1. Title Page

Title	TARGET-EU: Tolvaptan and Risk Associated to Hepatotoxicity in Autosomal Dominant Polycystic Kidney Disease
Research question & Objectives	The main objective of the trial is to investigate if there is an increase in the risk of hepatotoxicity with tolvaptan use among patients with ADPKD compared to those who are unexposed to tolvaptan.
Protocol version	V1.0
Last update date	4 March 2026
Contributors	**Primary investigator contact information:** Dr. Maarten Lambert m.lambert@uu.nl **Contributor names:** Dr. Daniala Weir Dr. Patrick Souverein Prof. Dr. Helga Gardarsdottir Prof. Dr. Olaf Klungel Prof. Dr. Ian Douglas
Study registration	**Site:** https://catalogues.ema.europa.eu/node/4440/administrative-details **Identifier:** EUPAS1000000539
Sponsor	**Organization:** EU PE&PV research network **Contact:** eupepv@uu.nl
Conflict of interest	None

Table of contents

1. Title Page **1**
2. Abstract **4**
3. Amendments and updates **5**
4. Milestones **5**
5. Rationale and background **5**
6. Research questions and objectives **7**
6.1 Primary Estimand 1 7
6.2 Supplementary Estimand 2 10
6.3 Supplementary Estimand 3 12
7. Research methods **14**
7.1 Study design 14
7.2 Study design diagram 16
7.3 Setting 17
7.3.1 Definition of time 0 (and other primary time anchors) 17
7.3.2 Context and rationale for study inclusion criteria: 18
7.3.3 Context and rationale for study exclusion criteria 19
7.4 Variables 20
7.4.1 Context and rationale for exposure(s) of interest 20
7.4.2 Context and rationale for outcome(s) of interest 20
7.4.3 Context and rationale for follow up 22
7.4.4 Context and rationale for covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications) 24
7.5 Core Emulation Table - Design Summary 25
7.6 Data analysis 34
7.6.1 Analysis Plan 34
7.6.2 Primary Estimand (1) Analysis 35
7.6.3 Supplemental Estimand (2) Analysis 39
7.6.4 Supplemental Estimand (3) Analysis 43
7.6.5. Sensitivity Analyses 47
7.6.6. Other Supplemental Analyses 51

7.6.7 Core Emulation Table – Estimation Summary 51
7.7 Data sources 56
7.7.1 Data sources and Quality 56
7.8 Data management 58
7.9 Quality control 61
7.10 Study precision 68
8. Limitation of the methods 72
9. Protection of human subjects 74
10. Reporting of adverse events 75
11. References 76
12. Appendices 77

2. Abstract

Background: Autosomal Dominant Polycystic Kidney Disease (ADPKD) is a chronic, inherited systemic disorder with manifestations that occur mainly renally, but also in other organs. Tolvaptan is the only approved drug for treatment of ADPKD in adults who are at risk of a rapidly progressive phenotype. Tolvaptan is a vasopressin V2 receptor agonist that blocks vasopressin signalling, which is an important driver of kidney cyst growth in ADPKD. A potentially serious side effect of tolvaptan is hepatotoxicity. In several clinical trials the risk of hepatotoxicity has been reported, and special warnings and safety measures have been included in the label.
Objectives: The current study aims to investigate the safety of tolvaptan use through assessing the risk of hepatotoxicity associated with tolvaptan treatment in ADPKD patients compared to the risk in patients who are unexposed to tolvaptan.
Methods: We will conduct a cohort study with a prevalent new-user design using linked electronic health records from the UK (CPRD) from 2015 until the latest available data. Eligible individuals are adults (≥ 18 years) with ADPKD who, for the intervention group, initiated tolvaptan treatment since market authorisation or, for the control group, who had a time-matched visit to a physician without a tolvaptan prescription. In the primary estimand, a treatment policy strategy is used for treatment discontinuation, a hypothetical strategy is used for tolvaptan initiation in the control group or the use of medication with hepatotoxic effects, and a while alive strategy is used for all-cause mortality. The primary analysis uses a Cox proportional hazards model, with supplemental analyses using an accelerated failure time model to estimate restricted mean survival time (RMST) at 3 years. Sensitivity analyses will be conducted to assess the impact of outcome misclassification using probabilistic bias analysis and exploring alternative assumptions about censoring.

3. Amendments and updates

Version date	Version number	Section of protocol	Amendment or update	Reason
4 March 2026	V1.0			

4. Milestones

Milestone	Date
Study protocol for RWD study	08 August 2025
Preliminary results RWD study	April 2026
Final Study report	10 June 2026

5. Rationale and background

What is known about the condition:

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is a chronic, inherited systemic disorder with manifestations that occur mainly renally, but also in other organs. ADPKD is the fourth leading cause of end-stage kidney disease in adults. [1] In the United Kingdom, one person in every 1000 to 2500 people is estimated to have ADPKD, adding to between 30 000 and 70 000 people with the condition. [2] ADPKD is caused by mutations in the genes encoding plasma membrane-spanning polycystin 1 and polycystin 2. These polycystins regulate vascular development in many organs, including kidneys, brain, heart, liver, and pancreas. [3] Although the disease starts developing before birth, in many cases, the first signs of the disease may not present itself to people who are affected for multiple decades. Phenotypes are heterogeneous, meaning that patients may show symptoms and disease progression from mild to severe. Although there is no established definition of rapidly progressive disease, several biomarkers may predict the rate of progression, including eGFR and total kidney volume. [4]

What is known about the exposure of interest:

Tolvaptan is the only approved drug for treatment of ADPKD in adults who are at risk of a rapidly progressive phenotype. [4] Tolvaptan is a vasopressin V2 receptor agonist that blocks vasopressin signalling, which is an important driver of kidney cyst growth in ADPKD. The aim of treatment with tolvaptan is to slow the loss of kidney function to delay kidney failure. [5] The efficacy of tolvaptan has been shown in several clinical trials, with reductions in eGFR decline as most important endpoint. [6,7] Although these trials do not provide long-term data, a retrospective study followed tolvaptan users for up to 11 years, showed similar long-term results. [8]

Gaps in knowledge:

While the efficacy of tolvaptan has been well studied, the safety profile has not be investigated to the same extent. As ADPKD is a relatively rare disease and tolvaptan is only indicated in patients at risk for a rapidly progressive disease, data on the effects of tolvaptan use are limited. Especially for investigating uncommon side effects, the relatively low number of participants in clinical trials is a limitation. A potentially serious side effect of tolvaptan is hepatotoxicity. In several clinical trials the risk of hepatotoxicity has been reported, although the extent to which it occurs varies from less than 1% to more than 4%. [7,9,10] . Although current safety measures are in place to identify hepatotoxicity as a side effect of tolvaptan treatment, better quantification of the incidence and development of hepatotoxicity in tolvaptan users may improve the benefit-risk assessment by clinicians before treatment initiation, improve accuracy of patient counselling, change hepatotoxicity testing burden and resource allocation, and provide more tailored care for specific subgroups.

What is the expected contribution of this study?

The current study aims to investigate the safety of tolvaptan use through assessing the risk of hepatotoxicity associated with tolvaptan treatment in ADPKD patients compared to the risk in patients who are unexposed to tolvaptan. By conducting a retrospective observational study, safety effects can be studied on a longer term than possible in clinical trials. Moreover, including all tolvaptan users since the introduction of tolvaptan to the market increases the ability to quantify the incidence of rare adverse events.

6. Research questions and objectives

The overall aim is to quantify the risk of hepatotoxicity in ADPKD patients who are treated with tolvaptan compared to patients who do not use tolvaptan.

6.1 Primary Estimand 1

Research question answered by the estimand: What is the hazard ratio of hepatotoxicity for tolvaptan use versus no tolvaptan use in adults with Autosomal Dominant Polycystic Kidney Disease, while alive, regardless of treatment discontinuation and in the absence of tolvaptan for patients assigned to the unexposed group and the absence of hepatotoxic medication in both groups?

Table 1. Primary Estimand (Estimand 1)

	Target Trial	Target Trial Emulation	Comment
Population	Adults with ADPKD	Adults with ADPKD	
Treatment Conditions	Tolvaptan vs. unexposed (i.e. treated with symptomatic treatments or untreated)	Tolvaptan vs. unexposed (i.e. treated with symptomatic treatments or untreated)	Exposure is defined based on first observed prescription within treatment episode
Endpoint	Time to first occurrence of hepatotoxicity, defined as the first occurrence of any of the following events: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN	Time to first occurrence of hepatotoxicity. Hepatotoxicity will be defined as the first registration of an ICD-10, Read, or SNOMED code related to hepatotoxicity or a recorded increase in the any of the following liver enzymes (if data are available): • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN + elevated gamma-glutamyl transpeptidase	Although the endpoint is the same, hepatotoxicity will be identified differently in the emulation compared to the target trial. It is expected that, at least in the unexposed group, liver enzyme levels are not registered regularly. In the emulation, hepatotoxicity will therefore be identified through increases in liver enzymes with respect to the last available values prior to the index date or the registration of ICD-10, Read, or SNOMED codes related to hepatotoxicity . Composite outcome increases the chance of capturing hepatotoxicity even when enzyme levels are not available.

		(GGT) and/or elevated 5'nucleotidase (5-NT)	Additionally, alkaline phosphatase elevation is specified in a more elaborate way in the emulation, but this is not a deviation from the target trial, where the same elaborate specification should have been used. Single elevation of alkaline phosphatase may not be a sufficiently specific biomarker for hepatotoxicity as it can also increase through various bone pathologies. However, alkaline phosphatase is a specific biomarker for hepatotoxicity if combined with gamma-glutamyl transpeptidase and/or 5'nucleotidase. If GGT and 5-NT levels are normal, elevated AP suggests bone disease. If GGT and/or 5-NT levels are elevated, elevated AP suggests hepatotoxicity.
Summary Measure	Hazard ratio	Hazard ratio	Target trial and emulation identical
Intercurrent events and strategies to handle them	Treatment discontinuation: treatment policy Tolvaptan initiation in control group: hypothetical Same for both treatment strategies: Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	Treatment discontinuation: treatment policy Tolvaptan initiation in control group: hypothetical Same for both treatment strategies: Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	Target trial and emulation identical Treatment policy reflects real-world effectiveness. Treatment discontinuation might be difficult to accurately identify in data source. Any mismeasurement of treatment discontinuation is not an issue as we are interested in the treatment effect regardless of whether the IE occurred. Hypothetical strategy reflects a hypothetical scenario in which the intercurrent event would not occur. To implement this strategy, we need to identify the occurrence (and date) of the IE and there might be measurement errors for the IE,

			which is important as we are not interested in outcomes after occurrence of the IE.

ADPKD: Autosomal Dominant Polycystic Kidney Disease, ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal

Rationale for handling of intercurrent events:

Treatment policy strategy for discontinuation: the strategy evaluates the risk of hepatotoxicity based on the treatment assigned regardless of whether participants stopped their treatment. As such, it estimates safety effects after initiating tolvaptan as they occur in clinical practice.
Hypothetical strategy for tolvaptan initiation in control group or use of any medication with known hepatoxicity effects: the strategy evaluates what the risk of hepatotoxicity would have been if tolvaptan had not been initiated in the unexposed group.
While-alive strategy for all-cause mortality. The interest lies in the risk of hepatotoxicity while patients are alive.

6.2 Supplementary Estimand 2

Research question answered by the estimand: What is the hazard ratio of hepatotoxicity for tolvaptan versus no treatment in adults with Autosomal Dominant Polycystic Kidney Disease, while alive and while on their initial treatment and in the absence of hepatotoxic medication in both groups?

Table 2. Estimand 2

	Target Trial	Target Trial Emulation	Comment
Population	Adults with ADPKD	Adults with ADPKD	
Treatment Conditions	Tolvaptan vs. unexposed (i.e. treated with symptomatic treatments, or untreated)	Tolvaptan vs. unexposed (i.e. treated with symptomatic treatments, or untreated)	Exposure is defined based on first observed prescription within treatment episode
Endpoint	Time to first occurrence of hepatotoxicity, defined as the first occurrence of any of the following events: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN	• Time to first occurrence of hepatotoxicity. • Hepatotoxicity will be defined as the first registration of an ICD-10, Read, or SNOMED code related to hepatotoxicity or a recorded increase in the any of the following liver enzymes (if data are available): • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN+ elevated gamma-glutamyl transpeptidase (GGT) and/or elevated 5'nucleotidase (5-NT)	Although the endpoint is the same, hepatotoxicity will be identified differently in the emulation compared to the target trial. It is expected that, at least in the unexposed group, liver enzyme levels are not registered regularly. In the emulation, hepatotoxicity will therefore be identified through increases in liver enzymes with respect to the last available values prior to the index date or the registration of ICD-10, Read, or SNOMED codes related to hepatotoxicity. Composite outcome increases the chance of capturing hepatotoxicity even when enzyme levels are not available. Additionally, alkaline phosphatase elevation is specified in a more elaborate way in the emulation, but this is not a deviation from the target trial, where the same elaborate specification should have been used. Single

			elevation of alkaline phosphatase may not be a sufficiently specific biomarker for hepatotoxicity as it can also increase through various bone pathologies. However, alkaline phosphatase is a specific biomarker for hepatotoxicity if combined with gamma-glutamyl transpeptidase and/or 5'nucleotidase. If GGT and 5-NT levels are normal, elevated AP suggests bone disease. If GGT and/or 5-NT levels are elevated, elevated AP suggests hepatotoxicity.
Summary Measure	Hazard ratio	Hazard ratio	Target trial and emulation identical
Intercurrent events and strategies to handle them	Treatment discontinuation: while on treatment Tolvaptan initiation in control group: while on treatment Same for both treatment strategies: Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	Treatment discontinuation: while on treatment Tolvaptan initiation in control group: while on treatment Same for both treatment strategies: Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	Target trial and emulation identical While on treatment strategy evaluates the risk of hepatotoxicity while patients are under the assigned treatment. For tolvaptan users this means while they actually use tolvaptan. Treatment discontinuation might be difficult to accurately identify in data source, which may lead to errors in IE measurement. Hypothetical strategy reflects a hypothetical scenario in which the intercurrent event would not occur. To implement this strategy, we need to identify the occurrence (and date) of the IE and there might be measurement errors for the IE, which is important as we are not interested in outcomes after occurrence of the IE.

ADPKD: Autosomal Dominant Polycystic Kidney Disease, ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal

Rationale for why selected strategies to handle intercurrent events are chosen

Hypothetical strategy for use of any medication with known hepatoxicity effects: the strategy evaluates what the risk of hepatotoxicity would have been if this medication had not been initiated.
While on treatment strategy for discontinuation and initiation of tolvaptan in the control group: evaluates the risk of hepatotoxicity while patients are under the assigned treatment. For tolvaptan users this means while they actually use their tolvaptan. This implies that the hepatotoxicity is of interest only if it occurs when participants use tolvaptan. Any data collected after discontinuation or dose changes are not relevant for this estimand and will not be included in the statistical analyses. Whenever tolvaptan is down titrated until stopping and restarted afterwards, only the first treatment episode is considered, and data afterwards are censored.
While-alive strategy for all-cause mortality: The interest lies in the risk of hepatotoxicity while patients are alive.

