argue that a major educational effort is necessary, and that Bayesian networks may be a useful tool for this.

In our opinion, the general information about errors that should be easily accessible should convey the following message:

- errors are rare but possible
- if there is any reason to believe an error was made please contact the DNA expert
- in some cases it may be a realistic scenario that the match was caused by an error, in others this may merely be an academic discussion
- general error rates can be found at the NFI website however they may not be representative for the case at hand

- the expert may provide a personal estimate of case specific error rates, however as this is a lot of work this should only be requested when deemed relevant
- errors are not the only issue to consider when interpreting DNA evidence.
- forensic statistics models have been developed for the evidential value of the observations, that take the various sources of uncertainties into account. This analysis can be requested from the NFI

This message is best conveyed through various communication channels in general publications and educational settings. In the context of a case errors are discussed only when requested.

5. Discussion

Forensic DNA casework is conducted worldwide in a large number of laboratories, both private companies and in institutes owned by the government. Quality procedures are in place in all laboratories, but the nature of the quality system varies a lot between the different labs. In particular, there are many forensic DNA laboratories that operate without a quality issue notification system like the one described in this paper. In our experience, such a system is extremely important for the detection and proper handling of errors. This is crucial in forensic casework that can have a major impact on people's lives. We therefore propose that the implementation of a quality issue notification system is necessary for any laboratory that is involved in forensic DNA casework.

Such system can only work in an optimal way, however, when there is a blame-free culture in the laboratory that extends to the police and the legal justice system. People have a natural tendency to hide their mistakes, and it is essential to create an atmosphere where there are no adverse personal consequences when mistakes are reported. The management should take the lead in this culture change. Quality failures should be considered as a normal part of the system that should be registered and handled properly without the need to blame someone for it. At the NFI, the management has succeeded in creating such a culture. This is a major step in the culture shift that is necessary in a state of the art forensic science (DNA) laboratory [22].

Like Mnookin, Thompson [5] and many others, we call for the disclosure of data on error rates. As argued above, such information is important, even if it is not directly relevant to the criminal case at hand. The raw data of the quality issue notification system can be downloaded (in Dutch) at: http://www.nederlandsforensischinstituut.nl/over_het_nfi/organisatie/kwaliteit/kwaliteitsrapportages.aspx. Moreover, the yearly reports of the Dutch Accreditation Council can be downloaded here.

As far as we know, the NFI is the first forensic DNA laboratory in the world to reveal such detailed data and reports. It shows that this is possible without any disasters or abuse happening, and there are no reasons for nondisclosure. As mentioned in the introduction, in laboratory medicine publication of data on error rates has become standard practice. Quality failure rates in this domain are comparable to ours.

As mentioned in the introduction, the assessment of potential impact was done by the quality control manager in cooperation with a senior DNA expert, and is approved by the head of the department. One can criticize this as not being impartial, because these people can have an interest in downscaling the potential impact. To avoid this, the blame-free culture mentioned above is again important. Moreover, we do not see a practical alternative way to obtain the assessment with the same degree of know-how and accuracy.

Bayesian networks seem a promising tool to analyze the effect of uncertainties like errors on the evidential value of a DNA match. Le et al. [32] show that they may also be useful for the detection of errors. Bayesian networks are thus very flexible probabilistic models that are potentially also very effective in communicating a probabilistic analysis to lay persons. In our limited experience, this does not require major training efforts. First showing a very simple example like a coin tossed several times convinces most people that the tool works correctly and according to their intuition. Subsequently showing a very simple forensic problem like a matching blood type between stain and suspect convinces them it also works for practical forensic problems. They can then follow the line of reasoning with more complex situations rather easily, and much better than with formulas or complicated schemes. A different question is whether such ultimate transparent reasoning serves the needs of the court. Anyway, we consider the use of these networks an important topic for further research.

6. Closing comment

The main objective of continuously scrutinizing the QIN's of the NFI is to reduce the error rate of the forensic DNA analysis process. The data however may give the impression that in-depth monitoring of QIN's and the process of continuous improvements in the process have not led to any significant quality improvements at the NFI. The long-term lack of an effective reduction of the number of registered QIN's has a number of feasible and practical reasons.