6.3 Supplementary Estimand 3

Research question answered by the estimand: What is the difference in restricted mean survival time to hepatotoxicity for tolvaptan use versus no-treatment in adults with Autosomal Dominant Polycystic Kidney Disease, while alive, regardless of treatment discontinuation and in the absence of tolvaptan for patients assigned to the unexposed group and the absence of hepatotoxic medication in both groups?

Table 3. Estimand 3

	Target Trial	Target Trial Emulation	Comment
Population	Adults with ADPKD	Adults with ADPKD	
Treatment Conditions	Tolvaptan vs. unexposed (i.e. treated with symptomatic treatments or untreated)	Tolvaptan vs. unexposed (i.e. treated with symptomatic treatments or untreated)	Exposure is defined based on first observed prescription within treatment episode
Endpoint	Time to first occurrence of hepatotoxicity, defined as the first occurrence of any of the following events: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN	• Time to first occurrence of hepatotoxicity. • Hepatotoxicity will be defined as the first registration of an ICD-10, Read, or SNOMED code related to hepatotoxicity or a recorded increase in the any of the following liver enzymes (if data are	Although the endpoint is the same, hepatotoxicity will be identified differently in the emulation compared to the target trial. It is expected that, at least in the unexposed group, liver enzyme levels are not registered regularly. In the emulation, hepatotoxicity will therefore be

	• Alkaline phosphatase > 2x ULN	available): • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN+ elevated gamma-glutamyl transpeptidase (GGT) and/or elevated 5'nucleotidase (5-NT)	identified through increases in liver enzymes with respect to the last available values prior to the index date or the registration of ICD-10, Read, or SNOMED codes related to hepatotoxicity. Composite outcome increases the chance of capturing hepatotoxicity even when enzyme levels are not available. Additionally, alkaline phosphatase elevation is specified in a more elaborate way the emulation, but this is not a deviation from the target trial, where the same elaborate specification should have been used. Single elevation of alkaline phosphatase may not be a sufficiently specific biomarker for hepatotoxicity as it can also increase through various bone pathologies. However, alkaline phosphatase is a specific biomarker for hepatotoxicity if combined with gamma-glutamyl transpeptidase and/or 5'nucleotidase. If GGT and 5-NT levels are normal, elevated AP suggests bone disease. If GGT and/or 5-NT levels are elevated, elevated AP suggests hepatotoxicity.
Summary Measure	Difference in restricted mean survival time	Difference in restricted mean survival time	Target trial and emulation identical
Intercurrent events and strategies to handle them	Treatment discontinuation: treatment policy Tolvaptan initiation in control group: hypothetical Same for both treatment strategies:	Treatment discontinuation: treatment policy Tolvaptan initiation in control group: hypothetical Same for both treatment strategies:	Target trial and emulation identical Treatment policy reflects real-world effectiveness. Treatment discontinuation might be difficult to accurately identify in data source. Any mismeasurement of treatment discontinuation is not an issue as we are not

	Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	interested in the treatment effect regardless of whether the IE occurred. Hypothetical strategy reflects a hypothetical scenario in which the intercurrent event would not occur. To implement this strategy, we need to identify the occurrence (and date) of the IE and there might be measurement errors for the IE, which is important as we are not interested in outcomes after occurrence of the IE.

ADPKD: Autosomal Dominant Polycystic Kidney Disease, ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal

Rationale for handling of intercurrent events: Equal to estimand 1.

7. Research methods

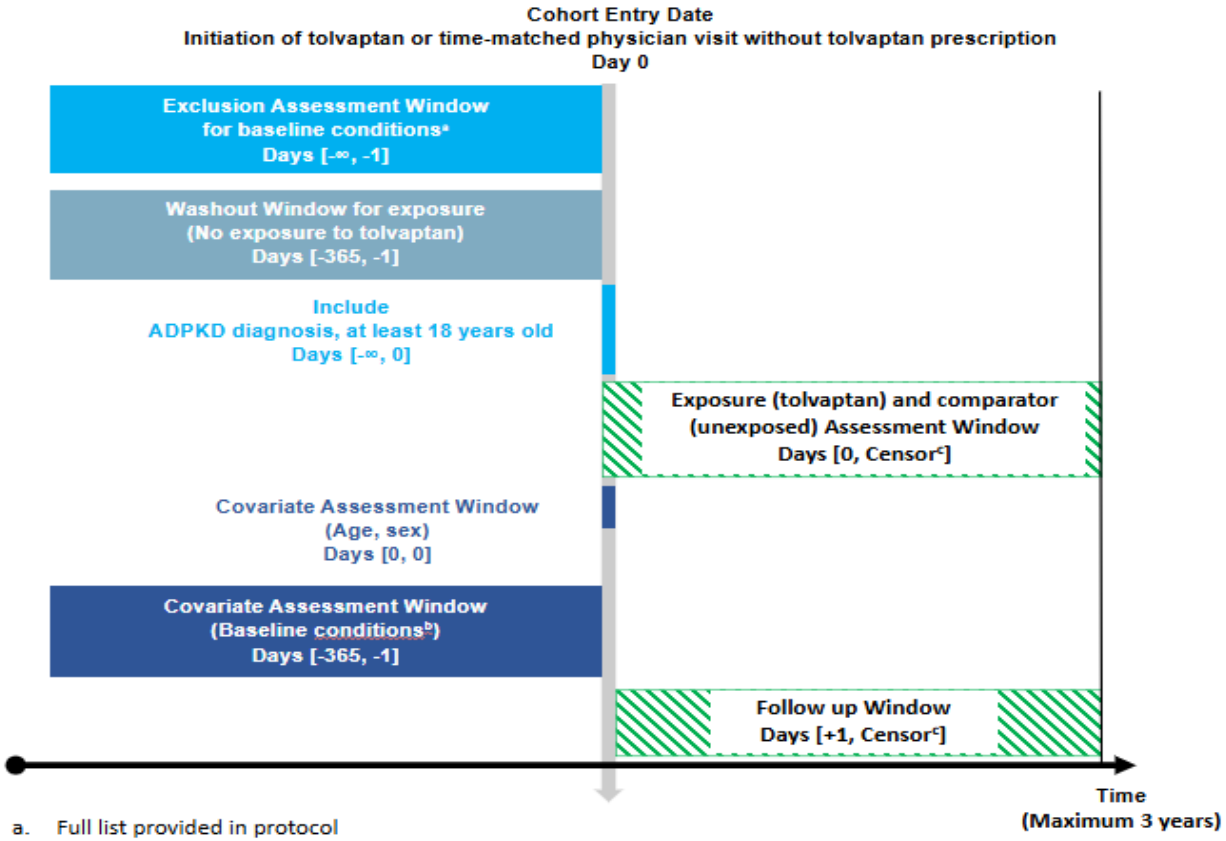
7.1 Study design

Research design (e.g. cohort, case-control, etc.): A cohort study with a prevalent new-user design

Rationale for study design choice: A pragmatic design with real-world data captures care as it is delivered in practice. Although the observational emulation cannot perfectly mimic randomisation in the target trial, several measures are taken to minimise bias and confounding. The prevalent new-user design avoids immortal time bias and confounding by disease severity which could be an issue in a new user study with no active comparator. Inclusion and exclusion criteria are picked to closely mirror those in the target trial. Additionally, calendar time and propensity score matching will increase the chances of having participants with similar characteristics in the intervention and control groups. Moreover, the prevalent new-user design aligns start of follow-up with treatment initiation. This ensures that all included patients are at risk of the outcome from a common time point. Inclusion of tolvaptan initiators of any dose reflects the hypothetical target trial and the dose and duration flexibility that is seen in routine care. The prevalent new-user design matches new tolvaptan users with ADPKD with time-matched (time from indication diagnosis) unexposed ADPKD patients, which is the closest way to mirroring a placebo-controlled RCT. There is no alternative treatment to tolvaptan to which the risk of hepatotoxicity could

be compared. Clinically it is most relevant to compare tolvaptan use to unexposed ADPKD patients (who only receive symptomatic care when appropriate but do not receive any treatment for ADPKD itself) as this is what patients receive in clinical practice.

7.2 Study design diagram



a. Full list provided in protocol
b. Full list provided in protocol.
c. Earliest of: outcome of interest (hepatotoxicity), death, disenrollment, end of the study period (3 years after treatment initiation). Censoring is estimand specific: no censoring after IE's for estimands 1&3, censoring after IE's for estimand 2.

ADPKD = Autosomal Dominant Polycystic Kidney Disease

7.3 Setting

This study is conducted using routinely collected electronic health records from 2015 to 2024, reflecting the period of tolvaptan use in routine clinical practice. The study is set primarily in primary care, drawing on longitudinal data from general practices with linkage to hospital data. Data are sourced from the United Kingdom (Clinical Practice Research Datalink [CPRD]), providing population-based and representative coverage of real-world clinical care.

7.3.1 Definition of time 0 (and other primary time anchors)

For the intervention group, time 0 is defined as the first prescription of tolvaptan. This is when participants enter the study, which mimics initiation of tolvaptan therapy in the target randomised controlled trial. For participants in the control group, time 0 is the date of a physician visit without a tolvaptan prescription, matched on time since ADPKD diagnosis to the date of a treatment receipt in the intervention group. Additionally, the control and intervention time 0 are also matched on the calendar time scale. The matching will not be on the exact value but on the quintile. In the first stage of matching, on calendar time, matching will be done 5 : 1 for control group: intervention group. This is necessary to end up with a 1:1 control: intervention group allocation for the second matching step on propensity score. Individuals can only be eligible as matched control once and can only be assigned to either the control group or the intervention group once (at time 0).

Table 4. Operational Definition of Time 0 (index date) and other primary time anchors

Study population name(s)	Time Anchor Description (e.g. time 0)	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Intervention: tolvaptan use	First prescription of tolvaptan	Single	Incident	[-365, -1]	OP	ATC	n/a	1 year washout period prior to first use	n/a	n/a
Comparator	Time-matched time of diagnosis until time of physician visit	Single	Incident/ prevalent	[-365, -1]	OP	n/a	n/a	1 year washout period for tolvaptan use is applied.	n/a	n/a

	without tolvaptan prescription									

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable

[2]See appendix for listing of clinical codes for each study parameter

7.3.2 Context and rationale for study inclusion criteria:

- Have a recorded confirmed diagnosis of Autosomal Dominant Polycystic Kidney Disease: the study focuses on the use of tolvaptan and risk of hepatotoxicity in patients with ADPKD specifically.
- Be at least 18 years old at time of recorded diagnosis: tolvaptan is only registered for use in adults

Table 5. Operational Definitions of Inclusion Criteria

Criterion	Details	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
ADPKD diagnosis	Any recorded ADPKD diagnosis	[∞, -1]	OP, IP	ICD-10, Read codes, SNOMED	primary	Intervention, comparator	n/a	n/a
At least 18 years old	At time 0	At time 0	n/a	n/a	n/a	Intervention, comparator	n/a	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)
ADPKD: Autosomal Dominant Polycystic Kidney Disease

7.3.3 Context and rationale for study exclusion criteria

Table 6. Operational definitions of exclusion criteria

All exclusion criteria are mirrored from the hypothetical target trial.

Criterion	Details	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Use of medication with hepatotoxic effects	A current prescription of any medication known for potentially causing hepatotoxic effects	[-90 days, 0]	OP, IP	ATC	n/a	Intervention, comparator	n/a	n/a
Hepatotoxicity	Diagnosis of hepatotoxicity identified by a relevant ICD-10, Read, or SNOMED code	[-365, 0]	OP, IP	ICD-10, Read, or SNOMED	Primary or secondary	Intervention, comparator	n/a	n/a
AST or ALT elevation	AST or ALT levels > 1.5 times ULN	At time 0	OP, IP	n/a	n/a	Intervention, comparator	n/a	n/a
Total bilirubin,alkaline phosphatase, GGT, 5-NT elevation	Total bilirubin, alkaline phosphatase, GGT, 5-NT levels higher than ULN	At time 0	OP, IP	n/a	n/a	Intervention, comparator	n/a	n/a
Use of diuretics	A current prescription of any diuretic	[-90 days, 0]	OP, IP	ATC	n/a	Intervention, comparator	n/a	n/a

ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal, GGT: gamma-glutamyl transpeptidase, 5-NT: 5'nucleotidase

7.4 Variables

7.4.1 Context and rationale for exposure(s) of interest

There is no alternative treatment to tolvaptan to which the risk of hepatotoxicity could be compared. Clinically it is most relevant to compare tolvaptan use to ADPKD patients unexposed to tolvaptan, who may receive symptomatic treatment, as per clinical practice. Any dose or regimen of tolvaptan is eligible for inclusion.

Algorithm to define duration of exposure effect:

Duration of prescriptions will be used to define exposure effect. Duration of a single prescription will be based on the prescribed amount and the dosing regimen. If refills occur before the predicted end of the prescription, we will add the overlapping days to the end of the treatment episode, if the number of days to be added is not more than 90 days. We will allow a gap of 30 days between the end of the prescription and the subsequent refill.

Table 7. Operational Definitions of Exposure

Exposure group name(s)	Details	Washout window	Assessment Window	Care Setting[1]	Code Type[2]	Diagnosis position[3]	Applied to study populations:	Incident with respect to…	Measurement characteristics/ validation	Source of algorithm
Exposure	Tolvaptan	[-365, 0]	[1, end of FU]	n/a	ATC	n/a	Intervention		No validation study	n/a
Comparator	Non tolvaptan users	[-365, 0]	[1, end of FU]	n/a	ATC	n/a	Control		No validation study	n/a

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)

7.4.2 Context and rationale for outcome(s) of interest

Clinical trials have suggested that there is an association between tolvaptan treatment and elevations in liver enzyme levels (alanine aminotransferase (ALT) and aspartate aminotransferase (AST)) [2,6]. In these studies, hepatotoxicity is defined as increases in ALT or AST levels to more than 3 times the upper limit of normal (ULN) [2,6]. It is expected that, at least in the unexposed group, liver enzyme levels are not registered regularly. In this emulation, hepatotoxicity will be defined as the registration of ICD-10, Read, or SNOMED codes related to hepatotoxicity or increases in the following liver enzymes if data are available:

- ALT > 3x ULN
- AST > 3x ULN
- Total bilirubin > 2x ULN
- Alkaline phosphatase > 2x ULN + elevated GGT and/or 5-NT

Table 8. Operational Definitions of Outcome

Outcome name	Details	Primary outcome?	Type of outcome	Washout window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source of algorithm
Hepatotoxicity	Composite: ICD-10, Read, or SNOMED, ALS or AST > 3x ULN, total bilirubin or alkaline phosphatase > 2x ULN + elevated GGT and/or 5-NT	yes	Time-to-event	n/a	IP, OP	ICD-10, Read, SNOMED	Any	Intervention and control	n/a	[5,10]

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)
ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal, GGT: gamma-glutamyl transpeptidase, 5-NT: 5'nucleotidase

7.4.3 Context and rationale for follow up

Hepatotoxicity related to tolvaptan may develop after prolonged exposure [6], a multi-year study allows for identification of early-onset and late-onset hepatotoxicity events. Most events of hepatotoxicity are expected to develop within the first 18 months after initiating tolvaptan treatment [10]. Moreover, tolvaptan is a long-term therapy, so patients in clinical practice will remain on the treatment for several years.