1. There is a yearly increase in the number of DNA samples that has been processed.
2. There has been a move from the SGM plus to the far more sensitive NGM analysis system which has led to a sharp increase in the number of contaminations.
3. There is a continuous increase in the total and relative number of low template DNA samples that are requested for DNA analysis. Low template DNA samples are more prone to contamination than traditional traces from blood, semen and saliva.
4. Inherent to the QIN notification system of the NFI is that errors are registered in the year of discovery. A number of recent registrations therefore concern (samples from) cases from the past.
5. The growth of the elimination database with the DNA profiles from NFI staff and forensic workers from the police. This has resulted in the discovery and subsequent registration of a number of contaminations from the past.
6. In spite of the fact that new techniques and methods are extensively validated before introduction in actual casework the switch to a different platform is sometimes associated with a rise in the number of registrations. For example, the introduction of automated DNA extraction from reference samples, a technology that is developed to assist in error reduction, was associated with a relatively high number of QIN's.
7. Changes in the criteria for registrations of QIN's. For example, minor administrative errors in NFI DNA reports were not recorded in the past. From 2010 these errors are however registered.

The QIN monitoring however has brought a number of major improvements in the DNA typing process. For example when a relatively large number of sample mix-ups were noticed the NFI implemented large scale automation of the DNA extraction and typing process. The registration made clear at an early stage that the introduction of a new DNA typing system (NGM) led to a higher number of contaminations. Immediate measures were taken to reduce the number of contaminations, such as improvement of bench cleaning. In addition the GeneMapper® ID-X Software version 1.1.1 profile automated comparison tool was implemented to scrutinize batches of samples for DNA sample to sample carry over and contamination. Hence the QIN monitoring system has become an important tool to initiate preventive measures at an early stage in order to minimize error rates in forensic DNA analysis.

Acknowledgements

The authors wish to thank Maartje Elderenbosch for her work in the evaluation of the data. Thanks are also due to Prof. dr. Arian van Asten for his suggestions on the collection of the data.