Table 9. Operational Definitions of Follow Up

Estimand 1 and 3

Follow up start	Time 0	
Follow up end[1]	**Select all that apply**	**Specify**
Date of outcome	Yes	First occurrence of hepatotoxicity, specified in Table 7
Date of death	Yes	Used to determine date for censoring
End of observation in data	Yes	Last collection date in data source
Day X following index date *(specify day)*	Yes	3 years after time 0
End of study period (specify date)	Yes	31 December 2024
End of exposure *(specify operational details, e.g. stockpiling algorithm, grace period)*	No	Treatment policy ignores treatment discontinuatio
Date of add to/switch from exposure *(specify algorithm)*	Yes	Hypothetical strategy for starting tolvaptan in the control group leads to censoring
Other date *(specify)*	Yes	Receipt of other hepatotoxic medication in either treatment group

[1] Follow up ends at the first occurrence of any of the selected criteria that end follow up.

Estimand 2

Follow up start	Time 0	
Follow up end[1]	**Select all that apply**	**Specify**
Date of outcome	Yes	Specified in Table 7
Date of death	Yes	Used to determine date for censoring
End of observation in data	Yes	Last collection date in data source
Day X following index date *(specify day)*	Yes	3 years after time 0
End of study period (specify date)	Yes	31 December 2024
End of exposure *(specify operational details, e.g. stockpiling algorithm, grace period)*	Yes	Stockpiling algorithm: If refills occur before predicted end of prescription, count overlapping days and add at the end of the treatment episode, if the number of additional days is not more than 90 days. Grace period: Bridge gaps ≤ 30 days between end of current prescription and refill.
Date of add to/switch from exposure *(specify algorithm)*	Yes	To implement the hypothetical strategy for initiating medication with hepatotoxic effects in the intervention group and control group; and in the control group also for initiation of tolvaptan.
Other date *(specify)*	Yes	Receipt of other hepatotoxic medication in either treatment group

[1] Follow up ends at the first occurrence of any of the selected criteria that end follow up.

7.4.4 Context and rationale for covariates (confounding variables and effect modifiers, e.g. risk factors, comorbidities, comedications)

List of key confounders relevant to the emulation study

- Patient characteristics
    - Age
    - Sex
    - Baseline liver function (enzyme levels)
    - Alcohol use disorders
    - Smoking status
    - Comorbidities
        - Any liver disease or liver function disorder
        - Hypertension

Table 10. Operational Definitions of Covariates

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Age at diagnosis	Date of diagnosis, date of birth	Continuous	[0,0]	n/a	n/a	n/a	Intervention and control	n/a	n/a
Sex	Male, female, other	Binary	[0,0]	n/a	n/a	n/a	Intervention and control	n/a	n/a
Baseline liver function	AST, ALT, total bilirubin, alkaline phosphatase, gamma-glutamyl transpeptidase, 5'nucleotidase	Continuous	[-365,0] most recent	OP	n/a	n/a	Intervention and control	n/a	[5,10]
Alcohol use disorders	Most recent known status	Binary	[-365,0] most recent	OP	ICD-10, Read, SNOMED	n/a	Intervention and control	n/a	n/a
Smoking status	Most recent known status	Categorical	[-365,0] most recent	OP	ICD-10, Read, SNOMED	n/a	Intervention and control	n/a	n/a

Characteristic	Details	Type of variable	Assessment window	Care Settings[1]	Code Type[2]	Diagnosis Position[3]	Applied to study populations:	Measurement characteristics/ validation	Source for algorithm
Comorbidity: liver disease	Any liver disease, disorder, or injury	Binary	[start of data, -1]	OP	ICD-10, Read, SNOMED	n/a	Intervention and control	n/a	[5,10]
Comorbidity: hypertension		Binary	[-365, -1]	OP	ICD-10, Read, SNOMED	n/a	Intervention and control	n/a	n/a
Baseline creatinine clearance	Creatinine clearance in ml/min	Continuous	[-365,0] most recent	OP	n/a	n/a	Intervention and control	n/a	[6,7]

[1] IP = inpatient, OP = outpatient, ED = emergency department, OT = other, n/a = not applicable
[2] See appendix for listing of clinical codes for each study parameter
[3] Specify whether a diagnosis code is required to be in the primary position (main reason for encounter)
ALT: alanine aminotransferase, AST: aspartate aminotransferase

7.5 Core Emulation Table - Design Summary

Table 11. Comparison of Target Trial and Proposed Target Trial Emulation Design Elements

Estimand 1 and 3

	Target Trial	Target Trial Emulation	Comment
Inclusion criteria	Be diagnosed with Autosomal Dominant Polycystic Kidney Disease according to Ravine criteria Be at least 18 years old at time of screening Have a total kidney volume of 750 ml or more	Have a recorded confirmed diagnosis of Autosomal Dominant Polycystic Kidney Disease (ICD-10 code Q61.2 or a relevant Read code) Be at least 18 years old at time 0	The target trial used a very specific and relatively healthy population, we are interested in the entire population that may receive a tolvaptan prescription as this reflects the real situation better, hence we have relaxed the inclusion criteria on kidney function and volume. In England, Wales, and Scotland, the Kidney Association Guidance Commentary on

	Have a creatinine clearance of 60 ml/min or more		tolvaptan use mentions the possibility for tolvaptan use at a creatinine clearance of 30 ml/min or more. Broadening this criterium to 30 ml/min instead of 60 ml/min better reflects the reality of the patient population and clinical usage. Measuring kidney volume is not part of the prescribing safety measures when starting tolvaptan, using kidney volume as an inclusion criterion therefore creates a subgroup that does not reflect practice accurately.
Exclusion criteria	AST or ALT levels > 1.5x ULN at screening Total bilirubin or alkaline phosphatase > ULN Use or have used tolvaptan in the past year Use of medication known for potentially causing hepatotoxic effects Use of diuretics	AST or ALT levels > 1.5x ULN at screening Total bilirubin, alkaline phosphatase, gamma-glutamyl transpeptidase, 5'nucleotidase > ULN Or current hepatotoxicity event identified through relevant ICD-10, Read or SNOMED codes. Use or have used tolvaptan in the past year Use of medication known for potentially causing hepatotoxic effects Use of diuretics	All exclusion criteria from the hypothetical target trial are used in the emulation. These exclusion criteria will be identified through recorded values of liver enzymes, or the registration of ICD-10, Read, or SNOMED codes related to hepatotoxicity during the screening period.
Setting	Multicenter, parallel-group, randomized controlled, open-label trial.	Routine care data from CPRD database capturing prescriptions, outcomes, and covariates	Real-world data captures care as delivered and reflects the setting where patients are most likely to be recruited from. In combination with the relaxed in- and exclusion criteria, the study outcomes relate to the entire population of tolvaptan users for ADPKD, thus supporting broad policy changes if applicable.

			Even if tolvaptan was initially prescribed in hospital/by specialist, the general practitioner (GP) can renew the prescription in primary care and that's where we will see the record. The deviation from the target trial is that we would miss prescriptions that were never transferred to the GP, or if they are initiated by the specialist we would mismeasure the initiation date.
Time *(when follow up begins and ends):*	Begins at randomization, ends at: outcome, tolvaptan initiation in control group, initiation of medication with hepatotoxic effects, all-cause mortality, loss to follow-up, study withdrawal or 36 months.	Begins at date of registration of tolvaptan initiation (index date) for intervention group and at time-matched physician visit (without tolvaptan initiation, duration since ADPKD diagnosis must be within the same quintile) for control group. Ends at outcome, tolvaptan initiation in control group, initiation of medication with hepatotoxic effects, all-cause mortality, loss to follow-up, end of data availability or 36 months.	Challenges to identifying intercurrent events include incomplete capturing of prescribing in the database, delayed or missing updates, or incomplete linkage between databases. Additionally, a prescribing database does not provide information on dispensing or adherence, thus misclassification may occur.
Study treatment conditions	Intervention group: starting with daily morning and afternoon doses of 45 mg and 15 mg, respectively, with weekly increases to 60 mg and 30 mg and then to 90 mg and 30 mg, up to highest patient-reported tolerability Control group: unexposed	Intervention group: initiation of any dose of tolvaptan Control group: any included participant with a time-matched visit to a physician without initiating tolvaptan treatment, matched on time since ADPKD diagnosis (same quintile) at the date of tolvaptan initiation.	Reflects new user design, dose or duration flexibility mirrors routine care.
Outcome (including operational definition)	Time to first occurrence of one of the following:	Time to first recording of registered confirmed diagnosis of liver injury following	Definition of endpoint differs between TT and TTE.

	Liver enzyme levels: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN	a relevant ICD-10, SNOMED or t Read code Liver enzyme levels: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN + gamma-glutamyl transpeptidase and/or 5'nucleotidase elevation	Measurement of liver enzymes is mandatory before tolvaptan initiation and will be available for patients in the tolvaptan group monthly for the first 18 months and three-monthly after that. In the control group, it is expected that these data are often not present. Therefore, in the TT, hepatotoxicity is based on lab values of liver enzymes, but for the TTE, ICD-10, SNOMED, or Read codes regarding liver injuries will be used primarily.
Method of Assignment to Trial Intervention	1:1 randomisation after screening	Randomisation cannot be directly emulated.	At the analysis stage, matching will be performed. Matching will comprise two stages, first a time match based on time since ADPKD diagnosis (in quintiles) combined with a calendar match of tolvaptan prescription (nearest match within 12 months). Second, a propensity score matching. The time matching will be at a 5 : 1 ratio for control : intervention. Propensity score matching will be used to adjust for baseline covariates at a 1 : 1 ratio for control : intervention. Individuals can only be eligible as matched control once and can only be assigned to either the control group or the intervention group once (at time 0). To adjust for baseline covariates and balance confounders, emulates randomization. Prevalent new user design prevents immortal

			time bias by time-controlled matching of intervention and control groups. Time matching is therefore done in quintiles to allow for a sufficiently large pool of participants for the second matching step, this may increase the precision of the estimated treatment effect.
Handling of Intercurrent Events	Treatment discontinuation: treatment policy Tolvaptan initiation in control group: hypothetical Same for both treatment strategies: Use of any medication with known hepatoxicity effects: hypothetical All-cause mortality: while alive	Same strategies implemented based on prescribing data, mortality data and using administrative censoring (or lack of for these intercurrent events). Operational definitions: Treatment discontinuation is identified using prescription refill data, where a gap of more than 90 days between refills is considered a discontinuation. Treatment initiation of tolvaptan or medication with hepatotoxic effects is identified using prescription refill data. All-cause death is determined using cause-of-death data.	Medication usage is estimated using prescribing records which may not accurately reflect usage. Treatment discontinuation in particular is prone to timing errors.
Loss to follow up	A participant will be classified as lost to follow-up when after multiple attempts contact with the patient remains impossible and health status remains unknown.	Date patient left the practice.	Loss to follow-up will be defined using real-world proxies, recognizing that in some cases patients may appear to remain under follow-up despite having effectively left (e.g., if they do not formally de-register from their GP). This risk is expected to be low, where unique patient

			identifiers ensure automatic de-registration upon re-registration at a new practice.

ADPKD: Autosomal Dominant Polycystic Kidney Disease, ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal

Estimand 2

	Target Trial	Target Trial Emulation	Comment
Inclusion criteria	Be diagnosed with Autosomal Dominant Polycystic Kidney Disease according to Ravine criteria Be at least 18 years old at time of screening Have a total kidney volume of 750 ml or more Have a creatinine clearance of 60 ml/min or more	Have a recorded confirmed diagnosis of Autosomal Dominant Polycystic Kidney Disease (ICD-10 code Q61.2 or a relevant Read code) Be at least 18 years old at time 0	The target trial used a very specific and relatively healthy population, we are interested in the entire population that may receive a tolvaptan prescription as this reflects the real situation better, hence we have loosened the inclusion criteria on kidney function and volume. In England, Wales, and Scotland, the Kidney Association Guidance Commentary on tolvaptan use mentions the possibility for tolvaptan use at a creatinine clearance of 30 ml/min or more. Broadening this criterium to 30 ml/min instead of 60 ml/min could increase the total number of participants in the TTE and better reflect the reality of the patient population and clinical usage. Measuring kidney volume is not part of the prescribing safety measures when starting tolvaptan, using kidney volume as an inclusion criterion therefore creates a subgroup that does not reflect practice accurately.

Exclusion criteria	AST or ALT levels > 1.5x ULN at screening Total bilirubin or alkaline phosphatase > ULN Use or have used tolvaptan in the past year Use of medication known for potentially causing hepatotoxic effects Use of diuretics	AST or ALT levels > 1.5x ULN at screening Total bilirubin or alkaline phosphatase > ULN Or current hepatotoxicity event identified through relevant ICD-10, Read or SNOMED codes. Use or have used tolvaptan in the past year Use of medication known for potentially causing hepatotoxic effects Use of diuretics	All exclusion criteria from the hypothetical target trial are used in the emulation. These exclusion criteria will be identified through recorded values of liver enzymes, or the registration of ICD-10, Read, or SNOMED codes related to hepatotoxicity during the screening period.
Setting	Multicenter, parallel-group, randomized controlled, open-label trial.	Routine care data from CPRD database capturing prescriptions, outcomes, and covariates.	Real-world data captures care as delivered and reflects the setting where patients are most likely to be recruited from. In combination with the loosened in- and exclusion criteria, the study outcomes relate to the entire population of tolvaptan users for ADPKD, thus supporting broad policy changes if applicable.
Time *(when follow up begins and ends):*	Begins at randomization, ends at: outcome, tolvaptan initiation in control group, initiation of medication with hepatotoxic effects, all-cause mortality, loss to follow-up, study withdrawal or 36 months.	Begins at date of registration of tolvaptan initiation (index date) for intervention group and at time-matched physician visit (without tolvaptan initiation, duration since ADPKD diagnosis must be within the same quintile) for control group, ends at outcome, treatment discontinuation, tolvaptan initiation in control group, initiation of medication with known hepatotoxic effects, all-cause mortality, censoring, or 36 months.	Aligns start of follow-up with treatment initiation in intervention group to mimic randomization, Prevalent new-user design avoids immortal time bias. Challenges to identifying intercurrent events include incomplete capturing of prescribing in the database, delayed or missing updates, or incomplete linkage between databases. Additionally, a prescribing database does not

			provide information on dispensing or adherence, thus misclassification may occur.
Study treatment conditions	Intervention group: starting with daily morning and afternoon doses of 45 mg and 15 mg, respectively, with weekly increases to 60 mg and 30 mg and then to 90 mg and 30 mg, up to highest patient-reported tolerability Control group: unexposed	Intervention group: initiation of any dose of tolvaptan. Control group: any included participant with a time-matched visit to a physician without initiating tolvaptan treatment, matched on time since ADPKD diagnosis (same quintile) at the date of tolvaptan initiation.	Reflects new user design, dose or duration flexibility mirrors routine care.
Outcome (including operational definition)	Time to first occurrence of one of the following: Liver enzyme levels: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN	Time to first recording of registered confirmed diagnosis of liver injury following a relevant ICD-10, SNOMED, or Read code. Liver enzyme levels: • ALT > 3x ULN • AST > 3x ULN • Total bilirubin > 2x ULN • Alkaline phosphatase > 2x ULN	Definition of endpoint differs between TT and TTE. Measurement of liver enzymes is mandatory before tolvaptan initiation and will be available for patients in the tolvaptan group monthly for the first 18 months and three-monthly after that. In the control group, it is expected that these data are often not present. Therefore, in the TT, hepatotoxicity is based on lab values of liver enzymes, but for the TTE, ICD-10, SNOMED, or Read codes regarding liver injuries will be used primarily.
Method of Assignment to Trial Intervention	1:1 randomisation after screening	Randomisation cannot be directly emulated.	At the analysis stage, matching will be performed. Matching will comprise two stages, first a time match based on time since ADPKD diagnosis (in quintiles) combined with a calendar match of tolvaptan prescription (nearest match within 12 months). Second, a

			propensity score matching. The time matching will be at a 5 : 1 ratio for control : intervention. Propensity score matching will be used to adjust for baseline covariates at a 1 : 1 ratio for control : intervention. Individuals can only be eligible as matched control once and can only be assigned to either the control group or the intervention group once (at time 0). To adjust for baseline covariates and balance confounders, emulates randomization. Prevalent new user design prevents immortal time bias by time-controlled matching of intervention and control groups. Time matching is therefore done in quintiles to allow for a sufficiently large pool of participants for the second matching step this may increase the precision of the estimated treatment effect.
Handling of Intercurrent Events	Treatment discontinuation: while on treatment Tolvaptan initiation in control group: while on treatment Same for both treatment strategies: Use of any medication with known hepatoxicity effects: hypothetical	Same strategies implemented based on prescribing data, mortality data and using administrative censoring (or lack of for these intercurrent events) Operational definitions: Treatment discontinuation is identified using prescription refill data, where a gap of more than 90 days between refills is considered a discontinuation.	Medication usage is estimated using prescribing records which may not accurately reflect usage. Treatment discontinuation in particular is prone to timing errors.

	All-cause mortality: while alive	Treatment initiation of tolvaptan or medication with hepatotoxic effects is identified using prescription refill data. All-cause death is determined using cause-of-death data.	
Loss to follow up	A participant will be classified as lost to follow-up when after multiple attempts contact with the patient remains impossible and health status remains unknown.	Date patient left the practice)	Loss to follow-up will be defined using real-world proxies, recognizing that in some cases patients may appear to remain under follow-up despite having effectively left (e.g., if they do not formally de-register from their GP). This risk is expected to be low, where unique patient identifiers ensure automatic de-registration upon re-registration at a new practice

ADPKD: Autosomal Dominant Polycystic Kidney Disease, ALT: alanine aminotransferase, AST: aspartate aminotransferase, ULN: Upper Limit of Normal

7.6 Data analysis

7.6.1 Analysis Plan

Overview

The analyses are conducted within a target trial emulation framework to estimate the effect of tolvaptan exposure compared with non exposure on the risk of hepatotoxicity.

For **Estimand 1**, the main estimand supporting decision making, the primary causal effect summary measure is the hazard ratio for time to first hepatotoxicity event, estimated using a stratified Cox proportional hazards model in the propensity score-matched sample.

Sensitivity analyses will assess robustness of the primary findings to key assumptions, including inverse probability of censoring weighting (IPCW) (assuming censoring is independent of the outcome, with all common causes of both the outcome and censoring being accounted for conditional on the treatment, time-varying covariates and survival up to the time of censoring as well as indirectly baseline covariates used to estimate the treatment weights), best/worst case scenario analysis, and outcome misclassification (details in Section 7.6.5).

Two supplemental estimands are also defined: **Estimand 2**, applying a while-on-treatment strategy for intercurrent events, and **Estimand 3**, estimating treatment effects using restricted mean survival time (RMST) derived from a Propensity score matched Weibull accelerated failure time (AFT) model. In addition, supplemental analyses (e.g., crude and PS-matching adjusted Kaplan–Meier curves, crude Cox models, event counts and incidence rates, propensity score and weight distributions, covariate balance before and after weighting, censoring and intercurrent event patterns, proportional hazards diagnostics, positivity checks, and multiple-imputation diagnostics) will be conducted to support interpretation of the main analysis.

For all estimands, conditional propensity scores will be calculated, and patients will be matched accordingly. For estimand 1 and 2, the Cox model will be used to calculate the HR of hepatotoxicity of tolvaptan users vs. nonusers. For estimand 3, an accelerated failure time model with Weibull distribution will be used to calculate the mean (CI 95%) hepatotoxicity free time in tolvaptan users vs. nonusers and the difference in restricted mean survival time.

7.6.2 Primary Estimand (1) Analysis

i. Objective

This study will investigate the extent to which Tolvaptan exposure increases the risk of hepatotoxicity relative to non-exposure

ii. Exposure contrast

Tolvaptan users vs. nonusers

iii. Outcome

Time to first occurrence of hepatotoxicity

iv. Analytic software

R

v. Handling of intercurrent events

- Treatment discontinuation: treatment policy – data after IE is of interest for analysis
- Tolvaptan initiation in control group: hypothetical – data after IE is not of interest for analysis
- Use of any medication with known hepatoxicity effects: hypothetical – data after IE is not of interest for analysis

- All-cause mortality: while alive – data after IE is not of interest for analysis

vi. Outcome Modelling

Outcome model: Cox proportional hazards

Follow-up time

- start of follow-up: date of tolvaptan treatment initiation or time-matching physician visit
- endpoint: time from treatment to first occurrence of hepatotoxicity
- censoring:
    - Non-administrative: censoring due to loss of follow-up, initiation of tolvaptan in control group, initiation of hepatotoxic medication in either treatment group
    - Administrative: censoring at end of study follow-up in absence of hepatotoxicity (maximum 3 years), censoring at death, database end
- Model covariate: treatment group (tolvaptan vs. unexposed)

Assumptions of Cox Model

- Proportional Hazards:
    - The effect of treatment is assumed to be constant over time.

- Non-informative Censoring:
    - Censoring is assumed to be independent of the outcome, conditional on the treatment, survival up to the time of censoring and indirectly baseline covariates used to estimate the treatment weights

Diagnostics for Cox Model

- Proportional hazards assessed using log(-log) survival plots or Schoenfeld residuals.

The assumption of non-informative censoring cannot be verified with observed data; it will be addressed through sensitivity analyses.

vii. Confounding Adjustment

We will use logistic regression to estimate the propensity score of receiving tolvaptan versus being unexposed. The model will include baseline covariates selected a priori based on clinical relevance and prior evidence as potential confounders—specifically, variables considered to be causes of

the outcome, potentially associated with treatment but not affected by treatment initiation. The linear predictor of the model will include the following covariates: age, sex, smoking status, alcohol disorders, baseline liver function, history of hypertension, history of liver disease, baseline creatinine clearance. We will perform nearest-neighbour propensity score matching, using a calliper of 0.2.

Assumptions Underlying propensity score matching;

- No unmeasured confounding (all relevant baseline confounders are included in the propensity score model).
- Positivity (each individual has a non-zero probability of receiving either treatment, given their covariates).
- Correct model specification (the propensity score model is correctly specified [functional form, covariate inclusion]).
- Consistency (each individual's potential outcome under the observed treatment equals their actual outcome).

Diagnostics for propensity score matching:

- Covariate balance: Check that baseline characteristics are balanced across treatment groups after matching
- Evaluate standardized mean differences (SMDs): SMDs < 0.1 will be considered acceptable.
- Positivity check: Ensure adequate overlap in propensity score distributions between tolvaptan users and potential unexposed comparators

viii. Missing Data Handling

Censoring-at-random assumption implicitly made in the Cox regression model. Censored participants contribute partial information, i.e. time at risk up to the time of censoring.

Missing Exposure Data

We assume that missing refill or prescription records for tolvaptan reflect true treatment discontinuation after 30 days, and not incomplete data capture or prescriptions issued outside the database

Missing Outcome Data

The Cox proportional hazards model implicitly assumes non-informative censoring, meaning that censored participants contribute time at risk up to the time of censoring and their censoring is unrelated to the outcome, **conditional** on model covariates and survival up to the time of censoring (i.e., outcome data is missing at random under these assumptions).
Additionally, it is assumed that no outcome event observed before censoring (i.e., during observed follow-up) implies no outcome event happened.

Missing Covariate Data

The absence of a disease diagnosis code will be interpreted as the absence of the corresponding condition. Missing values for lifestyle factors (e.g., smoking status) and lab variables (e.g, AST/ALT values) will be addressed using multiple imputation with chained equations (MICE) under the Missing at Random (MAR) assumption.

Assessment of Missingness

Before performing imputation, we will examine the extent and patterns of missingness to evaluate whether imputation is appropriate. Specifically, we will:

- Quantify the percentage of missing data for each covariate.
- Compare the proportion of missing values across treatment groups to assess differential missingness.
- If a covariate has more than 40% missing data, we will consider alternative approaches (e.g., exclusion of the variable, sensitivity analyses) and justify the decision. Thresholds of 40% have been cited because effect estimates begin to be less reliable under imputation as the level of missingness increases beyond this threshold.[12]

Imputation Model

The MICE procedure will include all covariates used in the outcome and treatment models. The treatment and outcomes of interest will also be included.

Key covariates included in the imputation model will be:

- Demographics (age, sex)
- Clinical history and comorbidities (e.g., liver diseases or function disorders)
- Laboratory values and vital signs (e.g., creatinine clearance, blood pressure)
- Lifestyle factors (e.g., smoking, alcohol use)

Full Conditional Distributions

MICE will use variable-specific conditional models:

- Logistic regression for binary variables (e.g., smoking yes/no).
- Multinomial logistic regression for categorical variables with >2 categories.
- Predictive mean matching for continuous variables

Number of Imputations and Diagnostics

We will generate at least 10 imputed datasets (to ensure stable estimates given the level of missingness) and pool results across imputations using Rubin's rules. Diagnostics will include:

- Checking whether imputed values are plausible and consistent with observed distributions.
- Evaluating convergence of the chained equations.
- Assessing stability and consistency of results across imputed datasets.

Effect Estimation Under Multiple Imputation

The imputation model will be applied prior to effect estimation and outcome models will then be fitted in each imputed dataset, and treatment effect estimates (e.g., hazard ratios) will be pooled across datasets using Rubin's rules.

ix. Subgroup Analyses

Not applicable

7.6.3 Supplemental Estimand (2) Analysis

i. Objective

This study will investigate the extent to which Tolvaptan exposure increases the risk of hepatotoxicity relative to non-exposure

ii. Exposure contrast

Tolvaptan users vs. nonusers

iii. Outcome

Time to first occurrence of hepatotoxicity

iv. Analytic software

R

v. Handling of intercurrent events

- Treatment discontinuation: while on treatment – data after IE is not of interest for analysis
- Tolvaptan initiation in control group: while on treatment – data after IE is not of interest for analysis
- Use of any medication with known hepatoxicity effects: hypothetical – data after IE is not of interest for analysis
- All-cause mortality: while alive – data after IE is not of interest for analysis

vi. Outcome Modelling

Outcome model: Cox proportional hazards
Follow-up time

- start of follow-up: date of tolvaptan treatment initiation or time-matching physician visit
- endpoint: time from treatment to first occurrence of hepatotoxicity
- censoring:
    - Non-administrative: censoring due to loss to follow-up, initiation of hepatotoxic medication in either treatment group
    - Administrative: censoring at end of study follow-up in absence of hepatotoxicity (maximum 3 years), initiation of tolvaptan in control group, treatment discontinuation, censoring at death, database end
- Main covariate: treatment group (tolvaptan vs. unexposed)

Assumptions of Cox Model

- Proportional Hazards:
    - The effect of treatment is assumed to be constant over time.

- Non-informative Censoring:
    - Censoring is assumed to be independent of the outcome, conditional on the treatment, survival up to the time of censoring and indirectly baseline covariates used to estimate the treatment weights

Diagnostics for Cox Model

- Proportional hazards assessed using log(-log) survival plots or Schoenfeld residuals.

The assumption of non-informative censoring cannot be verified with observed data; it will be addressed through sensitivity analyses.
We will also conduct descriptive analyses to characterize censoring patterns overall and across treatment groups. This will include median time to censoring overall and according to the reason for censoring. This will be estimated separately for the overall study population and by treatment arm (tolvaptan vs. unexposed).

vii. Confounding Adjustment

We will use logistic regression to estimate the propensity score of receiving tolvaptan versus being unexposed. The model will include baseline covariates selected a priori based on clinical relevance and prior evidence as potential confounders—specifically, variables considered to be associated with the outcome, plausibly associated with treatment but not affected by treatment initiation. The linear predictor of the model will include the following covariates: age, sex, smoking status, alcohol disorders, baseline liver function, history of hypertension, history of liver disease, baseline creatinine clearance. We will perform Logistic regression to estimate propensity score using the below listed covariates followed by nearest-neighbour propensity score matching, using a calliper of 0.2.

Assumptions Underlying propensity score matching;

No unmeasured confounding (all relevant baseline confounders are included in the propensity score model).
Positivity (each individual has a non-zero probability of receiving either treatment, given their covariates).
Correct model specification (the propensity score model is correctly specified [functional form, covariate inclusion]).
Consistency (each individual's potential outcome under the observed treatment equals their actual outcome).

Diagnostics for propensity score matching:
Covariate balance: Check that baseline characteristics are balanced across treatment groups after matching
Evaluate standardized mean differences (SMDs): SMDs < 0.1 will be considered acceptable.
Positivity check: Ensure adequate overlap in propensity score distributions between tolvaptan users and potential unexposed comparators

viii. Missing Data Handling

Censoring-at-random assumption, censored participants contribute partial information.

Missing Exposure Data
We assume that missing refill or prescription records for tolvaptan reflect true treatment discontinuation after 30 days, and not incomplete data capture or prescriptions issued outside the database

Missing Outcome Data
The Cox proportional hazards model implicitly assumes non-informative censoring, meaning that censored participants contribute time at risk up to the time of censoring and their censoring is unrelated to the outcome, **conditional** on model covariates and survival up to the time of censoring (i.e., outcome data is missing at random under these assumptions)..
Additionally, it is assumed that no outcome event observed before censoring (i.e., during observed follow-up) implies no outcome event happened.

Missing Covariate Data
The absence of a disease diagnosis code will be interpreted as the absence of the corresponding condition. Missing values for lifestyle factors (e.g., smoking status) and lab variables (e.g, AST/ALT values) will be addressed using multiple imputation with chained equations (MICE) under the Missing at Random (MAR) assumption.

Assessment of Missingness
Before performing imputation, we will examine the extent and patterns of missingness to evaluate whether imputation is appropriate. Specifically, we will:

- Quantify the percentage of missing data for each covariate.
- Compare the proportion of missing values across treatment groups to assess differential missingness.

- If a covariate has more than 40% missing data, we will consider alternative approaches (e.g., exclusion of the variable, sensitivity analyses) and justify the decision. Thresholds of 40% have been cited because effect estimates begin to be less reliable under imputation as the level of missingness increases beyond this threshold.[12]

Imputation Model

The MICE procedure will include all covariates used in the outcome and treatment models. The treatment and outcomes of interest will also be included.