References

[1] M. Possley, S. Mills, F. McRoberts, Scandal Touches Even Elite Labs: Flawed Work, Resistance to Scrutiny Seen Across U.S., Tribune, Chicago, 2004, Oct. 21, 2004.
[2] NFI, DNA-onderzoek Zuuk, NFI, 2004, http://www.om.nl/algemene_onderdelen/uitgebreid_zoeken/@133442/dna-onderzoek_zuuk/.
[3] P.J. Koppen, H. van en Elffers, De mythe van het DNA-bewijs, Advocatenblad 86 (2006) 607–618.
[4] A. Carracedo, P.M. Schneider, J. Butler, M. Prinz, Analysis and biostatistical interpretation of complex and low template DNA samples (editorial), Forensic Sci. Int. Genet. 6 (December (6)) (2012) 677–678. , http://dx.doi.org/10.1016/j.fsigen.2012.08.010 (Epub 2012 Sep 19).
[5] W.C. Thompson, The potential for error in forensic DNA testing (and how that complicates the use of DNA databases for criminal identification), Gene Watch 21 (3–4) (2008), http://www.gene-watch.org.
[6] M. Lynch, S. Cole, R. McNally, K. Jordan, Truth Machine: The Contentious History of DNA Fingerprinting, University of Chicago Press, Chicago, 2008.
[7] M. Lynch, Science, truth, and forensic cultures: the exceptional legal status of DNA evidence, Stud. History Philos. Biol. Biomed. Sci. 44 (2013) 60–70.
[8] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
[9] J.J. Koehler, Error and exaggeration in the presentation of DNA evidence, Jurimetrics 34 (1993) 21–39.
[10] W.C. Thompson, Subjective interpretation, laboratory error and the value of forensic DNA evidence: three case studies, Genetica 96 (1995) 153–168.
[11] W.C. Thompson, F. Taroni, C.G.G. Aitken, How the probability of a false positive affects the value of DNA evidence, J. Forensic Sci. 48 (1) (2003) 47–54.
[12] M.J. Saks, J.J. Koehler, The individualization fallacy in forensic science evidence, Vanderbilt Law Rev. 61 (1) (2008) 199–219.
[13] B. Schiffer, The Relationship between Forensic Science and Judicial Error: A Study Covering Error Sources, Bias, and Remedies, (PhD thesis), University of Lausanne, 2009, http://www.unil.ch/webdav/site/esc/shared/These_Schiffer.pdf.
[14] National Research Council, DNA Technology in Forensic Science, National Academy Press, Washington, DC, 1992, pp. 88–89.
[15] National Research Council, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996 (Chapter 3).
[16] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, National Academy Press, Washington, DC, 2009.
[17] R. Lapworth, T.K. Teal, Laboratory blunders revisited, Ann. Clin. Biochem. 31 (1994) 78–84.
[18] M. O'Kane, The reporting, classification and grading of quality failures in the medical laboratory, Clin. Chim. Acta 404 (2009) 28–31.
[19] I. Kindt, R. Huijgen, M. Boekel, K.J. van der Gaag, J.C. Defesche, J.P.J. Kastelein, P. de Knijff, Quality assessment of the genetic test for familial hypercholesterolemia in The Netherlands, Cholesterol (2013), http://dx.doi.org/10.1155/2013/531658.
[20] B. Budowle, M.C. Bottrel, S.G. Bunch, R. Fram, D. Harrison, S. Meagher, C.T. Oien, P.E. Peterson, D.P. Seiger, M.B. Smith, M.A. Smrz, G.L. Soltis, R.B. Stacey, A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement, J. Forensic Sci. 54 (2009) 798–809.
[21] R. Ansell, Internal quality control in forensic DNA analysis, Accred. Qual. Assur. 18 (4) (2013) 279–289.
[22] J.L. Mnookin, S.A. Cole, I.E. Dror, B.A.J. Fisher, M.M. Houck, K. Inman, D.H. Kaye, J.J. Koehler, G. Langenburg, M.D. Risinger, N. Rudin, J. Siegel, D.A. Stoney, The need for a research culture in the forensic sciences, UCLA Law Rev. 58 (2011) 725–779.
[23] A.A. Westen, J.H. Nagel, C.C. Benschop, N.E. Weiler, B.J. de Jong, T. Sijen, Higher capillary electrophoresis injection settings as an efficient approach to increase the sensitivity of STR typing, J. Forensic Sci. 54 (2009) 591–598.
[24] R. Harmon, B. Budowle, Questions about forensic science, Science 311 (2006) p607.
[25] D.J. Balding, Error and misunderstandings in the second NRC report, Jurimetrics 37 (1997) 469–476.
[26] A.J. Meulenbroek, De Essenties van forensisch biologisch onderzoek; Humane biologische sporen en DNA, Zutphen, Paris, 2009.
[27] A.J. Meulenbroek, Leidraad en praktische handvatten voor de jurist bij het doorgronden van conclusies forensisch DNA-onderzoek, Expertise en Recht 3 (2011) 73–90.
[28] C.P. van der Beek, A.D. Kloosterman, M.J. Sjerps, De detectie van vals positieve en de preventie van vals negatieve matches bij grootschalige DNA-databankvergelijkingen, Expertise en Recht 6 (2011) 219–221.
[29] D.J.J. Balding, Weight-of-Evidence for Forensic DNA Profiles, John Wiley & Sons, Ltd., Chichester, 2005.
[30] F. Taroni, C.G.G. Aitken, P. Garbolino, A. Biedermann, Bayesian Networks and Probabilistic Inference in Forensic Science, Wiley, Chichester, 2006.
[31] N.E. Fenton, M. Neil, A. Hsu, Calculating and understanding the value of any type of match evidence when there are potential testing errors, Artif. Intell. Law 22 (2014) 1–28.
[32] Q.A. Le, G. Strylewicz, J.N. Doctor, Detecting blood laboratory errors using a Bayesian network: an evaluation on liver enzyme tests, Med. Decis. Making 31 (2011) 325–337.
[33] M. Plebani, P. Carraro, Mistakes in a stat laboratory: types and frequency, Clin. Chem. 43 (1997) 1348–1351.
[34] M. Stahl, E.D. Lund, I. Brandslund, Reasons for a laboratory's inability to report results for requested analytical tests, Clin. Chem. 44 (1998) 2195–2197.
[35] W.T. Hofgärtner, J.F. Tait, Frequency of problems during clinical molecular-genetic testing, Am. J. Clin. Pathol. 112 (1999) 14–21.
[36] P. Carraro, M. Plebani, Errors in a stat laboratory: types and frequencies 10 years later, Clin. Chem. 53 (2007) 1338–1342.