Key covariates included in the imputation model will be:

- Demographics (age, sex)
- Clinical history and comorbidities (e.g., liver diseases or function disorders)
- Laboratory values and vital signs (e.g., creatinine clearance, blood pressure)
- Lifestyle factors (e.g., smoking, alcohol use)

Full Conditional Distributions

MICE will use variable-specific conditional models:

- Logistic regression for binary variables (e.g., smoking yes/no).
- Multinomial logistic regression for categorical variables with >2 categories.
- Predictive mean matching for continuous variables

Number of Imputations and Diagnostics

We will generate at least 10 imputed datasets (to ensure stable estimates given the level of missingness) and pool results across imputations using Rubin's rules. Diagnostics will include:

- Checking whether imputed values are plausible and consistent with observed distributions.
- Evaluating convergence of the chained equations.
- Assessing stability and consistency of results across imputed datasets.

Effect Estimation Under Multiple Imputation

The imputation model will be applied prior to effect estimation and outcome models will then be fitted in each imputed dataset, and treatment effect estimates (e.g., hazard ratios) will be pooled across datasets using Rubin's rules.

ix. Subgroup Analyses

Not applicable

7.6.4 Supplemental Estimand (3) Analysis

i. Objective

This study will investigate the extent to which Tolvaptan exposure increases the risk of hepatotoxicity relative to non-exposure

ii. Exposure contrast

Tolvaptan users vs. nonusers

iii. Outcome

Time to first occurrence of hepatotoxicity

iv. Analytic software

R

v. Handling of intercurrent events

- Treatment discontinuation: treatment policy – data after IE is of interest for analysis
- Tolvaptan initiation in control group: hypothetical – data after IE is not of interest for analysis
- Use of any medication with known hepatoxicity effects: hypothetical – data after IE is not of interest for analysis
- All-cause mortality: while alive – data after IE is not of interest for analysis

vi. Outcome Modelling

Weibull Accelerated Failure Time (AFT) model, followed by estimation of Restricted Mean Survival Time (RMST) at 3 years.

Start of follow-up: Date of tolvaptan initiation.

Endpoint: Time from tolvaptan initiation to first occurrence of hepatotoxicity

Censoring:

- Non-administrative: censoring due to loss of follow-up, initiation of tolvaptan in control group, initiation of hepatotoxic medication in either treatment group

- Administrative: censoring at end of study follow-up in absence of hepatotoxicity (maximum 3 years), censoring at death, database end

Model Assumptions:

- Survival times follow a Weibull distribution
- Censoring is assumed to be independent of the outcome, conditional on model covariates and not experiencing hepatotoxicity up to the time of censoring.
- Log-linear relationship between covariates and log survival time

Diagnostics:

- Log(-log(S(t))) vs log(t) should be linearQ-Q plot of residuals

We will also conduct descriptive analyses to characterize censoring patterns overall and across treatment groups. This will include median time to censoring overall and according to the reason for censoring. This will be estimated separately for the overall study population and by treatment arm (tolvaptan vs. unexposed).

The assumption of non-informative censoring cannot be directly verified with observed data; it will be addressed through sensitivity analyses.

To estimate the RMST at 3 years from the Weibull AFT model, we first use the model to obtain the predicted survival curve for each treatment group. The RMST is then calculated as the average survival time up to a fixed time point, which corresponds to the area under the survival curve between time zero and the chosen time horizon (3 years).

- Fit the Weibull AFT model, which gives the shape and scale of the survival curve for each group.
- From this model, generate the predicted survival probability at each time.
- Integrate (i.e., add up) the survival probabilities from time 0 to 3 years. The result is the expected survival time lived within those windows.
- Compare the RMST values between treatment groups to obtain the difference in average survival time over 3 years.

To account for uncertainty from both missing data and sampling variability, we will combine multiple imputation with nonparametric bootstrapping. We will use the Boot MI approach proposed by Schomaker & Heumann [12] in which bootstrap resampling is performed first, followed by multiple imputation within each bootstrap sample.

Specifically, we will generate B = 1,000 bootstrap samples of the original dataset (with missing data) and perform M = 10 imputations within each bootstrap sample using the prespecified imputation model. Within each bootstrap sample we will combine the M estimates using Rubin's rules to produce one bootstrap estimate. The empirical distribution of the resulting B bootstrap estimates will be used to construct percentile-based 95% confidence intervals. To obtain final point estimates, we will pool results across M = 40 imputations in the original dataset.

vii. Confounding Adjustment

We will use logistic regression to estimate the propensity score of receiving tolvaptan versus being unexposed. The model will include baseline covariates selected a priori based on clinical relevance and prior evidence as potential confounders—specifically, variables considered to be associated with the outcome, plausibly associated with treatment but not affected by treatment initiation. The linear predictor of the model will include the following covariates: age, sex, smoking status, alcohol disorders, baseline liver function, history of hypertension, history of liver disease, baseline creatinine clearance. We will perform nearest-neighbour propensity score matching, using a calliper of 0.2.

Assumptions Underlying propensity score matching;

- No unmeasured confounding (all relevant baseline confounders are included in the propensity score model).
- Positivity (each individual has a non-zero probability of receiving either treatment, given their covariates).
- Correct model specification (the propensity score model is correctly specified [functional form, covariate inclusion]).
- Consistency (each individual's potential outcome under the observed treatment equals their actual outcome).

Diagnostics for propensity score matching:

- Covariate balance: Check that baseline characteristics are balanced across treatment groups after matching
- Evaluate standardized mean differences (SMDs): SMDs < 0.1 will be considered acceptable.

Positivity check: Ensure adequate overlap in propensity score distributions between tolvaptan users and potential unexposed comparators

viii. Missing Data Handling

Censoring-at-random assumption, censored participants contribute partial information.

Missing Exposure Data

We assume that missing refill or prescription records for tolvaptan reflect true treatment discontinuation after 30 days, and not incomplete data capture or prescriptions issued outside the database

Missing Outcome Data

The Cox proportional hazards model implicitly assumes non-informative censoring, meaning that censored participants contribute time at risk up to the time of censoring and their censoring is unrelated to the outcome, **conditional** on model covariates and survival up to the time of censoring (i.e., outcome data is missing at random under these assumptions).

Additionally, it is assumed that no outcome event observed before censoring (i.e., during observed follow-up) implies no outcome event happened.

Missing Covariate Data

The absence of a disease diagnosis code will be interpreted as the absence of the corresponding condition. Missing values for lifestyle factors (e.g., smoking status) and lab variables (e.g, AST/ALT values) will be addressed using multiple imputation with chained equations (MICE) under the Missing at Random (MAR) assumption.

Assessment of Missingness

Before performing imputation, we will examine the extent and patterns of missingness to evaluate whether imputation is appropriate. Specifically, we will:

- Quantify the percentage of missing data for each covariate.
- Compare the proportion of missing values across treatment groups to assess differential missingness.
- If a covariate has more than 40% missing data, we will consider alternative approaches (e.g., exclusion of the variable, sensitivity analyses) and justify the decision. Thresholds of 40% have been cited because effect estimates begin to be less reliable under imputation as the level of missingness increases beyond this threshold.[13]

Imputation Model

The MICE procedure will include all covariates used in the outcome and treatment models. The treatment and outcomes of interest will also be included.

Key covariates included in the imputation model will be:

- Demographics (age, sex)
- Clinical history and comorbidities (e.g., liver diseases or function disorders)
- Laboratory values and vital signs (e.g., creatinine clearance, blood pressure)
- Lifestyle factors (e.g., smoking, alcohol use)

Full Conditional Distributions

MICE will use variable-specific conditional models:

- Logistic regression for binary variables (e.g., smoking yes/no).
- Multinomial logistic regression for categorical variables with >2 categories.
- Predictive mean matching for continuous variables

Number of Imputations and Diagnostics

We will generate at least 10 imputed datasets (to ensure stable estimates given the level of missingness) and pool results across imputations using Rubin's rules. Diagnostics will include:

- Checking whether imputed values are plausible and consistent with observed distributions.
- Evaluating convergence of the chained equations.
- Assessing stability and consistency of results across imputed datasets.

Effect Estimation Under Multiple Imputation
The imputation model will be applied prior to effect estimation and outcome models will then be fitted in each imputed dataset, and treatment effect estimates (e.g., hazard ratios) will be pooled across datasets using Rubin's rules.

ix. Subgroup Analyses

Not applicable

7.6.5. Sensitivity Analyses

Sensitivity analyses – rationale, strengths and limitations

Due to the challenges in measuring hepatotoxicity, outcome misclassification may occur. This could be differential with respect to the exposure, i.e., the likelihood or pattern of misclassification may depend on whether patients used tolvaptan or not. This is especially likely as liver enzyme levels will be more regularly recorded in tolvaptan users than unexposed patients. Sensitivity analyses will be conducted for the primary estimand.

Table 12. Sensitivity analyses - Outcome Misclassification

Analysis Method	• To evaluate the potential impact of exposure initiation misclassification at index date, we will conduct a probabilistic bias analysis at the summary estimate level. This analysis will be performed after pooling hazard ratios across imputed datasets using Rubin's rules so that uncertainty due to missing data and sampling variability is incorporated into the input effect estimate. The pooled hazard ratio and its variance will then serve as the starting point for the bias analysis. • Plausible distributions for the sensitivity and specificity of prescription-based exposure classification will be specified based on the known limitations of prescription records as proxies for true medication use (e.g., prescriptions may not be filled or taken). This sensitivity analysis addresses misclassification of initial exposure assignment only (i.e., at treatment initiation), not exposure discontinuation or adherence during follow-up. • Within each Monte Carlo iteration, a hazard ratio will be sampled from a probability distribution informed by the pooled hazard ratio and its variance from the main analysis. This step propagates uncertainty due to sampling variability and multiple imputation into the bias analysis. • In each iteration, a new pair of sensitivity and specificity values will be drawn for each treatment group from the prespecified distributions. These values will be used to probabilistically adjust the pooled hazard ratio for exposure misclassification using established bias-adjustment methods for dichotomous exposure classification. Repeating this process across many iterations (e.g., 10,000) will generate a distribution of bias-adjusted hazard ratios that incorporates both (1) the uncertainty from the primary analysis (due to sampling variation and multiple imputation) and (2) the uncertainty associated with assumptions about misclassification parameters.

	• Final uncertainty will be represented by the percentile-based 95% confidence interval from the distribution of simulated bias-adjusted hazard ratios, thereby reflecting combined uncertainty from missing data, sampling variability, and exposure misclassification.
Assumptions	• Misclassification is differential • Sensitivity and specificity of outcome classification are known or reasonably estimated from external data or expert opinion. • Misclassification affects only the outcome, not exposure or covariates. • The misclassification process can be simulated accurately. • Sensitivity and specificity are considered constant over the follow-up period, i.e., the probability of correctly classifying an event does not depend on time

What is being varied? How?	• The assumption that the outcome is measured without error is relaxed.
	• Varying values of sensitivity and specificity are drawn from defined distributions in each simulation iteration.
	• For differential misclassification, separate Sensitivity/Specificity values are specified for individuals with and without exposure:
	  ○ Exposed: sensitivity is high due to regular monitoring: 0.90 - 0.95, specificity: 0.90 - 0.99
	  ○ Unexposed: sensitivity is low due to irregular monitoring: 0.55 - 0.65, specificity: 0.90 - 0.99

Why? (What do you expect to learn?)	• To assess the robustness of the estimated treatment effect to plausible levels of outcome misclassification. • To determine whether conclusions change under realistic measurement error for outcome.
Strengths Compared to Primary Analysis	• Explicitly accounts for uncertainty in outcome measurement. • Provides a distribution of adjusted estimates rather than a single corrected value. • Can reflect differential or non-differential misclassification. • Enhances transparency around the impact of measurement error.
Limitations Compared to Primary Analysis	• Requires external data or expert assumptions to specify sensitivity and specificity. • Results are only as reliable as the plausibility of input parameters. • May be computationally intensive. • Does not account for other sources of bias (e.g., unmeasured confounding, exposure misclassification) unless jointly modeled.

Table 13. Sensitivity analyses - Best/worst case scenario

<table>
<tr><td>Analysis Methods</td><td>

Non-informative censoring. Best/worst case scenario under the censoring not at random assumption. The goal is to assess the impact on the estimated treatment effect of selected censoring non-at-random assumptions. The assumptions chosen represent the 4 extremes of a tipping point sensitivity analysis.
For non-administrative censored individuals, repeat the analysis under four scenarios, assuming a) all censored individuals had the outcome of interest at the censoring date and b) none of the censored individuals had the outcome of interest by the end of the study, respectively for each treatment group. This equates to the following scenarios

<table>
<tr><td>Scenario</td><td>Exposed Group (tolvaptan)</td><td>Unexposed Group (unexposed)</td><td>Interpretation</td></tr>
<tr><td>1</td><td>Best case (lowest event rate)</td><td>Worst case (highest event rate)</td><td>Maximally favors tolvaptan</td></tr>
<tr><td>2</td><td>Worst case (highest event rate)</td><td>Best case (lowest event rate)</td><td>Maximally favors unexposed</td></tr>
<tr><td>3</td><td>Best case</td><td>Best case</td><td>Optimistic for both groups</td></tr>
<tr><td>4</td><td>Worst case</td><td>Worst case</td><td>Pessimistic for both groups</td></tr>
</table>
</td></tr>
<tr><td>Assumptions</td><td>

Censoring not at random assumptions whereby a extreme increase/decrease in the hazard post-censoring is assumed leading to all or none of the censored individuals had the outcome of interest
Correct model specification for the primary analysis

</td></tr>
<tr><td>What is Being Varied?</td><td>

The assumption made in the primary analysis that censoring is non-informative.

</td></tr>
<tr><td>Why (Objective)</td><td>

To assess the robustness of the treatment effect estimate to violations of the non-informative censoring assumption.
If results are stable across scenarios, confidence increases that findings are not driven by bias (due to censoring).
To understand the range of values of the treatment effect under the four extreme scenarios described above

</td></tr>
<tr><td>Strengths</td><td>

Offers a principled method for exploring alternative assumptions about censoring.
Sets bounds on the extent of maximum possible bias due to informative censoring

</td></tr>
<tr><td>Limitations</td><td>

Best/worst case scenarios are extreme assumptions.

</td></tr>
</table>

Table 14. Sensitivity analyses - Inverse Probability of Censoring Weighting

Analysis Methods	In the primary analysis we assumed censoring is independent of the outcome, conditional on the treatment, survival up to the time of censoring and indirectly baseline covariates used to estimate the treatment weights. An additional assumption is that the propensity score matching indirectly balances covariates between the censored and uncensored (i.e., outcome data is missing at random conditional on observed exposure and outcome). In the IPCW analysis, we use inverse probability of treatment weights and inverse probability of

	censoring weights. This analysis assumes censoring is independent of the outcome, with all common causes of both the outcome and censoring being accounted for Follow-up will be divided into equal 30-day intervals. At the start of each interval, we will update the information on baseline covariates available on each patient and assess whether they remain followed or have been censored. If they remain under observation, they contribute to the risk set for that interval. The weight for each participant at each interval is calculated as: The estimated probability of remaining uncensored, given a set of baseline and time-updated covariates that could affect both censoring and the outcome. Characteristics that could affect both censoring and the outcome include: treatment group (tolvaptan vs. unexposed), demographics (age, sex), baseline liver functions (e.g., AST, ALT) and comorbidities (liver diseases and disfunctions) The weight is calculated separately for each interval, and then multiplied together across all intervals of follow-up to give each participant's cumulative weight. The denominator probability will be estimated using pooled logistic regression fit to the person-interval dataset. In this model, the outcome is whether the participant was censored in that interval. Time will be modelled flexibly using restricted cubic splines. We will truncate weights at prespecified percentiles (1st and 99th).
Assumptions	- Censoring is conditionally independent of the outcome given covariates (i.e., non-informative censoring/censoring at random, conditional on other covariates beyond those in the analysis model). - Correct model specification - Positivity (every individual has a non-zero probability of remaining uncensored at each time point given their covariates). - Outcome does not directly influence its own missingness (would imply informative censoring via MNAR mechanism)
What is Being Varied?	- The assumption made in the primary analysis that censoring is non-informative given treatment group and not experiencing hepatotoxicity up to the time of censoring. - IPCW explicitly models informative censoring based on additional observed covariates.
Why (Objective)	- To assess the robustness of the treatment hazard ratio to violations of the non-informative censoring assumption made in the primary analysis. - If results are stable across scenarios, confidence increases that findings are not driven by bias (due to censoring).

Strengths	- IPCW adjusts for measured predictors of censoring. - Offers a principled method for recovering unbiased estimates under informative censoring.
Limitations	- IPCW is sensitive to model misspecification. - Cannot account for unmeasured factors affecting censoring. - Weighting can increase variance, especially if weights are unstable

7.6.6. Other Supplemental Analyses

We will report the baseline characteristics of the total population and by treatment group. For continuous variables we will report the median along with the interquartile range, while for categorical variables we will report the percentages. We will compare the two treatment groups by calculating the standardized mean differences for each baseline covariate, before and after matching. In addition, we will provide crude and adjusted Kaplan-Meier curves for the primary outcome (hepatotoxicity) in both study groups. We will also conduct descriptive analyses to characterize censoring patterns overall and across treatment groups. This will include median time to censoring overall and according to the reason for censoring. This will be estimated separately for the overall study population and by treatment arm (tolvaptan vs. unexposed).

Reasons for censoring will include:

Administrative censoring: reaching the maximum follow-up period of 3 years or the end of the study period (31 December 2024).

End of data availability: last recorded healthcare encounter, database end date, or practice withdrawal.

Loss to follow-up: deregistration from the contributing practice or migration out of the healthcare system.

We will also compare baseline characteristics between the eligible population and resulting study population after Propensity Score matching.

7.6.7 Core Emulation Table – Estimation Summary

Table 15. Core Emulation Table Estimand 1: Estimation Summary

Attribute	Target Trial	Target Trial Emulation	Comment
Analysis Method	Cox Proportional Hazards model.	Weighted Cox model to estimate HR, For all estimands, conditional propensity scores	Time matching and PS weighting used to emulate randomization in observational data.

	Endpoint: time from randomization to occurrence of first hepatoxicity event. Censoring: loss to follow-up, initiation of tolvaptan in control group, initiation of hepatotoxic medication in either treatment group (non-administrative censoring), censoring at death (administrative censoring) study end without hepatotoxicity (administrative censoring), database end (administrative censoring). Covariates: Treatment group (tolvaptan vs. control.	will be calculated, and patients will be matched accordingly. Endpoint: time from randomization to occurrence of first hepatoxicity event. Censoring: loss to follow-up (non-administrative censoring), censoring at death (administrative censoring) study end without hepatotoxicity (administrative censoring). Covariates: Treatment group (tolvaptan vs. control plus covariates selected for the propensity score model.	
Missing Data Assumptions and Methods to Handle	Censoring-at-random assumption conditional on treatment group and not experiencing hepatotoxicity up to the time of censoring. Censored participants contribute partial information with their time at risk up to the time of censoring.	Outcome: Same assumption, covariates included in the condition are different (PS matching included); administrative censoring used. Exposure: for missing exposure data, assume absence of refill or prescription/dispensation records for tolvaptan indicates true treatment discontinuation after 30 days. Covariates: absence of a diagnosis code will be interpreted as absence of the condition, while missing lifestyle variables (i.e. smoking status) will be imputed using multiple imputation by chained equations (MICE) under the missing at random assumption.	Mechanisms of missing exposure, covariate and outcome data differs between target trial and emulation (e.g., rather than leaving study, patients could be part of GP practice that no longer contributes data). Missing exposure data not possible in target trial but could be as a result of missing or incomplete prescription records in emulation. Multiple imputation would not occur for missing covariate data in target trial.

Statistical Model Assumptions	Proportional hazards assumption for Cox model Diagnostics: log-cumulative hazard plots	PH assumption tested using Schoenfeld residuals and log(-log) survival plots.	Diagnostics confirm appropriateness of Cox model; violations addressed in supplemental estimands and analyses (RMST).
Sensitivity Analyses	A sensitivity analysis under the CNAR assumption will be carried out regardless of whether assumption of CAR holds.	**Outcome misclassification** Methods: Probabilistic Bias Analysis using Monte Carlo Simulation at the summary level measure Purpose: To assess the robustness of the estimated treatment effect to plausible levels of outcome misclassification. Assumptions: • Misclassification is differential • Sensitivity and specificity of outcome classification are known or reasonably estimated from external data or expert opinion. • Misclassification affects only the outcome, not exposure or covariates. • The misclassification process can be simulated accurately. **Non-Informative Censoring conditional on additional baseline and time-varying covariates** ***Inverse probability of censoring weighting***	

		Methods: Estimate the probability of remaining uncensored at each time point *t* using baseline and time-varying covariates using logistic regression Purpose: To assess the robustness of the treatment hazard ratio to violations of the non-informative censoring assumption Adjusts for potential bias when censoring depends on additional measured baseline and time-varying covariates beyond those included in the primary analysis model. Assumptions: • Censoring is conditionally independent of the outcome given covariates and survival up to time t. • No unmeasured confounding for censoring. • Correct model specification and positivity. ***Informative censoring. Best/worst case scenario under censoring not at random assumption.*** Methods: For non-administrative censored individuals, repeat the analysis under four scenarios, assuming a) all censored individuals had the outcome of interest at the censoring date and b) none of the censored individuals had the outcome of	

		interest by the end of the study, separately for each treatment group. Purpose: To assess the robustness of the treatment effect estimate to violations of the non-informative censoring assumption. Assumptions: • All or none of the censored individuals had the outcome of interest • Correct primary analysis model specification and positivity.	

Estimand 2: Same as estimand 1 but without sensitivity analyses, except for censoring rules for intercurrent events as the while on treatment strategy is used for treatment discontinuation in the intervention group and for tolvaptan initiation in the control group. When these intercurrent events occur we will apply administrative censoring because the data after the intercurrent event are no longer of interest. Therefore, there is no 'missing outcome' data.

Table 16. Core Emulation Table Estimand 3: Estimation Summary

Attribute	Target Trial	Target Trial Emulation	Comment
Analysis Method	Accelerated failure time (AFT) model assuming Weibull distribution	Weighted Weibull AFT model estimating marginal restricted mean survival time (RMST) at years 3	AFT model selected to relax proportional hazards assumption; RMST improves interpretability of time-to-event differences.
Missing Data Assumptions and Methods to Handle	Assumes non-informative censoring; censored participants contribute partial information under AFT	Same assumption; administrative and non-administrative censoring applied; multiple imputation for missing covariates	Missing data handled using multiple imputation for covariates; censoring assumed non-informative and assessed in sensitivity analyses.
Statistical Model Assumptions	Weibull survival distribution; log-linear relationship between covariates and log survival time	Same; assessed using diagnostics such as log(-log(S(t))) vs log(t) for Weibull assumption and Q-Q plot for residuals	AFT assumptions (e.g., distributional form, non-informative censoring) assessed graphically; cannot be fully tested empirically.
Sensitivity Analyses	A Kaplan-Meier-based RMST will be used as non-parametric estimator as a sensitivity analysis instead of the AFT model	A Kaplan-Meier-based RMST will be used as non-parametric estimator as a sensitivity analysis instead of the AFT model	Target trial and emulation are the same.

7.7 Data sources

7.7.1 Data sources and Quality

Reason for selection:

CPRD offers large, high-quality, population based electronic health records with national coverage in the United Kingdom. The database provides sufficient data elements to conduct the study, including demographics, diagnoses, prescriptions and laboratory test results. Hospitalizations and mortality data are linked through NHS Digital. CPRD has been extensively used for post authorization safety studies and is known to support the emulation of clinical trial eligibility criteria and endpoints. The extensive follow-up time is easily sufficient for this current study with a three-year study period. Data extraction and feasibility assessments confirm that the study variables are sufficient in completeness and coverage to achieve the research objectives.

Strengths of data source:

CPRD covers over 19 million patients in total, with 16.5 million current acceptable patients in Aurum and nearly 3 million in GOLD. The data are updated frequently (monthly for GOLD, quarterly for Aurum), and mortality, prescribing, and diagnosis information are available with high completeness. In CPRD, there are 917 recorded tolvaptan users, of whom 704 have a diagnosis of ADPKD. 877 of 917 have had at least one AST/ALT measurement recorded and all have a recorded eGFR and creatinine clearance.

Limitations of data source:

A major limitation of using a database to assess hepatotoxicity is that patients without a tolvaptan prescription will not have regular measurements of liver enzyme values. Other missing values of lab measurements or diagnosis may also pose potential limitations. In CPRD, only the year of birth is available for adults, which may slightly impact precision in age-based eligibility criteria. Medication adherence cannot be determined from this prescription database, recorded tolvaptan users may therefore not actually use tolvaptan which could bias the results.

Data quality:

CPRD is managed by the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK. Data from participating general practices are extracted monthly (GOLD) or quarterly (Aurum) and undergo multi-level validation and quality checks. The CPRD provides data through its secure Trusted Research Environment or via multi-study licenses. Full metadata and SOPs are available at https://www.cprd.com/data-access and https://www.cprd.com/data-quality.

Databases' suitability and case-study feasibility assessments followed three key steps: (I) characterization of data source systems and processes, using the EMA data quality checklist to evaluate foundational aspects and their maturity; (II) assessment of data quality metrics for each data source

(data reliability), based on published research and open-access catalogues; and (III) fitness-for-use evaluation (data relevance), assessing database suitability for each case study based on question-specific determinants. Steps 1 and 2 were database-specific, while step 3 was both database- and case-specific, i.e. it could only be assessed in view of the specific research question to be addressed. From these steps, two tables containing qualitative information (I and III) and one with quantitative metrics (II) were created. The overall feasibility of the case studies using the candidate data sources was determined by critically analysing the collected information. Additional insights were gathered from DEAPs. All of the information was compiled into a report accompanying the generated tables, with our narrative assessment (appendix).

The overall feasibility assessment is summarised in Table 17. CPRD is deemed feasible to use as a data source for studying hepatotoxicity risks involved with tolvaptan use. It is expected that the target sample size for intervention and control groups will be reached. Elements with high criticality are available and fairly reliable with sufficient recency, although dispensing data is unavailable, discontinuation must be inferred from prescription duration, and diagnostic coding may be incomplete in some emergency room settings. Some other limitations are present, but these are manageable within the study design.

Table 17. Overall feasibility assessment summary

RWD source	Sample size estimation form the hypotetical trial protocol	Feasibility assessment *(yes/yes, with limitations/no)*	Rationale for the feasibility assessmernt	Limitations identified during the feasibility assessment and categorisation	Description of potential impact of the identified limitations on the study results
CPRD	With an approximate estimated sample size of 552 (based on a 1:1 ratio between intervention and control groups, with 276 participants in each), and considering that CPRD includes data from approximately 4.4 million inhabitants (as of 2014), autosomal dominant polycystic kidney disease (ADPKD) affects 1 in 400 to 1,000 people, and tolvaptan at its 90mg and 60mg dosages is prescribed to approximately 2,273 patients with ADPKD in England, the target sample size is anticipated to be reached. [10,11]	Yes	Elements with high criticality are available and fairly reliable. Data recency of 3 months before extraction, reasonably enough for the research question. Based on published information, the sample size proposed in the hypothetical trial protocol seems achievable.	·Minor: From the DEAPs experience, they use diagnostic of alcohol abuse rather than use alcohol consumption in units. ·Minor: Kidney volume, if not directly available, could be assessed though an internal algorithm that assesses lab tests/diagnoses of renal function. ·Minor: Liver function test is available as requested by GP, but this will surely have testing bias (higher availability for those clinically worse or with symptoms). ·Minor: The DEAP stated will probably only have basline liver function enzyme levels available for the tolvaptan users as it is mandatory before starting, but not in the unexposed patients. ·Minor: Treatment discontinuation not directly available but can be assessed using standard adherence calculation methods.	Given alcohol consumption detection is based on diagnostic codes, more severe cases are likely to be identified. So, alcohol use not associated with a particular disorder might be underrepresented and bias towards the most severe cases. As kidney volume is not readily available and needs to be derived as proxy, resulting values might be innacurate.

Table 18. Metadata about data sources and software

	Data 1
Data Source(s):	CPRD
Study Period:	2015 – latest available data
Eligible Cohort Entry Period:	2013
Data Version (or date of last update):	CPRD GOLD: November 2024 CPRD Aurum: September 2024
Data sampling/extraction criteria:	ADPKD patients at least 18 years old with a tolvaptan prescription or matched ADPKD patients with a time-matched physician visit without a tolvaptan prescription
Type(s) of data:	Primary care prescription data, linked secondary care prescription data, lab values, demographics
Data linkage:	Yes: deterministic linkage with HES, ONS mortality, NCRAS, etc. (linkage via NHS number, DOB, postcode, sex)
Conversion to CDM*:	Yes: deterministic linkage with HES, ONS mortality, NCRAS, etc. (linkage via NHS number, DOB, postcode, sex)
Software for data management:	EMIS Web for Aurum, Vision for GOLD; analyses via CPRD secure TRE or local TRE
HMA data catalogue link:	Clinical Practice Research Datalink (CPRD) GOLD | HMA-EMA Catalogues of real-world data sources and studies

*CDM = Common Data Model

7.8 Data management

The study will be conducted in a distributed manner using the UMCU, ARS Toscana and VAC4EU tools, procedures, and pipeline. Figure 1 specifies the data sets (D) and transformation processes (T), programming follows this pipeline, with involvement of different types of experts.

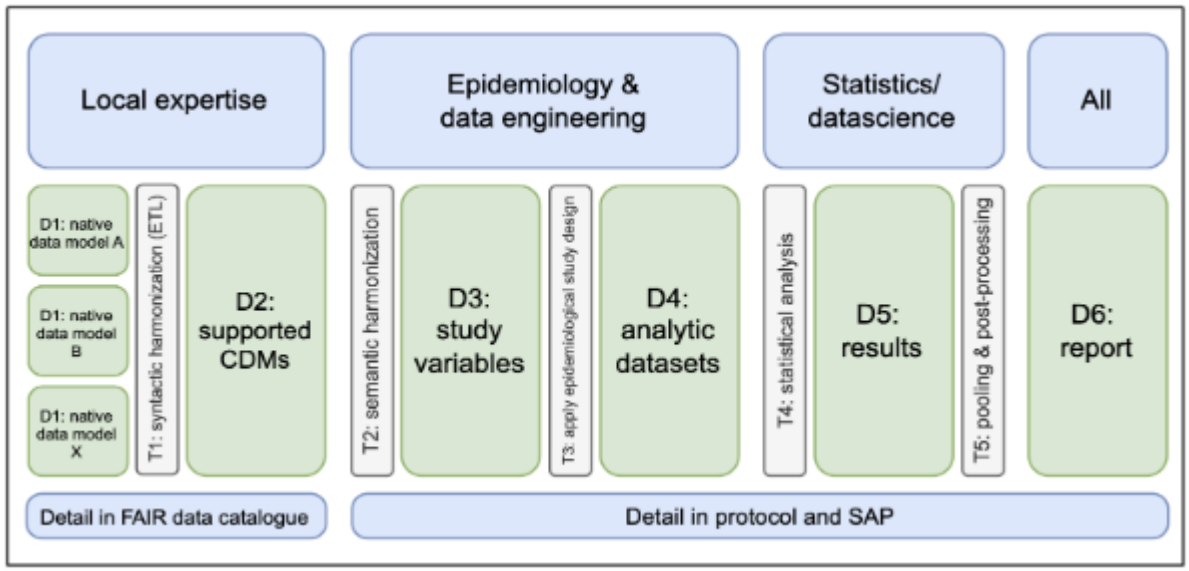


Figure 1. Data Management from the data transformation perspective

D1: Original data can be in any native format

The RWD-RWE pipeline used by VAC4EU starts with data banks that are controlled by the Data Expert and Access Partner (DEAP) and can be in any format. Data always stays local and never leaves the secure environments of the DEAPs. The ETL (extract, transform, load, see below for more details under 'T1') design is shared in a searchable FAIR VAC4EU catalogue. The VAC4EU FAIR Molgenis data catalogue is a meta-data management tool designed to contain searchable meta-data describing organisations that can provide access to specific data sources.

T1: Syntactic harmonisation (ETL)

T1: Syntactic harmonisation is conducted through an extraction, transformation, and loading (ETL) process of native data into the ConcePTION common data model (CDM) (see section 'D2: Common data model'). To harmonise the structure of the data sets stored and maintained by each data partner, a shared syntactic foundation is used. The ETL process has various structured steps as described by Thurin et al (2021)[8]

- DEAPs are asked to share the data dictionaries of their data banks (selected tables and variable names/structure)
- Metadata (descriptive data about the data sources and databanks) & data dictionaries, are uploaded in FAIR data catalogue (Molgenis).

D2: Common data model

For this project, the CDM (D2) is the ConcePTION common data model. The CDM version that is used is v2.2, which is available as an open-source CDM. In this CDM, data are represented in a common structure, but the values of the data remain in their original language (e.g. codes will have either ICD9/10/ICPC/SNOMED or MEDCODEID values).

T2: Semantic harmonisation

During the T2 step, many data transformations occur related to the completion of missing features in the data. Based on the relevant diagnostic medical codes and keywords, as well as other relevant concepts (e.g., medications), one or more phenotype algorithms are constructed (typically one sensitive, or broad, algorithm and one specific, or narrow, algorithm) to operationalise the identification and measurement of each event. In this step we conduct time anchoring (observation periods, look back periods), clean the data such as the dose of vaccines, sort on record level, aggregate across multiple records, and combine concepts for implantation of algorithms, and rule-based creation of study variables.

In this phase of the creation of study variables, semantic mapping is conducted. This semantic mapping across different vocabularies is conducted as part of the R-study script using different functionalities. To reconcile differences between different terminologies and native data availability, machine-readable code lists are used that comprise the terminologies that are used in the network (e.g. ICD-9, ICD10, SNOMED, ICPC and DEAP specific adaptations). This is combined with the BRIDGE metadata file that defines risk windows, look-back periods, and algorithms for each study variable (10).

D3: Study variables

D3 datasets are interim data sets with information on study variables for each study participant, the unit may be a person, a medicine, or episode of time. The design of these datasets is described in codebooks. Examples of D3 datasets are the outputs of the ConcePTION pregnancy algorithm (https://github.com/IMI-ConcePTION/ConceptionTools/wiki#conception-pregnancy-algorithm), and outputs of functions that define smoking. Multiple functions/packages exist within the VAC4EU, for different study variables.

T3: Application of epidemiological design

In the T3 step epidemiological designs are applied such as sampling, matching (on specific variables and/or propensity scores), and selection based on inclusion and exclusion criteria using the study variables in the D3 datasets. The designs will be implemented for the various study objectives using R-scripts, and these may use the existing functions (R-cran) or functions that have been developed in the VAC4EU community (e.g. matching).

D4: Analytical data set

D4 is an analytical dataset, and multiple D4 data sets may be produced based on the objectives of the study. The format is described initially in a code book for communication between programmers and statisticians.

T4: Statistical analysis

This step in the data transformation pipeline will produce statistical estimates such as descriptives (counts, percentages), distributions (mean, percentiles), rates (prevalence, incidence), regression coefficients, or other relevant estimates. This will be conducted using R.

D5: Results

D5 is the set of estimates, tables or aggregate data that is transferred from the DEAPs to the Digital Research Environment (DRE). The aggregated results produced by these scripts at the DEAP's site will be uploaded to the UMCU DRE for post-processing, pooling and visualisation (Figure 1). The DRE is a cloud-based, globally available research environment where data are stored and organised securely and where researchers can collaborate. The DRE is made available through UMCU. The DRE applies double authentication where researchers can collaborate using data that are stored and organised securely [ref]. UMCU is responsible for data processing and data security.

All researchers who need access to the DRE will be granted access to study-specific secure workspaces by UMCU. Access to the workspaces will be possible only after double authentication using an identification code and password together with the user's mobile phone for authentication.

Uploading files will be possible for all researchers with access to the workspace within the DRE. Downloading of files will be possible only after requesting and receiving permission from a workspace member with an "owner" role, who will be a UMCU team member.

T5: Post-processing/pooling

In this step, the result from different DEAPs is pooled and converted into tables and figures for reporting.

7.9 Quality control

For all data sources and for each data instance we will conduct *INSIGHT*[9] level 1-2 quality checks, detailed statistical analysis plans for the indicators are available on the public repositories:

- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level1 Hoxhaj, V. (2023). UMC-Utrecht-RWE/INSIGHT-Level1: https://doi.org/10.5281/zenodo.10035167
- https://github.com/UMC-Utrecht-RWE/INSIGHT-Level2 Hoxhaj, V., & van den Bor, R. (2023). UMC-Utrecht-RWE/INSIGHT-Level2: https://doi.org/10.5281/zenodo.10035169

Briefly, level 1 verifies Data Completeness and level 2 Data Consistency.

Level 1 – Data Completeness

The purpose of the level 1 check is to verify the completeness of the ETL process and the data in the variables. Examples of tests are:

- Presence of variables in each of the CDM tables in D2
- Checks for misspellings and letter case in variable names in each of the CDM tables
- Verification of vocabularies
- Check date formats
- Check conventions of values
- Missing data analysis
- Frequency tables for categorical variables

Level 2 – Data Consistency

Real data is not random but follows certain logical constraints that reflect rules governing real-world situations. Examples of indicators generated by level 2 checks are:

- Event dates before date of birth
- Event dates after date of death
- Event dates out of observation periods
- Subjects having an observation but not present in the PERSONS table
- Observations associated with a visit id and occurred before/after the visit start/end date
- Subjects younger than 12 years old reported as parents
- Age at the observation period older than 115 y old Data

Code Quality

These coding practices define how the TARGET programming team collaborates to write clean, reliable, and reproducible code for the VAC4EU Real-World Evidence (RWE) Analytical Pipeline. They aim to ensure clarity, consistency, and maintainability across all case studies within the project.

Coding conventions

To ensure clarity, consistency, and maintainability across the project, the following conventions will be applied to all codebases within the project:

- Consistent style: Code follows a consistent and readable style (see the tidyverse style guide for R).
- Meaningful names: Use clear, descriptive names for variables, functions, and files to convey their purpose.
- Modular code: Break down code into small, reusable functions where possible.
- No hardcoded paths: Use configuration files or relative paths to ensure portability.

Following these conventions makes the code easier to understand, test, and reuse across case studies and teams.

Documenting Code

Code documentation is used to promote good coding practices and ensure our work is understandable, maintainable, and reproducible. To achieve this, we will:

- Use descriptive comments that explain the purpose and rationale behind code sections, focusing on why something is done, not just what.
- Clearly document function inputs, outputs, and side effects, using standardized formats (e.g., roxygen2 in R) where appropriate and supported.
- Write meaningful variable and function names to make the code as self-explanatory as possible.

Version Control

We use Git and GitHub to manage version control. These tools support good coding practices by enabling collaboration, tracking changes, accessing a project's history, and ensuring code quality through review and documentation.

A dedicated GitHub organisation has been created for the project (https://github.com/target-roc19). Each case study is managed in its own repository within this organisation. Repositories are structured consistently across case studies, to reinforce modularity. Access to repositories is controlled through teams.

During development, all repositories remain private to ensure confidentiality. Once the project is finalised, relevant repositories will be made public and assigned a digital object identifier (DOI) via Zenodo to support transparency, reproducibility, and reuse by the wider research community.

To maintain code quality and clarity, we follow the git and GitHub guidelines below.

- Always use pull requests (PRs): never push directly to the main branch.
- Open an issue before creating a new branch. Ideally, one PR resolves one issue to keep changes focused and reviewable.
- Every PR must be reviewed by at least one other person before merging.
- The PR author merges the PR after it has been reviewed and approved.
- Write clear, descriptive commit messages.
- Write informative PR descriptions, including:
    - A concise title
    - Links to related issues
    - A summary of the changes

Continuous Integration

Continuous Integration (CI) is set up to automatically check code quality and run tests whenever changes are pushed to the repository or submitted through a pull request (PR). The CI workflow ensures that the package adheres to predefined style guidelines and that all automated tests pass before changes are merged.

Coding Template

Every case study follows the general coding template used across all code in the TARGET project. The folder structure is organised as follows:

```
case-study-template
|____data
| |____D2_cdm
| |____D3_study_variables
| |____D4_analytic_datasets
| |____D5_results
| |____D6_report
|____docs
|____logs
|____run
|____tests
|____transformations
| |____T2_semantic_harmonization
| |____T3_study_design
```

```
| |____T4_statistical_analysis
| |____T5_processing_results
|____CHANGELOG.md
|____LICENSE
|____README.md
```

Project Data Structure and Storage

The data folder follows the Real-World Evidence pipeline structure. Data conforming to the common data model is stored in the D2_cdm folder.

Results from transformations T2, T3, T4, and T5 are saved in the respective folders:

- D3_study_variables
- D4_analytic_datasets
- D5_results
- D6_report

Each dataset is associated with a codebook, explained in more detail below.

All data remain securely stored on the Data Expert and Access Partners (DEAPs) servers and are never transferred externally. For testing purposes, dummy datasets are created. These fall into two categories:

- Unit test data: Small, predefined input and output pairs used to test individual transformation steps. These are stored in the tests folder, not in data, and can support automated testing.
- Pipeline test data: Larger, more complex dummy datasets used to test whether the full pipeline runs as expected. These may be included in the repository only if they remain below GitHub's 100 MiB file size limit and will otherwise be shared via SharePoint.

Logging System

When the pipeline is executed, log files are saved in the logs folder. These logs are especially helpful when running the code in the DEAPs environment, as they help trace and diagnose potential errors. We recommend using the logger R package to handle logging throughout the pipeline. A sample logging setup can be found in the logger.R script located at the root of the project directory.

Executing the Analytical Pipeline

The run folder contains scripts used to execute each transformation step in the pipeline.

- A central script, run_pipeline.R, orchestrates the full pipeline from start to finish.
- Subscripts (e.g., run_T2.R or similar) are available to run individual transformation steps separately.

Typically, the run_pipeline.R script is the main entry point used by a DEAP to execute the full pipeline. Before running it in the DEAP environment, the pipeline may need to be adapted to local settings. This can be done using a configuration file that defines variables required to tailor the pipeline to a specific DEAP. Please note that configuration files should not include sensitive information.

Such a file might include variables like:

- The name of the DEAP
- The path to the local data instance
- The path to any required external resources

Testing and Quality Assurance

The tests folder contains scripts to test the analytical pipeline. Tests will be used to ensure code behaves as expected and remains stable over time. By systematically checking inputs, outputs, and edge cases, tests help catch errors early and make future changes safer. We use the testthat R package to structure and run unit tests.

Continuous integration (CI) is used to automate testing. With CI, tests are automatically run each time code is pushed to the repository (e.g., via GitHub Actions). This helps identify issues immediately, ensures that new changes do not break existing functionality, and supports better collaboration by enforcing consistent code quality across contributors.

Modular Data Transformation Workflow

The transformations folder follows the Real-World Evidence pipeline structure. It contains the source code for all transformation steps, which is typically written in R. Each subfolder corresponds to a specific step in the pipeline (e.g., T2_semantic_harmonization, T3_study_design, T4_statistical_analysis, T5_processing_results) and includes the relevant scripts and helper functions for that step.

During the T2 step, a database is usually created (e.g., using DuckDB). This database can be queried using SQL, and it is recommended that all SQL queries be saved as clearly named, standalone SQL script files to ensure readability and reusability.

The purpose of the transformations folder is to structure and modularise the processing logic, making it easier to maintain, test, and reuse across different case studies. By organising code by transformation step, teams can work in parallel, increasing efficiency.

Changelog

A changelog will be kept for all notable changes in the project. Changelogs help track the evolution of the project over time, making it easier for collaborators to understand what has changed between versions. We follow the structure and best practices outlined in Keep a Changelog.

Codebooks

Before developing code, codebooks are created to describe each dataset (D) within the pipeline. A codebook is a comprehensive document that outlines the structure, contents, and metadata of a dataset. It serves as a detailed reference guide for anyone working with the data and plays a crucial role in guiding the development of the analytical pipeline by clearly defining both the inputs and expected outputs.

All codebooks are summarized in a central index file, which provides a high-level overview of the pipeline's structure. For each codebook, the index file includes:

- A brief description of its purpose,
- A list of the scripts used to generate the corresponding dataset,
- A description of the input datasets and input parameters required.

The datasets D2, D3, D4, and D5 are typically subdivided into multiple smaller transformation steps, each detailed within their respective codebooks. These smaller transformation steps ensure that each part of the pipeline is clearly scoped and well-documented.

In addition to supporting development, codebooks help ensure quality control by making transformation logic transparent and verifiable, and they enhance reproducibility by documenting exactly how data is structured and used throughout the analytical pipeline.

Deployment

The analytical pipeline is delivered to DEAPs as a GitHub release, tagged with a version number. Versioning follows the format: vYYYYMMDD.XX, where the date indicates the release date and XX denotes the sub-version or revision number.

Any deployment issues can be reported via the GitHub repository using the issues feature, where the programming team responsible for the R code will collaborate with the local DEAP to resolve them as needed.

Reproducibility

It is recommended to locally use the renv R package to maintain the R version and version of packages for reproducibility purposes.

At this time, however, using renv reliably across different systems and environments remains challenging. For this reason, we currently recommend its use only in local development setups.

We are actively monitoring developments in the R ecosystem related to cross-platform reproducibility. As soon as a more stable and portable solution becomes available, we will revisit this guidance and promote broader adoption.

Licensing

The code will be made available under an open source license.

README Guidelines

Each case study repository includes a README that covers the following points:

- Project Overview: brief summary of the study goals and key research questions.
- Background: context and rationale for the study.
- Repository Structure: Outline of main folders and their contents.
- Data Overview: Description of data sources, formats, and data privacy considerations.
- How to Run: Instructions for running the pipeline and key scripts, plus where outputs are saved.
- Testing: How to run tests to verify code functionality.
- Contributing: Guidelines for code contributions and issue tracking.
- License: Information about the code license.
- Contact: Who to reach out to for help or questions.

7.10 Study precision

Sample size estimation from the hypothetical trial protocol:

In the target trial, the sample size was calculated for a hypothesis test as follows: Assuming an alpha of 0.05, a 1% withdrawal rate of the entire study period, a risk of hepatotoxicity of 1% in control group and 4% in intervention group, we will need a total of 552 participants and 17 events for 80% power.

Sample size estimation in this NIS protocol:

In this non-interventional study, no hypothesis test will be performed. The focus is on the precision of the estimated treatment effect. Assuming the study size in each RWD source will be similar to the sample size of the target trial, the precision is estimated as described below.
To estimate the level of precision that is achievable with a fixed sample size can estimate the expected width of the confidence interval for the effect estimate.

Estimation of the precision of the HR

To estimate the expected 95% CI for a HR from a Cox proportional hazards model, the standard error (SE) of the log(HR) is derived from the total number of events.

Assumptions
Equal allocation to treatment groups.
Large-sample normal approximation for log(HR).
Symmetric CI on the log scale.

The confidence interval (CI) width for the hazard ratio (HR) can be calculated using the following formula:

$$CI\_width\_HR = \exp(\hat{\beta} + 1.96 \times SE) - \exp(\hat{\beta} - 1.96 \times SE)$$

Where:
- $\hat{\beta}$ is the log hazard ratio (log(HR))
- SE is the standard error of the log(HR)
- 1.96 is the z-score for a 95% confidence interval

Calculation of 95% CI

1. Assume equal allocation:
Number of events per group: $d1 = d2 = d / 2$

2. Calculate SE of log(HR):

SE[log(HR)] = sqrt(1/d1 + 1/d2) = sqrt(2/d)

3. Construct the 95% CI on log scale:
log(HR) ± 1.96 × SE[log(HR)]

4. Convert back to HR scale:
CI_HR = exp(log(HR) ± margin)

Scenario 1

Calculation based on 17 events, HR = 4.0

SE = sqrt(2/17) = 0.3431
Log(HR)= log(4) = 1.3863

Margin = 1.96 X SE = 1.96 x 0.3431 = 0.6725
Lower bound = 1.3863 - 0.6725 = 0.7138
Upper bound = 1.3863 + 0.6725 = 2.0588

Lower CI = exp(0.7138) = 2.041
Upper CI = exp(2.0588) = 7.837

CI width = 7.837 - 2.041 = 5.796

Precision = (upper limit of CI / assumed HR) - 1 = 7.84/4 -1 = 96%

Scenario 2

Calculation based on 15 events (10% fewer), HR = 4.0

SE = sqrt(2/15) = 0.3651
Log(HR)= log(4) = 1.3863

Margin = 1.96 X SE = 1.96 x 0.3651 = 0.7156
Lower bound = 1.3863 - 0.7156 = 0.6707
Upper bound = 1.3863 + 0.7156 = 2.1019

Lower CI = exp(0.6707) = 1.956
Upper CI = exp(2.1019) = 8.170

CI width = 8.170 - 1.956 = 6.214

Precision = (upper limit of CI / assumed HR) - 1 = 8.17/4 -1 = 104%

Table 19. Power and sample size

Scenario	Number of Events	Hazard Ratio (HR)	log(HR)	Standard Error (SE)	Margin of Error	Lower CI (HR)	Upper CI (HR)	CI Width	Precision
Scenario 1	17	4	1.3863	0.3431	0.6725	2.041	7.837	5.796	96%
Scenario 2	15	4	1.3863	0.3651	0.7156	1.956	8.170	6.214	104%

8. Limitation of the methods

A limitation of the approach of comparing the original RCT with the TTE RWD study is the potential that the TTE cannot fully emulate all aspects of the original RCT, e.g. due to incomplete information of variables to assess in-/exclusion criteria, limited validity of outcomes, incomplete emulation of treatment strategies (e.g. dose escalating schemes). These limitations are inherent to the TTE approach, and an assessment will be made regarding the impact of these limitations on potential differences between RCT estimates and TTE RWD estimates. We will also perform quantitative bias analysis to better understand the potential roles of misclassification, confounding or selection bias.

Population, eligibility and setting

Eligibility criteria will be operationalised using diagnostic codes, lab measurements and prescriptions within the defined assessment windows relative to index date. These pragmatic decisions may lead to incomplete ascertainment of conditions compared with trial screening procedures. As a result, some patients included in the emulation may not have met the target trial's eligibility thresholds.

Treatment conditions and exposure

Treatment initiation is identified using the first prescription record, which may not fully capture adherence or prescription fill errors. Background symptomatic therapy is not restricted, reflecting real-world practice, but also introducing heterogeneity in treatment regimens. These factors represent departures from the controlled dosing and treatment monitoring procedures of a randomised trial.

Exposure is defined based on prescription records, which do not capture whether patients actually filled or took their medications. As a result, some individuals categorized as "exposed" may not have received the intended treatment. This limitation is acknowledged as a source of non-differential exposure misclassification, which would bias effect estimates toward the null. A sensitivity analysis will be implemented to assess the impact of varying levels of sensitivity and specificity in exposure classification, based on published estimates of prescription adherence.

Treatment assignment and follow-up

Randomisation cannot be emulated. Instead, matching on calendar time and propensity score is used to approximate exchangeability. While this matching balances measured baseline covariates, it cannot account for unmeasured confounding. Moreover, matching is done in quintiles which means that there will not be exact matching and some baseline differences between the control and intervention group may remain.

In the emulation, follow-up begins at the date of first prescription rather than randomisation. This aligns the start of follow-up with treatment initiation but differs from the controlled conditions of a trial. In addition, follow-up is based on data availability and practice registration rather than study visits, which may lead to imperfect measurement of loss to follow-up.

Outcomes and intercurrent events

Endpoints are identified using validated code lists in real-world data sources. These differ from the outcome in the target trial. Hepatotoxicity is in literature defined by elevations in liver enzyme levels. However, as liver enzyme levels are not regularly monitored in patients who do not use tolvaptan, this outcome cannot be used. Although liver injuries as captured through diagnostic codes are a reasonable alternative, this deviation in outcome may challenge extrapolation of the results of the study to clinical practice. Misclassification of hepatotoxicity remains possible, which may bias effect estimates.

Strategies for handling intercurrent events (treatment discontinuation, initiation of tolvaptan in the control group) are implemented using prescription records. While these reflect real-world effectiveness, the accuracy of operational definitions (e.g., gaps of >30 days to define discontinuation) is limited by prescribing and refill practices. For Estimand 2, in particular, misclassification of discontinuation or switching may bias estimates of while-on-treatment effects.

Loss to follow-up is defined using practice deregistration or database end, which are proxies for true loss. Although unique patient identifiers mitigate risks of missed de-registration, there remains potential for misclassification (e.g., patients who stop attending their practice but do not formally deregister).

Analysis methods and statistical assumptions

Unlike the target trial, which relied on randomisation to achieve balance, the emulation employs matching in Cox and AFT models. This approach requires unverifiable assumptions: no unmeasured confounding, correct model specification, and positivity. Departures from these assumptions could bias results.

The Cox proportional hazards model assumes that the relative treatment effect is constant over time, which may not be appropriate. This assumption is explicitly tested using Schoenfeld residuals and log-log survival plots. In addition, a supplemental analysis is conducted using an accelerated failure time (AFT) model with a Weibull distribution to estimate restricted mean survival time (RMST) at 3 years.

Missing data and censoring

Unlike the trial, where exposure and covariates were actively collected, the emulation must rely on real-world data. Missing exposure data may occur through incomplete prescribing records, while covariates (e.g., lifestyle factors) are imputed under a missing-at-random assumption. Misclassification is also possible when absence of a diagnostic code is assumed to reflect absence of a condition. These differences from trial data collection procedures may affect validity.

A key limitation of this study is the reliance on assumptions regarding the censoring mechanism. In the primary Cox model, we assume that censoring is non-informative, meaning that the probability of being censored is independent of the outcome, conditional on treatment assignment, baseline

covariates, and survival up to the time of censoring. This assumption may not always hold in practice. For example, patients may leave the database or deregister from a practice due to worsening health, which could be directly related to their risk of experiencing the outcome. If this relationship is not fully captured by measured covariates, effect estimates could be biased.

To address this, we perform sensitivity analyses using inverse probability of censoring weights (IPCW), which relax the primary assumption by conditioning not only on baseline confounders but also on time-varying covariates that predict both censoring and the outcome. While IPCW provides a more flexible and potentially more realistic adjustment, it remains dependent on correct model specification and the availability of sufficient data to capture predictors of censoring. If important determinants of loss to follow-up are unmeasured or poorly recorded, residual informative censoring may persist. In addition, a tipping point sensitivity analysis is conducted to assess how strong the association between censoring and outcome would need to be to change the study's conclusions.

9. Protection of human subjects

This is a non-interventional study using secondary data collection and does not pose any risks for individuals. Each data source research partner will apply for an independent ethics committee review according to local regulations. Data protection and privacy regulations will be observed in collecting, forwarding, processing, and storing data from study participants. This study involves data that exists in an anonymized structured format and contains no patient personal information. All parties will comply with all applicable laws, including laws regarding the implementation of organisational and technical measures to ensure the protection of patient personal data. Such measures will include omitting patient names or other directly identifiable data in any reports, publications, or other disclosures, except where required by applicable laws. Patient personal data will be stored at DAPs in encrypted electronic form and will be password protected to ensure that only authorised study staff have access. DAPs will implement appropriate technical and organisational measures to ensure that personal data can be recovered in the event of a disaster. In the event of a potential personal data breach, DAPs shall be responsible for determining whether a personal data breach has in fact occurred and, if so, providing breach notifications as required by law.

Patient consent

As this study does not involve data subject to privacy laws according to applicable legal requirements, obtaining informed consent from individuals is not required.

10. Reporting of adverse events

For studies in which the research team uses only data from automated healthcare databases, according to the International Society for Pharmacoepidemiology Guidelines for GPP. “Aggregate analysis of database studies can identify an unexpected increase in risk associated with a particular exposure. Such studies may be reportable as study reports, but typically do not require reporting of individual cases. Moreover, access to automated databases does not confer a special obligation to assess and/or report any individual events contained in the databases. Formal studies conducted using these databases should adhere to these guidelines.” For non-interventional study designs that are based on secondary use of data, such as studies based on medical chart reviews or electronic health records, systematic reviews, or meta-analyses, reporting of adverse events/adverse drug reactions is not required. Reports of adverse events/adverse drug reactions should only be summarized in the study report, where applicable. According to the EMA Guideline on GVP, Module VI – Management and Reporting of Adverse Reactions to Medicinal Products, “All adverse events/reactions collected as part of [non-interventional post authorization studies with a design based on secondary use of data], the submission of suspected adverse reactions in the form of [individual case safety reports] is not required. All adverse events/reactions collected for the study should be recorded and summarized in the interim safety analysis and in the final study report.” Module VIII – Post-Authorization Safety Studies echoes this approach. Legislation in the EU further states that for certain study designs such as retrospective cohort studies, particularly those involving electronic health records, it may not be feasible to make a causality assessment at the individual case level.

11. References

1. Mahboob M, Rout P, Leslie SW, et al. Autosomal Dominant Polycystic Kidney Disease. *StatPearls*. Published online March 20, 2024. Accessed July 10, 2025. https://www.ncbi.nlm.nih.gov/books/NBK532934/
2. National Health Service. Autosomal Dominant Polycystic Kidney Disease. NHS.
3. J GJ. Autosomal Dominant Polycystic Kidney Disease. *New England Journal of Medicine*. 2025;359(14):1477-1485. doi:10.1056/NEJMcp0804458
4. Jdiaa SS, Mustafa RA, Yu ASL. Treatment of Autosomal-Dominant Polycystic Kidney Disease. *American Journal of Kidney Diseases*. 2025;85(4):491-500. doi:10.1053/J.AJKD.2024.08.008
5. Müller RU, Messchendorp AL, Birn H, et al. An update on the use of tolvaptan for autosomal dominant polycystic kidney disease: consensus statement on behalf of the ERA Working Group on Inherited Kidney Disorders, the European Rare Kidney Disease Reference Network and Polycystic Kidney Disease International. *Nephrology Dialysis Transplantation*. 2021;37(5):825. doi:10.1093/NDT/GFAB312
6. Torres VE, Higashihara E, Devuyst O, et al. Effect of Tolvaptan in Autosomal Dominant Polycystic Kidney Disease by CKD Stage: Results from the TEMPO 3:4 Trial. *Clin J Am Soc Nephrol*. 2016;11(5):803-811. doi:10.2215/CJN.06300615
7. Torres VE, Chapman AB, Devuyst O, et al. Multicenter, open-label, extension trial to evaluate the long-term efficacy and safety of early versus delayed treatment with tolvaptan in autosomal dominant polycystic kidney disease: the TEMPO 4:4 Trial. *Nephrology Dialysis Transplantation*. 2017;33(3):477. doi:10.1093/NDT/GFX043
8. Edwards ME, Chebib FT, Irazabal M V., et al. Long-term administration of tolvaptan in autosomal dominant polycystic kidney disease. *Clinical Journal of the American Society of Nephrology*. 2018;13(8):1153-1161. doi:10.2215/CJN.01520218,
9. E TV, B CA, Olivier D, et al. Tolvaptan in Patients with Autosomal Dominant Polycystic Kidney Disease. *New England Journal of Medicine*. 2025;367(25):2407-2418. doi:10.1056/NEJMoa1205511
10. Watkins PB, Lewis JH, Kaplowitz N, et al. Clinical Pattern of Tolvaptan-Associated Liver Injury in Subjects with Autosomal Dominant Polycystic Kidney Disease: Analysis of Clinical Trials Database. *Drug Saf*. 2015;38(11):1103-1113. doi:10.1007/s40264-015-0327-3
11. Hoxhaj V, Navarro CLA, Riera-Arnau J, et al. INSIGHT: A Tool for Fit-for-Purpose Evaluation and Quality Assessment of Observational Data Sources for Real World Evidence on Medicine and Vaccine Safety. *medRxiv*. Published online October 30, 2023:2023.10.30.23297753. doi:10.1101/2023.10.30.23297753
12. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018 Jun 30;37(14):2252-2266. doi: 10.1002/sim.7654. Epub 2018 Apr 16. PMID: 29682776; PMCID: PMC5986623.
13. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med Res Methodol. 2017 Dec 6;17(1):162. doi: 10.1186/s12874-017-0442-1. PMID: 29207961; PMCID: PMC5717805.

12. Appendices

Figure 1. Flow diagram of patient selection and exclusion